Parse Scoring with Prosodic Information

N. M. Veilleux **M. Ostendorf** **C.W. Wightman***

Electrical, Computer and Systems Engineering
Boston University
Boston, MA 02215

Abstract

The relative size and location of prosodic phrase boundaries provides an important cue for resolving syntactic ambiguity, and can be used to improve the accuracy of automatic speech understanding. This paper describes an approach to scoring candidate sentence hypotheses and associated parses using prosodic phrase cues. Specifically, for each hypothesized parse, prosodic breaks are automatically detected and the probability of these breaks given the parse is computed based on a stochastic model of the prosody/syntax relationship. The parse probability can be used to rank sentence hypotheses and associated parses, optionally in combination with other scores. Both the prosodic break recognition algorithm and the prosody/syntax model can be automatically trained and can therefore be designed specifically for different speaking styles or task domains, given appropriate labeled data. We have demonstrated the potential of this approach in experiments with a corpus of ambiguous sentences spoken by FM radio announcers. Disambiguation performance is comparable to that of human listeners using the algorithm with hand-labeled breaks, although there is some performance degradation with the fully automatic system.

1 Introduction

Spoken language understanding is a difficult problem, in part because of the many ambiguities inherent in natural language. Syntactic ambiguity, which arises when a given expression corresponds to more than one syntactic structure, contributes substantially to the difficulty of the natural language processing problem. Several factors may be involved in resolving such ambiguities, including semantics, discourse and bias toward a specific syntactic structure. In spoken language, prosody, or the suprasegmental information in an utterance, is an important cue. Prosody is especially useful in automatic speech understanding, since computer representations of semantics and discourse are not as sophisticated as human knowledge.

Listeners can resolve several types of syntactic ambiguities by using prosodic information [2, 4]. The results of Price *et al.* [4] indicated that human listeners could reliably select the intended meaning of two target syntactic structures (86% correct identification for six out of seven types of structural ambiguity). Of the prosodic patterns studied in that work, the relative size and location of phrase boundaries seemed to provide the principal cue for resolving ambiguities. Thus, it seems likely that automatically detected prosodic phrase breaks could be used in speech understanding systems to reduce syntactic ambiguity.

There are many different ways in which prosody might be used for syntactic disambiguation for speech understanding. One approach would be to use automatically detected prosodic patterns directly in a parser as a grammatical constraint [5]. The approach taken here, an extension of [7], involves scoring an hypothesized parse in terms of the likelihood of prosodic breaks automatically detected based on the hypothesized word sequence from a speech recognizer. This probability can be used to rank parse hypotheses directly or in combination with parse probabilities. In a speech understanding system using an N-best search algorithm (e.g., [6]), the prosody/parse probability could also be used to score sentence hypotheses.

In the following section, we briefly describe the speech corpus and prosodic break index representation of prosodic phrase structure. Next, we will describe the parse scoring algorithm in more detail, and comment on different possible scoring variations. We then present experimental results on the task of automatically disambiguating confusable sentence pairs, comparing results to those of human listeners. Finally, we discuss the implications of the results for future work.

2 Labeling and Corpora

The parse scoring algorithm described here is based on an integer "break index" representation of prosodic constituent structure. Each word boundary in an utterance is labeled categorically with a break index from 0-6 that corresponds to a level in a constituent hierarchy, or equivalently, the amount of prosodic decoupling between words. A "0" represents the most tightly bound words, such as in clitic groups, while a "6" represents the prosodic break between sentences.[1] The correspondence between this seven-

*Now with the Department of Electrical Engineering at the New Mexico Institute of Mining and Technology.

[1] Although the "sentence" is not recognized as a prosodic constituent and utterance might be a more appropriate term, all the

level system of indices and various hierarchies proposed in the literature is discussed in depth in [12].

The experiments used to test the parse scoring algorithm are primarily based on the corpus of ambiguous sentences described in more detail in [4]. Four professional FM radio announcers were asked to read 35 pairs of sentences, where members of a pair were phonetically similar but associated with different syntactic structures and therefore different meanings. The sentences included five examples of each of seven types of structural ambiguity: parenthetical clauses vs. non-parenthetical subordinate clauses, appositions vs. attached noun (or prepositional) phrases, main clauses linked by coordinating conjunctions vs. a main clause and a subordinate clause, tag questions vs. attached noun phrases, far vs. near attachment of final phrase[2], left vs. right attachment of middle phrase, and particles vs. prepositions.

