Proceedinds of the 9th Conference on Speech Communication and Technology (INTERSPEECH 2005), Lisbon, 4-8 Sept. 2005

Tree-Based Prediction of Prosodic Phrase Breaks on top of Shallow Textual Features

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

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

Abstract

This paper reports on the evaluation of automatic prosodic phrase break assignment. We utilize two tree-structured predictors, the commonly used CART and a C4.5, to predict break placement from sequences of easy to extract shallow textual features. We are experimenting with two 500-utterance sized prosodic corpora developed by two Greek universities that originate from different domains in order to focus on the differences in prediction between generic and limited domain datasets. The evaluation shows that while the limited dataset achieves better accuracy than the generic one in the CART case, this difference is lowered with the introduction of C4.5. Minor breaks proved to be the most difficult class to predict in CART case, while we achieved a 50% of improvement with C4.5.

1. Introduction

Word juncture prediction accuracy in Text–to-Speech (TtS) synthesis heavily affects the structure of utterances, thus altering their understanding. Prosodic phrase breaks (PPB) divide utterances into meaningful 'chunks' of information [1] and thus variation in phrasing can change the meaning listeners assign to utterances of a given sentence. Situations where phrase breaks are missing when necessary or added in wrong places make the synthetic speech sound unnatural and boring.

In the past, such prediction 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 PPB 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 and in real time for TtS, such hand-crafted rule systems are extremely difficult to build and maintain. Some general proposals have been made which assume the availability of even more sophisticated syntactic and semantic information to be employed in PPB prediction [4].

Corpus-based synthesis has turned the attention of 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 most commonly used learning techniques are Hidden Markov models [6], neural networks [7], classification and regression trees (CART) [8], transformational rule-based learning (TRBL) [9] and Bayesian [10].

In this work, we inspect on the performance of prosodic phrase breaks placement prediction commencing rules learned from decision tree classifiers.. Along with the commonly used CART approach, we introduce a C4.5 classifier to evaluate over rapid extracted sequences of shallow textual features, to oppose earlier work where, more delegate and hard to extract linguistic features showed to improve break prediction [16].

As the most common problem of machine learning approaches is the difficulty to classify unseen data, especially when using undersized training data, our experiments were carried out by utilizing two speech corpora in the Greek language provided by the University of Athens (Speech Group) and the University of Patras (Artificial Intelligence Group). The first corpus is considered to be limited to museum domains while the latter is a generic, phonologically balanced one.

2. Tree -based classifiers

The present study provides an insight into the prosodic parameter classification experiments conducted into ToBI annotated corpora of Greek speech for PPB prediction. Decision trees have been among the first successful machine learning algorithms applied to PPB and pitch accent prediction for TtS. We first experimented with (CART) [13] inducer. Furthermore, C4.5 [14] algorithm was employed.

CART have been widely used in speech technologies due to their ability to deal with incomplete data and multiple types of features. In C4.5, binary decision is carried out in the nodes of a decision tree producing a set of logical rules. Therefore, every path starting from the root of a decision tree and leading to a leaf is representing a rule. The number of rules embodied to a given tree is equal to the number of its leaf nodes. The premise of every rule is the conjunction of the decisions leading from the root node, through the tree, to that

leaf, and the conclusion of that rule is just the category that the leaf node belongs to.

For the growth of C4.5 trees the basic algorithm used was a greedy method constructing the tree in top-down recursive divide and conquer manner. In C4.5 tree algorithm the procedure of *pruning* is performed. Pruning is process that is not included in some of its antecedent, such as the *ID3 tree* [14]. Unlike the stop splitting strategy, pruning is performed when a tree is grown fully and all the leaf nodes have minimum impurity. C4.5 selects a working set of examples at random from the training data and the tree growing/pruning process is repeated several times to ensure that the most promising tree has been selected.

3. Data Resources

In corpus-based speech technologies, the quality of the selected data heavily affects the analysis results. Our experiments were conducted with the exploitation of two prosodic annotated datasets. The first one featured prosodic phenomena encountered in a museum guided tour and thus it is thought to be a "limited domain" corpus. This was developed by the Speech Group of the University of Athens within the M-PIRO project [17]. The second one has been derived from a generic and balanced textual environment. That one was created by the Artificial Intelligence Group of the University of Patras. Both speech corpora have been recorded using professional actors in the Athenian dialect. Segmentation was conducted automatically using Greek models for the HTK tool. The hand-annotations made by experienced linguists incorporate the full ToBI specification and were further cross-checked for their consistency.

3.1. Corpora description

A museum guided tour has been captured in the first database. The description of the museum's exhibits as well as indications concerning new or given information plus other enriched linguistic meta-information consist this corpus. The 5.484 words are distributed in 516 utterances. Half of the corpus data contains grammatically restricted texts, while the remaining half is unrestricted texts [8]. The corpus has been recorded appropriately in order to capture a big variety of emphatic events, like for example new mentioned information to the visitor.

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 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.

