Experimental Evaluation of Tree-Based Algorithms for Intonational Breaks Representation

Panagiotis Zervas[1], Gerasimos Xydas[2], Nikolaos Fakotakis[1], George Kokkinakis[1], and Georgios Kouroupetroglou[2]

[1] Electrical and Computer Engineering Dept., University of Patras, Greece
{pzervas,fakotaki,gkokkin}@wcl.ee.upatras.gr
[2] Department of Informatics and Telecommunications, University of Athens, Greece
{gxydas,koupe}@di.uoa.gr

Abstract. The prosodic specification of an utterance to be spoken by a Text-to-Speech synthesis system can be devised in break indices, pitch accents and boundary tones. In particular, the identification of break indices formulates the intonational phrase breaks that affect all the forthcoming prosody-related procedures. In the present paper we use tree-structured predictors, and specifically the commonly used in similar tasks CART and the introduced C4.5 one, to cope with the task of break placement in the presence of shallow textual features. We have utilized two 500-utterance prosodic corpora offered by two Greek universities in order to compare the machine learning approaches and to argue on the robustness they offer for Greek break modeling. The evaluation of the resulted models revealed that both approaches were positively compared with similar works published for other languages, while the C4.5 method accuracy scaled from 1% to 2,7% better than CART.

1 Introduction

In speech communication, intonational phrases (IP) are separated by breaks in the form of pauses in speech. Accurate prediction of IP breaks in Text-to-Speech (TtS) synthesis heavily affects utterances structure and thus alters their understandability. As IP breaks divide utterances into meaningful 'chunks' of information [1], variation in phrasing can change the meaning listeners assign to utterances of a given sentence. For example, the interpretation of a sentence like "I will come because I was told so." (in Greek "Ta 'erTo epiD'i mu to 'ipan") will vary, depending upon whether it is uttered as one phrase or two. Situations where phrase breaks are missing when necessary or added in wrong places make the synthetic speech sound unnatural and boring.

In the past, the prediction of intonational boundaries for text-to-speech systems that handle unrestricted text was conducted using simple phrasing algorithms [2] based on orthographic indicators, keywords or part-of-speech spotting, and simple timing information. Research on the location of IP breaks was predicated on the relationship of prosodic and syntactic structures. Rule-based approaches [3] applied to this particular task were most successful in applications where syntactic and semantic information was available during the generation process. A weakness of this particular approach is that even if accurate syntactic and semantic information could be obtained automatically

V. Matoušek et al. (Eds.): TSD 2005, LNAI 3658, pp. 334–341, 2005.

and in real time for TtS, such hand-crafted rule systems are extremely difficult to build and maintain.

In addition, the relationship between prosody and syntax is not fully understood, though it is generally accepted that there is such a relationship. No current proposal integrating such information into the phrase assigned process has been shown to work well, even from hand-corrected labeled input. Some general proposals have been made which assume the availability of even more sophisticated syntactic and semantic information to be employed in IP breaks prediction [4].

In the field of IP break prediction, attention has been given by researchers in derivation of phrasing rules for text-to-speech systems from large labeled corpora [5]; most recently, attempts have been made to use self-organizing procedures to compute phrasing rules automatically from such corpora. The primary learning techniques currently being used include Hidden Markov models [6], neural networks [7], classification and regression trees (CART) [8], transformational rule-based learning (TRBL) [9] and Bayesian techniques [10].

The most commonly used feature set in such training frameworks for IP break prediction include part-of-speech (POS), pitch accent, syntactic structure, duration, length of the current sentence, number of words and syllables from the last break, etc. From the above POS has been proved to be an effective and easy to derive feature.

In this work, we inspect on the performance of tree-structured predictors for IP breaks placement. Along with the commonly used CART approach, we introduce a C4.5 classifier to evaluate over a rapid extracted shallow textual feature set. The experiments were carried out by utilizing two speech corpora in the Greek language provided by the University of Patras (Artificial Intelligence Group) and the University of Athens (Speech Group).