In addition to the ambiguous sentence corpus, we also use a corpus of radio news speech for training the new probabilistic prosody/syntax model. The data consists of 14 news stories, six from one announcer and eight from a second announcer, both female, for a total of 211 sentences (4210 words). Since these news stories are used only for training a model of the relationship between categorical breaks and syntax, only the word transcriptions are used (no acoustic information). These sentences differ from the ambiguous sentence pairs (the test data) in several ways. The ambiguous sentences are, on the average, shorter (7.6 vs. 19.6 words) and have a flatter syntactic structure (4 vs. 7 levels). In addition, the ambiguous sentence pairs include specific syntactic structures, by design, some of which are not generally found in FM radio news stories. For example, there are no sentences with tag questions in the radio news corpus.

For training and evaluating our algorithms, utterances have been hand-labeled according to the seven-level break index system. Sentences were also annotated with skeletal parses as part of a preliminary version of the University of Pennsylvania Treebank project [3].

3 Parse Scoring

The parse scoring algorithm has two modules: an algorithm for automatic break classification and a probabilistic model of the relationship between breaks and syntactic structure. These modules are described below, together with the parse score computation.

Both the break detection algorithm and the probabilistic prosody/syntax model developed for this work make use of binary decision trees, though they are different trees and are used in different ways. A binary decision tree [1] is an ordered sequence of binary questions that lead to one of a set of terminal nodes or leaves. A decision tree is designed from labeled training data by successively choosing questions about a feature vector (from a specified set) to minimize some impurity criterion on the sample distribution of classes associated with the leaves of the tree. At each leaf, the training data defines a relative frequency estimate of the posterior class probabilities. For both modules, the classes are the seven levels of breaks. For the prosody/syntax model, the probability distributions associated with the leaves of the tree are used directly in computing the parse score. In break classification, on the other hand, the most frequent break at a leaf is used as the assigned label in classifying future observations. Decision trees are particularly useful for prosody applications because they can easily model non-homogeneous features, both categorical and continuous variables, without requiring independence assumptions.

3.1 Automatic Break Detection

The automatic break detection algorithm used here, a slight variation of that described in [9], uses a decision tree to classify a sequence of feature vectors, one for each word boundary. The features are computed from speech annotated with phoneme and word boundary markers found automatically during speech recognition. In this work, the features include: average normalized duration of the rhyme of the word-final syllable, (a measure of preboundary lengthening), difference in the normalized duration of the rhyme and onset of the word-final syllable, absolute duration of a pause, probability of a boundary tone being present (computed by another decision tree [10]), a measure of local speaking rate change, and a flag indicating recognized lexical stress. Durations are normalized according to phone-dependent means and variances that are themselves adapted according to an estimate of the long-term speaking rate, using the algorithm described in [12].

A decision tree is designed to classify each feature vector independently as one of seven levels of breaks, giving the recognized break sequence $\hat{\mathbf{b}} = \{\hat{b}_1, \ldots, \hat{b}_n\}$ for a sequence of n hypothesized words. In this experiment with isolated sentences, sentence boundaries are known and can be assigned "6" deterministically. The break indices produced by the automatic labeling algorithm are reasonably accurate. Evaluating the algorithm on the ambiguous sentence corpus in speaker-independent mode, the correlation between the automatic labels and the hand-labeled indices is 0.84, and the two labels differ by 2 or more in 16% of the cases.

3.2 Probabilistic Break/Syntax Model

For computing the parse score, we need a model to represent the probability of the break index sequence given a parse. The parse is represented by a sequence of feature vectors, again, one for each word boundary, computed from the hypothesized word sequence and associated parse without use of acoustic information. The feature vector includes content/function word labels, labels of the syntactic constituents associated with different levels of left and right bracketing, a measure of distance in branches from the top and the bottom of the syntactic tree, and relative word location in the sentence (e.g., in the ith eighth of the

utterances used in this work are syntactically well-formed sentences.

[2] High vs. low attachment is probably more accurate syntactic terminology, but "far" vs. "near" is used in [4] as more descriptive.

word sequence). The questions about syntactic structure are based on Treebank skeletal parses [3], but in principle could be based on any parse representation.

