Decision-Tree-Based Markov Model for Phrase Break Prediction

Sanghun Kim and Seung Shin Oh

ABSTRACT—In this paper, a decision-tree-based Markov model for phrase break prediction is proposed. The model takes advantage of the non-homogeneous-features-based classification ability of decision tree and temporal break sequence modeling based on the Markov process. For this experiment, a text corpus tagged with parts-of-speech and three break strength levels is prepared and evaluated. The complex feature set, textual conditions, and prior knowledge are utilized; and chunking rules are applied to the search results. The proposed model shows an error reduction rate of about 11.6% compared to the conventional classification model.

Keywords—Prosody, phrasing, TTS, speech synthesis.

I. Introduction

Prosodic phrasing plays a decisive role in making text-to-speech (TTS) natural and understandable because prosodic events have a great influence on how to correctly assign a phrase break to a sentence. A great deal of research has dealt with this issue. Most successful approaches use statistically driven methods based on the classification and regression tree (CART) and the hidden Markov model (HMM) [1]-[4]. In [1], a binary decision tree is used to predict the presence or absence of a phrase break. As classification features, parts of speech (POS) and syntactic constituent structure are introduced. In [2], a POS sequence Markov model is proposed. The most likely break sequence, given the input POS tags for a sentence, is sub-optimally determined by decoding Viterbi search. In [3], a hierarchical stochastic model based on binary tree classification and maximum probability prosodic parsing is proposed. The constituent length probability distributions and the probability of a specific prosodic parse are introduced. Recently, a Korean phrasing model based on CART has been proposed in [4].

Manuscript received Jan. 8, 2007; revised Mar. 28, 2007.
This research was supported by the Ministry of Information and Communication, S. Korea.
Sanghun Kim (phone: + 82 42 860 5141, email: ksh@etri.re.kr) and Seung Shin Oh (email: sshinoh@gmail.com) are with the Embedded S/W Research Division, ETRI, Daejeon, S. Korea.

II. Proposed Model

The system architecture of our proposed model, shown in Fig. 1, is similar to that of the hierarchical stochastic model. However, our model forms two function-based parts reflecting the static and dynamic characteristics of phrasing. The static term, or decision tree part, concentrates on the classification of the break strength type. It chooses the most probable breaks based on linguistic/textual context information. The dynamic term, or break sequence N-gram model, corresponds to the transition probability in HMM and reflects the temporal characteristics of break sequences. It prevents the predictors allocating unrealistic break sequences. By adopting this approach, we can utilize various heterogeneous features and extend a break sequence model to a complex model simply.

For the given linguistic observation $\overline{X} = \{x_1,...,x_n\}$, the goal of phrase break prediction is to find the corresponding phrase break sequences $\overline{B} = \{b_1,...,b_n\}$ that have the maximum posterior probability $p(\overline{B} \mid \overline{X})$ as expressed by

$$\overline{B}' = \arg\max_{\overline{B}} p(\overline{B} \mid \overline{X}) = \arg\max_{\overline{B}} \frac{p(\overline{X} \mid \overline{B}) p(\overline{B})}{p(\overline{X})}. \quad (1)$$

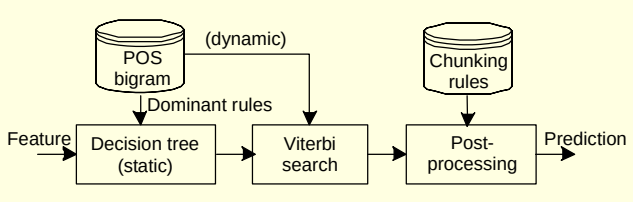

Fig. 1. Block diagram of the proposed system.