2 Data Resources

For the analysis of the proposed approaches experiments were conducted with the exploitation of two prosodic annotated datasets. The first one featured prosodical phenomena encountered in a generic textual environment while the other was derived from a museum domain text corpus. Professional speakers uttered both corpora in Athenian dialect. Both corpora were annotated to the full ToBI specification and checked for their consistency.

2.1 Corpora Description

The generic corpus consists of 5.500 words, distributed in 500 paragraphs, each one of which may be a single word utterance, a short sentence, a long sentence, or a sequence of sentences. For the corpora creation we used newspaper articles, paragraphs of literature and sentences constructed and annotated by a professional linguist. The corpus was recorded under the instructions of the linguist, in order to capture the most frequent intonational phenomena of the Greek language.

The museum domain corpus includes exhibits' descriptions from a museum guided tour. It consisted of 5484 words, distributed in 516 utterances. Half of the corpus contains grammatically restricted texts, while the remaining half is unrestricted text [8].

As the original corpus included enriched linguistic information provided by a Natural Language Generator, the corpus was recorded appropriately in order to capture a big variety of emphatic events, for example by the introduction of new or old mentioned information to the visitor.

2.2 Shallow Features

In order to predict the juncture class of an IP, textual features were incorporated. Apart from POS, researchers have stressed the important role of syntactic and morphological information for several languages. Taking into account that in real-time IP break prediction tasks, fully syntactic parsing would be time-consuming and would produce many syntactic trees, as well as that in several languages, including MG, syntactic tools are not freely available, a syntactic feature labeling each word with the syntactical chunk which belongs in a sentence was introduced [10]. The phrase boundary detector [12], or chunker, is based on very limited linguistic resources, i.e. a small keyword lexicon containing some 450 keywords (articles, pronouns, auxiliary verbs, adverbs, prepositions etc.) and a suffix lexicon of 300 of the most common word suffixes in MG. In the first stage the boundaries of non-embedded, intra-sentential noun (NP), prepositional (PP), verb (VP) and adverbial phrases (ADP) are detected via multi-pass parsing. Smaller phrases are formed in the first passes, while later passes form more complex structures. In the second stage the head-word of every noun phrase is identified and the phrase inherits its grammatical properties.

2.3 Task and Feature Definition

For the purpose of IP breaks prediction within TtS, it is common to flatten the prosodic hierarchy, hence a word juncture is considered to be a break or a non break.

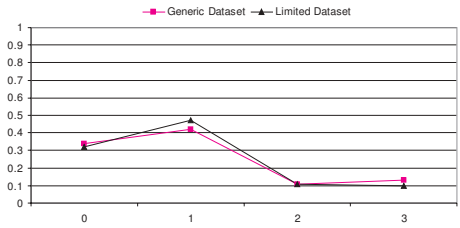

Fig. 1. IP breaks distribution in corpora

In an effort to deviate from that, we considered word junctures of the whole IP break marks proposed by ToBI transcription. Therefore our phrase break label files contain break indices ranging from 0 to 3 (b0, b1, b2 and b3), describing the strength of the juncture between each two lexical items; where b0 is representing that cliticization has merged two lexical items into a prosodic word while b3 is indicating a maximal, or fully-marked, intonational phrase boundary.

Our task was the derivation and application of a common set of shallow textual features extracted rapidly from text for both corpora and the application to the decision tree

classifiers for IP breaks placement. Previous works have shown the optimized performance of both models using their full feature set [10], [8] in predicting prosodic phrase breaks, pitch accents and endtones.