The relationship between prosodic breaks and syntax is again represented using a decision tree, in this case one originally designed for speech synthesis applications. Wang and Hirschberg [8] first proposed the use of decision trees to predict the presence or absence of a prosodic break, with very successful results. Although their experiments involved predicting only one type of break, the model is general and can be extended to predict an arbitrary number of levels. In our case, the classes are the different levels of prosodic breaks (i.e., 0 – 6). Unlike the application to synthesis, which uses decision trees in a classification mode, the parse scoring application associates a leaf t of the decision tree with the probability distribution of breaks $p(b|t)$. Once the tree is designed, it can be used to encode the feature vector sequence, which represents the parse information, into a sequence of leaf node labels $\mathbf{t} = \{t_1, \ldots, t_n\}$. The probability distributions given by each leaf node in the sequence will be used to compute the parse score.

3.3 Probability Score

To compute the score of a candidate word sequence and associated parse, automatic break detection is used to obtain the recognized breaks $\hat{b}_i$ and the parse is encoded as a sequence of decision tree terminal nodes t_i, where $\hat{b}_i$ and t_i correspond to the boundary after the ith word in the sentence. The parse score is then given by

$$S_P = \frac{1}{n} \sum_{i=1}^{n} \log p(\hat{b}_i | t_i),$$

where $p(b|t)$ is the distribution associated with terminal node t and we assume that the breaks are conditionally independent given the parse feature vectors. The factor $1/n$ normalizes the score to account for differences in word length in comparing sentence hypotheses; otherwise, the score is biased to favor shorter sentences.

Rather than computing the probability of the sequence of breaks, we could have explicitly predicted a sequence of breaks from text, taking the most probable sequence, and then used the correlation scoring approach from our previous work [11] with these predicted breaks. However, if the predicted break indices are the same for both parses, then it is impossible to distinguish them using the correlation score, though it is possible to choose between them using the probability score if the syntactic structures correspond to different decision tree leaf sequences $\mathbf{t}$. This does occur for the ambiguous sentences, and therefore correlation scoring with the decision tree synthesis model has lower disambiguation performance than probability scoring.

4 Experiments

We have tested our parse scoring algorithm by using it to select between two interpretations of a spoken utterance, the same task that human subjects were asked to perform in [4]. Each spoken utterance yielded two sequences of recognized break indices, one for each of the two hypothesized word sequences, which were then scored according to the prosody/syntax model using the syntactic structure of the corresponding sentence hypothesis. We chose the most probable parse as the intended interpretation. In the event of a tie, the first sentence in the pair was chosen. In addition, we repeated the experiment using the hand-labeled break sequence instead of the detected breaks, in order to better understand the performance of the two model components. The parameters of the break classification tree were designed for a specific speaker using the ambiguous sentence data from the three other speakers in the corpus. A single prosody/syntax model was designed using the FM radio news stories, and the corresponding decision tree can be found in [7]. The results of these experiments are summarized in Table 1 for each of the 14 types (7 pairs) of syntactic ambiguity. For comparison, Table 1 also contains the results reported for the human subjects [4].

Ambiguity	Hand Labels	Machine Labels	Human Perception
+ Parenthetical	70	35	77
– Parenthetical	70	70	96*
+ Apposition	100	90	92*
– Apposition	70	50	91*
Main-Main	60	85	88
Main-Subordinate	65	60	54
+ Tag	100	90	95*
– Tag	80	55	81
Far Attach	65	65	78
Near Attach	70	55	63
Left Attach	80	80	94*
Right Attach	95	75	95*
Particle	100	55	82*
Preposition	95	80	81*
Average	80	68	84
Average for *	89	74	91

Table 1: Percent correct disambiguation as a function of syntactic ambiguity for: hand-labeled breaks, automatically labeled breaks and human perceptual results. Those categories which were identified by human listeners with significant reliability are marked with asterisks. Percentages are based on 5 sentences from each of 4 speakers, giving 20 utterances in each category and 280 utterances total.

Not all syntactic differences can be disambiguated by prosodic information, and such cases obviously cannot be handled by our algorithms. Therefore, Table 1 indicates with an asterisk the categories that were most reliably identified by human listeners (those for which mean response minus standard deviation was greater than chance). On this test subset, the results based on the hand-labeled break indices give disambiguation accuracy comparable to the performance of human listeners. When calculating the probability score based on the automatically detected breaks, there is significant degradation in performance, but the algorithm still offers an improvement over random guessing.