Since the maximization of (1) is carried out with the observation $\overline{X}$ fixed, the above maximization is equivalent to the maximization of

$$\overline{B}' = \arg\max_{\overline{B}} p(\overline{X} \mid \overline{B})\, p(\overline{B}). \quad (2)$$

To simplify the model, the first-order Markov chain and the output independence are assumed; therefore, (2) will be rewritten as

$$\overline{B}' = \arg\max_{\overline{B}} p(\overline{B} \mid \overline{X})$$
$$= \arg\max_{\overline{B}} p(T(x_1) \mid b_1)\, p(b_1) \prod_{i=2}^{n} p(T(x_i) \mid b_i)\, p(b_i \mid b_{i-1}), \quad (3)$$

where $p(T(x_i) \mid b_i)$ is provided by the decision tree and $p(b_i \mid b_{i-1})$ is the bi-gram probability of the phrase break sequence. The maximization problem (3) can be decoded using a Viterbi search algorithm. In the post-processing step, the search result information is chunked into two kinds of phrasing patterns. In the first pattern, the boundary of some consecutive words, for example, '느다고 그러셨는데(ɯndago gɯrʌsjʌn-nɯnde, in IPA code)', '와 마찬가지로(wɑ mɑçangɑziro)' never has a break. In the second pattern, the boundary of other consecutive words, for example, '과도 같이(gwɑdo gɑçi)', '과는 달리(gwɑnɯn dɑlli)' always has a major break.

III. Database

1. Corpus

A large text corpus containing broadcast news, call center dialogue, and intelligent robot dialogue was prepared. The text corpus was marked manually with three break strength levels by an expert. Since broadcast news announcers speak faster than typical people do, the virtual speaking rate was controlled to the speaking rate of typical people while tagging the break strength. Finally, 200,000 words were prepared for training, and 17,000 words were prepared for testing.

2. POS Tagging

In Korean, a word phrase called a 'euijeol' consists of a stem word (content word) and an ending word (function word). Since an ending word determines the syntactic structure of a sentence, the POS tag of an ending word has been commonly adopted as a morphological feature for phrase break prediction. We introduce a new POS tag set to reflect several distinctive features of prosodic phrasing in Korean [5].

a) *Combining stem word with ending word*: In a complex sentence, the combined POS tags of the stem word and the ending word are needed. For example, '가기를(gɑgirɯl)' has a verb term '가기(gɑgi)' and an objective functional particle /를(rɯl)/. In this case, the sentence has a connotation phrase. Thus, the following word boundary should be a major phrase break.

b) *Differentiating auxiliary verb from verb*: Auxiliary verbs specify the meaning of the verbs which follow them, such as '주다(zuda)', '버리다(bʌrida)', '내다(næda)', '보다(boda)', '말다(mɑlda)', and so on. In this case, the word boundary between the auxiliary verb and the main verb should not have a break. Thus, the auxiliary verb POS tag should be differentiated from the main verb POS tag.

c) *Differentiating case particle from other particles*: A case particle normally represents the subject of a sentence. The word boundary following the case particle normally forms a prosodic phrase. Thus, the case particle POS tag should be differentiated from other auxiliary particle POS tags.

d) *Differentiating specific dependent nouns from other dependent nouns*: The word boundary following several specific dependent nouns forms clause boundaries. (such as '뿐(bˈun)', '동안(doŋan)', '만큼(mɑnkɯm)', and so on). It distinguishes them from other dependent nouns.

The new POS tag set of 40 tags represents the syntactic information of a sentence.

IV. Experiment

1. Tree Conditions

A decision tree expands the depth and the number of leaf nodes which are asked the series of questions. Each parent node splits into child or leaf nodes with the best splitting rule minimizing the degree of impurity compared to that of previous tree status. The final node (leaf node) contains the purest phrase break results in terms of impurity measure (Gini index, Entropy). The posterior probability of the leaf nodes can be used for the phrase break classification.

For our experiment, we set the following tree conditions. For the impurity measure, we used the Gini index. The maximum number of nodes was set to 5,000, and the tree depth was 20. The minimum number of cases per parent node was set to 10, and the minimum number of cases per terminal node was set to 1. Five surrogates were used to construct the tree, and all surrogates counted equally.

2. Question Set

To grow the decision tree, several linguistic and shallow textural features are extracted and a question set is asked at each node. The syntactic information is already reflected in the newly defined POS information. To incorporate prior knowledge, POS bigram rules with high probability (> 0.9) are included.

The morphological features include preceding/current/following POS information (P_POS, C_POS, F_POS). The shallow textual features include the numbers of words to and from a comma (NWTC, NWFC) and the number of syllables to a comma (NSTC), as shown in Fig. 2, as well as the existence of a comma, the numbers of syllables in the preceding/current/following word, the number of syllables from a comma, and the numbers of words and syllables from the start and to the end of a sentence.