Eventually, in order to facilitate the evaluation of the IP break prediction models, we adapted both databases according to the following feature vector:

- *pos:* the part of speech of the word. Values: verb (V), noun (N), adjective (ADJ), adverb (ADV) and a function word (FW) class holding non-content word pos types. For our experiments, the POS of the words in a window of -2,+1 words was employed.
- *chunk:* a syntactic feature that has been successfully applied to intonational phrase break detection [10]. These information is considered as shallow syntactic information, it is unambiguous and can be extracted rapidly [13]. In this work we introduce some combinational features extracted from syntactic chunking and information provided by punctuation. These features are described below:
  - *parent_chunk:* a binary indicator showing whether a word belongs to a different syntactic chunk than its previous one. A window of -1,+1, around the word, was utilized.
  - *chunk_break:* the distance in words from the beginning of the next syntactic chunk or of a major punctuation break.
  - *neigh_chunk:* a binary indicator that shows whether a word belongs to the same syntactic chunk with its next one. A window of -1,+1, around the word, was utilized.
- *word_in:* feeds the classifier with the information of words position from previous major punctuation break.
- *word_out:* presents the number of words until a major punctuation break.
- *syll_num:* the number of syllables in the present word. The values of this feature ranges from 1 to 5 where the last class (5) includes any polysyllabic words with 5 or more syllables. The latter group contains all the low frequency classes of word syllables.
- *syll_str_strct:* indicates the index of the syllable that holds the lexical stress in the word. The values for the Greek language are final, penultimate, antepenultimate and none. The above features were applied to the word level.

3 Phrase Break Prediction Schema

The present study provides an insight into the prosodic parameter classification experiments conducted into ToBI annotated corpora for IP break prediction. The windowed data described above was firstly applied to a decision tree inducer (CART) [13]. Furthermore, C4.5 [14] algorithm was employed. Decision trees have been among the first successful machine learning algorithms applied to IP break and pitch accent prediction for TtS. The three basic elements that a decision tree is composed of are:

- a decision node specifying a test feature.
- an edge or a branch corresponding to the one of the possible attribute values which means one of the test attribute outcomes.
- a leaf which is also named an answer node contains the class to which the object belongs.

In decision trees two major phases should be ensured; the phase of tree building on a given training set, and the classification. In order to classify a new instance, we start by the root of the decision tree, then we test the attribute specified by this node. The result of this test allows the tree branch relative to move down to the attribute value of the given instance. This process will be repeated until a leaf is encountered. The instance is then classified in the same class as the one characterizing the reached leaf. Several algorithms have been developed in order to ensure the construction of decision trees and its use for the classification task.

3.1 Classification and Regression Trees (CART)

The Regression trees, induced by the CART method, are a statistical approach for predicting data from a set of feature vectors. In particular, a CART uses a binary decision tree to model a conditional distribution. CART contains yes/no questions regarding the features and provides either the probability distribution or a mean and standard deviation.

CART analysis consists of four basic steps. Initially a tree is built by means of recursive splitting of nodes. All resulting nodes are assigned with a predicted class, based on the distribution of classes in the learning dataset which would occur in that node and the decision cost matrix. In each node a predicted class assigned whether or not that node is subsequently split into child nodes. The next step consists of stopping the tree building process. At this point a "maximal" tree has been produced, which probably greatly overfits the information contained within the learning dataset. The resulted "maximal" tree is "pruned", which results in the creation of a sequence of simpler trees, through the cutting of increasingly important nodes. Optimal tree selection of the resulted simpler trees is the fourth step, during which the tree fitting the information in the learning dataset, but does not overfit the information, is selected among the sequence of pruned trees.

3.2 C4.5 Algorithm

C4.5 is an improvement of the ID3 [14] algorithm being able to handle numerical data. The first task for C4.5 is to decide which of the non-target variable is the best to split the instances. Then, every possible split is tried. The value of potential splits in C4.5 is calculated from a criterion called information ratio. Information ratio suggests an estimate of how probable split on a variable will lead the decision to a leaf containing the fewer errors or has low disorder. The concept of low disorder means that the node contains instances with one major target variable.