5 Discussion

In summary, we have introduced a prosody model that uses the probability of the acoustic observations given the candidate parse as a scoring criterion. We have evaluated this probabilistic parse scoring method on a set of 70 ambiguous sentences, each spoken by four radio announcers, with results under some conditions that are comparable to human listeners. The performance of the prosody/syntax model is very good, which is encouraging given the significant differences between the training data and the test data. The break detection algorithm, however, needs further research. Other possibilities for improving performance include incorporating other prosodic cues, such as prominence, and using the acoustic model from the break detection algorithm jointly with the probabilistic prosody/syntax model in scoring a parse.

The parse scoring algorithm on hand-labeled data shows some loss in accuracy relative to human performance if we also consider the sentences that were less reliably identified by the human listeners. Several different factors probably account for this effect, including the fact that the different spoken versions of these sentences simply exhibit more variability in prosodic phrasing. It is also likely that humans are using other prosodic cues in addition to phrase breaks to resolve ambiguities, such as phrasal prominence, which could be accounted for in extensions of the model.

Although the algorithm has been tested in a limited domain, all components of the model can be automatically trained and therefore applied to other domains. However, there are several issues that may affect performance in a spoken language system. First, the syntactic parses were hand corrected. Second, the sentences here represent a narrow range of syntactic classes and performance outside of this set may vary. Finally, the break recognition algorithm used phone segmentations from a recognizer constrained to the correct word sequence. While these issues need to be investigated, it is possible that use of prosodic parse scoring may help overcome problems in other components of a spoken language system, rather than be limited by them. For example, it is possible that using a prosodic parse score would enhance the overall performance of the system because recognition errors would cause a mismatch between the orthography and the prosody, leading to prosodic patterns with low probability.

Acknowledgments

The authors gratefully acknowledge Patti Price and Stefanie Shattuck-Hufnagel for their valuable suggestions and insights. This research was jointly funded by NSF and DARPA under NSF grant number IRI-8905249.

References

[1] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone C. J. *Classification and Regression Trees.* Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA. 1884.

[2] Lehiste, I. "Phonetic Disambiguation of Syntactic Ambiguity." *Glossa 7,* 2, pp. 107–121. 1973.

[3] Marcus, M. P. & Santorini, B. "Building a very large natural language corpora: The Penn treebank." Submitted manuscript.

[4] Price, P., Ostendorf, M., Shattuck-Hufnagel, S. & Fong, C. "The Use of Prosody in Syntactic Disambiguation." *Journal of the Acoustical Society of America 90,* 6, pp. 2956–2970. 1991.

[5] Price, P., Wightman, C. W., Ostendorf, M., & Bear, J. "The Use of Relative Duration in Syntactic Disambiguation." *Proceedings of the International Conference on Spoken Language Processing,* pp. 13–16. 1990.

[6] Schwartz R. & Austin, S. "Efficient, High Performance Algorithms for N-Best Search." *Proceedings of the DARPA Workshop on Speech and Natural Language,* pp. 6–11. 1990.

[7] Veilleux, N. and Ostendorf, M. "Probabilistic Parse Scoring Based on Prosodic Phrasing," *Proceedings of the DARPA Workshop on Speech and Natural Language.* 1992.

[8] Wang, M. & Hirschberg, J. "Automatic classification of intonational phrase boundaries," *Computer Speech and Language,* 6-2, pp. 175-196. 1992.

[9] Wightman, C. W. & Ostendorf, M. "Automatic Recognition of Prosodic Phrases." *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. 321–324. 1991.

[10] Wightman, C. W. & Ostendorf, M. "Automatic Recognition of Intonation Features." *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. 221-224. 1992.

[11] Wightman, C. W., Veilleux, N. M. & Ostendorf, M. "Using Prosodic Phrasing in Syntactic Disambiguation: An Analysis-by-Synthesis Approach." *Proceedings of the DARPA Workshop on Speech and Natural Language,* pp. 384–389. 1991.

[12] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. & P. Price. "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries." *Journal of the Acoustical Society of America 91,* 3, pp. 1707—1717. 1992.