3. Tree Growing

Figure 2 shows the actual split conditions of the grown tree, which has 370 leaf nodes in optimal conditions. To determine which questions are more informative, we computed the entropy as bits for the given question set. The most important question is the bi-gram rule (0.49 bits) provided as prior knowledge. The current POS (0.47 bits)/following POS (0.44 bits) also contribute much to the phrase break prediction performance.

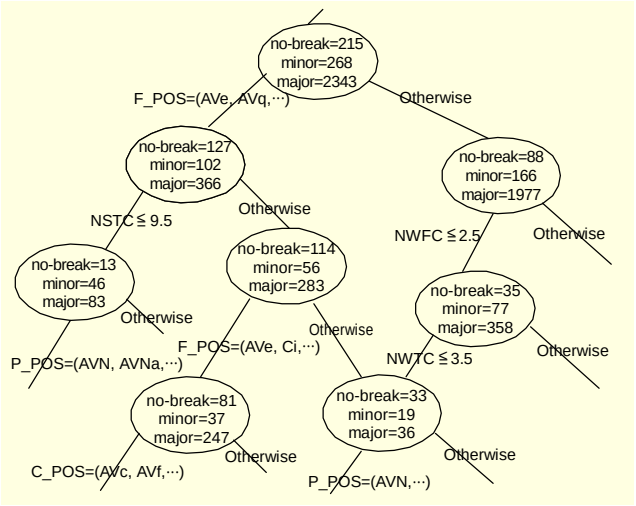

Fig. 2. Actual split conditions.

V. Results and Discussion

Table 1 shows the open test results of recall, precision, and F-score of the three prediction models: decision tree (DT), decision-tree-based Markov model (DT+MM), and decision-tree based Markov model with chunking information (DT+MM+chunking). In the cases of no break and major break strength prediction, the proposed DT+MM shows better performance than the conventional DT classification method. Though the performance of the minor break strength prediction is degraded, disagreement between the break strength predicted and the reference is negligible. For example, no break is predicted as minor, and no major break is predicted as a minor break. The overall performance of the proposed model, DT+MM+chunking, is 73.2% with an error reduction rate of 11.6%. In the chunking rules, 314 word pairs are classified as having a break and 2,155 word pairs are classified as having no break. Because a sentence can be phrased in various ways, the prediction results were checked subjectively by an expert. As a result, 80% of the breaks considered in our study were found to be acceptable.

Table 1. Recall (R), precision (P), and F-score (F) of three prediction models.

	no break	minor	major	percent correct
DT	R=78.5% P=87.1% F=82.6%	R=52.4% P=44.5% F=48.1%	R=69.0% P=62.4% F=65.5%	71.0%
DT+MM	R=89.8% P=79.7% F=84.4%	R=31.0% P=52.7% F=39.0%	R=69.2% P=63.6% F=66.3%	73.0%
DT+MM+ chunking	R=90.3% P=79.7% F=84.7%	R=30.7% P=53.1% F=38.9%	R=69.3% P=64.3% F=66.7%	73.2%

References

[1] M.Q. Wang and J. Hirschberg, "Automatic Classification of Intonational Phrasing Boundaries," *Computer Speech and Language*, vol. 6, no. 2, 1992, pp. 175-196.

[2] A.W. Black and P.A. Taylor, "Assigning Phrase Breaks from Part-of-Speech Sequences," *Proc. Eurospeech*, vol. 2, 1997, pp. 995-998.

[3] M. Ostendorf and N. Veilleux, "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location," *Computational Linguistics*, vol. 20, no. 1, 1994, pp. 27-52.

[4] K. Yoon, "A Prosodic Phrasing Model for a Korean Text-to-Speech Synthesis System," *Computer Speech & Language*, vol. 20, no. 1, 2006, pp. 69-79.

[5] S.S. Oh and S.H. Kim, "Modality-Based Sentence-Final Intonation Prediction for Korean Conversational-Style Text-to-Speech Systems," *ETRI Journal*, vol. 28, no. 6, Dec. 2006, pp. 807-810.