Calculation of information ratio is realized for all the variables and the 'winner' variable is the one with the largest information ratio and is chosen as the split variable. The tree will grow in a similar method. For each child node of the root node, the decision tree algorithm examines all the remaining attributes to find candidate for splitting. If the field takes on only one value, it is eliminated from consideration since there is no way it can be used to make a split. The best split for each of the remaining attributes is determined. When all cases in a node are of the same type, then the node is a leaf node.

C4.5 uses a method called pruning to avoid overfitting. There are two types of pruning applied in the C4.5 procedure: pre-pruning and post-pruning. Post-pruning refers to the building of a complete tree and pruning it afterwards, making the tree less complex and also probably more general by replacing a subtree with a leaf or with the most common branch. When this is done, the leaf will correspond to several classes but the label will be the most common class in the leaf. Post-pruning is affected by a parameter called confidence interval. The application of lower confidence results more drastic pruning. For our models we applied a confidence value of 25 %. Pre-pruning concerns the decision about when to stop developing subtrees during the tree building process. For example specifying the minimum number of observations in a leaf we can determine the size of the tree. A minimum number of 2 was utilized in our model. We have to also point out that in our case we applied post-pruning only while pre-pruning application showed any difference to the models resulted tree. After a tree is constructed, the C4.5 rule induction program can be used to produce a set of equivalent rules. The rules are formed by writing a rule for each path in the tree and then eliminating any unnecessary antecedents and rules.

4 Evaluation

The performance of the proposed approaches with the induction of the suggested features was measured by the utilization of f-measure metric per each IP break class, as they have been explained in Section 2. Results were obtained using the 10-fold cross validation method [15]. Defining f-measure (FM), is the harmonic mean of precision and recall, calculated as:

$$1/\left(\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}\right) \tag{1}$$

α is a factor determining the weighting of precision and recall. The value of $\alpha = 0.5$ has been used for the current evaluation for equal weighting of precision and recall.

4.1 Results

In an attempt to evaluate the IP break models we calculated the total accuracy, kappa statistic, mean average error (MAE) and root mean square error (RMSE).

Table 1. Total accuracy of the IP break models

Methods	Generic	Museum
CART	83.79%	87.71%
C4.5	86.02%	88.62%

All resulted models revealed a kappa statistic higher than 0.75 which is generally regarded as a good statistic correlation. Accuracy of the models is tabulated in Table 1. We can see that in all cases C4.5 performed better than CART especially in the case of generic dataset.

It is clear that MAE values for all models are close to the corresponding RMSE values giving us the insight that there were not test cases in which the prediction error was significantly greater than the average prediction error.

A detailed observation of the prediction of each class presented by each model, the following can be derived. In Figure 2 the f-measure for each IP break class is depicted for the generic dataset. For these models, classification for the IP break cases with the highest occurrence in the dataset along with class b3, performed better.

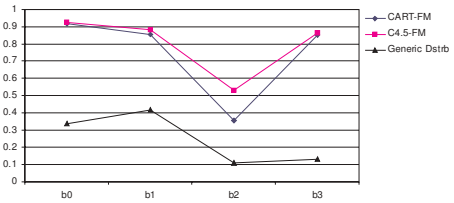

Fig. 2. F-measure for generic domain models

As regards b3 the high f-measure is a result of the fact that this class has low correlation with the other. C4.5 performed better than CART especially in the prediction of the b2 category. Non-breaks were predicted with an f-measure higher than 0.9 for both models.

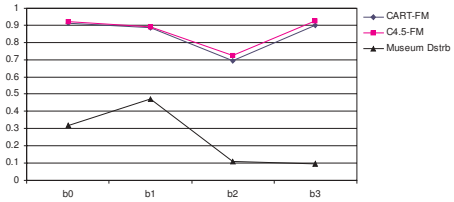

Fig. 3. F-measure for museum domain models

The next step of our exertion was the evaluation of the models derived from the museum domain dataset. F-measure of those models is illustrated in Figure 3. For this domain, both approaches performed sufficiently well with C4.5 performing slightly better almost in every category. For the prediction of non-breaks, both approaches achieved a score of f-measure more than 91% while the lowest was the prediction of b2 class with CART with an f-measure of 0.7.

5 Conclusions

We constructed IP break prediction models with the utilization of decision trees and the induction of shallow textual features. Specifically, we trained CART and C4.5 decision trees with a generic and a museum domain corpus. Both algorithms performed equally well in the prediction of all classes. As expected, museum domain models gave higher

prediction scores for all IP break classes as breaks are described by simpler "rules" due to the limitation of the domain. C4.5 performed slightly better than CART in all models. Furthermore, museum domain models showed high prediction accuracy for both approaches with C4.5 having the highest score.

The shallow textual feature set used in these experiments showed an improvement in the prediction of IP break prediction classes of b1, b2 and b3, especially in the case of the generic dataset, compared to the feature set used in earlier works by the two Universities [10], [8]. Further improvements in the tree-structured predictors can be achieved by the introduction of more delicate linguistic features as has been inspected on [16] for the CART approach.

References

1. Bolinger, D.: Intonation and its Uses: Melody in Grammar and Discourse, London, UK, Edward Arnold, 1989.
2. Anderson, M., Pierrehumbert, J., Liberman, M.: Synthesis by rule of English intonation patterns", ICASSP, pp. 281-284, 1984.
3. Prieto, P., Hirschberg, J.: Training Intonational Phrasing Rules Automatically for English and Spanish text-to-speech, Speech Communication, 18, ps. 281-290, 1996.
4. Bachenco, J., Fitzpatrick, E.: A Computational grammar of Discourse-Neutral Prosodic Phrasing in English, Computational Linguistics 16(3), 155-170, 1990.
5. Ostendorf, M., Veilleux, N.: A hierarchical stochastic model for automatic prediction of prosodic boundary location, Computational Linguistics, 20(1), 1989.
6. Taylor, P., Black, A.W.: Assigning Phrase Breaks from Part-of-Speech Sequences, Computer Speech and Language 12:99-117, 1998.
7. Muller, A. F., Zimmermann, H. G., and Neuneier, R.: Robust Generation of Symbolic Prosody by a Neural Classifier Based on Autoassociators, ICASSP-96, 1285-1288, 1996.
8. Xydas, G., Spiliotopoulos, D. and Kouroupetroglou, G.: "Modeling Prosodic Structures in Linguistically Enriched Environments", in "Text, Speech and Dialogue", Lecture Notes in Artificial. Intelligence. (LNAI), Springer-Verlag Berlin Heidelberg, Vol 3206, pp. 521-528, 2004.
9. Fordyce, C., S., Osterdorf, M.: Prosody Prediction for Speech Synthesis Using Transformational Rule-Based Learning, ICSLP-98, 682-685, 1998.
10. Zervas, P., Maragoudakis, M.: Fakotakis, N., Kokkinakis, G., Bayesian Induction of intonational phrase breaks, EUROSPEECH, Geneva, Switzerland, Sept. 1-4, 2003, pp. 113-116, 2003.
11. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.: ToBI: A standard for labeling English prosody,. ICSLP, pp. 867-870, 1992.
12. Stamatatos, E., Fakotakis, N. and Kokkinakis, G.: A Practical Chunker for Unrestricted Text, 2nd Int. Conf. of Natural Language Processing, pp. 139-150, 2000.
13. Breiman, L., Friedman, J.H., Olshen, R. A., Stone C. J.: Classification and Regression Trees, Belmont, CA: Wadsworth International Group, 1984.
14. Quinlan, J.R.: C4.5: Programs for Machine Learning, San Francisco: Morgan Kaufmann Publishers, 1993.
15. Stone, M.: Cross-validation choice and assessment of statistical predictions, Journal of the Royal Statistical Society, 36, 111-147, 1974.
16. Xydas G., Spiliotopoulos D., Kouroupetroglou G.: "Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora", in IEICE Transactions of Information and Systems, 2005 (to appear).