3.2. Shallow Features

Textual features were incorporated in order to predict the juncture class of a PPB. Apart from Part-of-Speech, researchers have raised the important role of syntactic and morphological information for several languages. Taking into account that in real-time PPB prediction tasks, fully syntactic parsing would be time-consuming and would produce many syntactic trees, as well as that in several languages, including Greek, syntactic tools are not freely available, a syntactic feature labeling each word with the syntactical chunk which belongs in a sentence [10] was introduced for our task. 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 Greek. 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.

3.3. Task and Feature Definition

Our main task was the prediction of the whole PPB 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 adjacent 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.

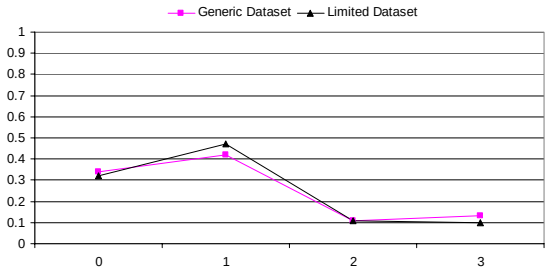

Figure 1: Prosodic phrase breaks distribution in corpora

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 PPB 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.

For the scope of evaluating the PPB 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.

4. Evaluation

Concerning the performance estimation of the PPB models, we calculated the f-measure per each PPB class, total accuracy, kappa statistic [18], mean average error (MAE) and root mean square error (RMSE). Results were obtained using the 10-fold cross validation method [15].

The f-measure is the harmonic mean of precision and recall, calculated as:

$$F = 1 \Big/ \left(\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R} \right) \qquad (1)$$

where α is a factor determining the weighting of precision and recall. Per class precision is defined as the number of correctly identified instances of a class, divided by the number of correctly identified instances, plus the number of wrongly selected cases for that class. Per class recall is estimated as the number of correctly identified instances of a class, divided by the number of correctly identified instances plus the number of cases the system failed to classify for that class. For the current evaluation we chose α=0,5 for equal weighting of precision and recall.

Figure 2 represents error metrics for CART and C4.5 approaches derived from both corpora. 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.

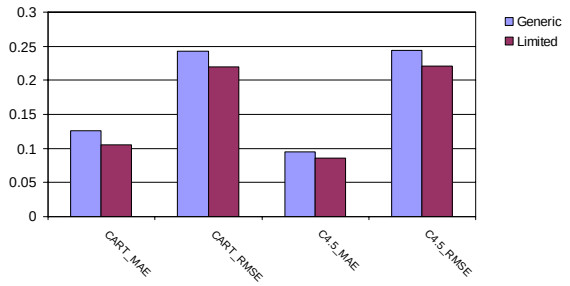

Figure 2: Mean Average Error (MAE) and Root Mean Square Error (RMSE) for both models.

The next step of our exertion was the evaluation of the models derived from the museum (limited) domain dataset. F-measure of those models is illustrated in Figure 3. For this domain, both approaches performed reasonably well with no significantly differences. The f-measure score for *non-breaks* is more than 91% while class *b2* achieved the lowest prediction (f-measure = 0.7). Though class *b3* has only a few instances (9%), its prediction is quite high due to its close relation to the punctuation marks and its low correlation with the other classes.

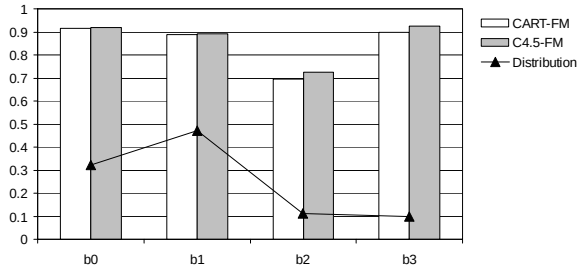

Figure 3: F-measure for museum (limited) domain models.

In Figure 4 the f-measure for each break class is depicted for the generic dataset. For these models, classification for the prosodic phrase break cases with the highest occurrence in the dataset along with class b3, performed better, which again showed low correlation with the other classes.

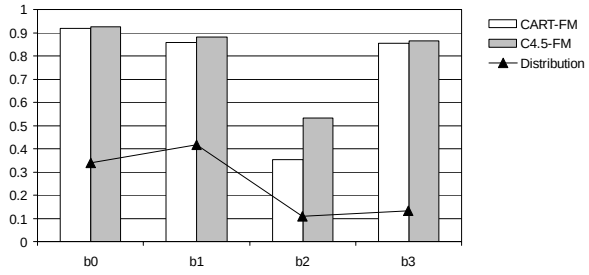

Figure 4: F-measure for generic domain models.

It is interesting to point out here that C4.5 performed better than CART especially in the prediction of the low-frequency *b2* category. In contrast to the limited dataset, *b2* prediction is significantly lower. We believe that this has been caused by the strict syntax that exhibits' description was formed (Object-Verb-Subject in 59% of the utterances). On the other hand, *non-breaks* were predicted with an f-measure higher than 0.9 for both methods.

The total accuracy of all models is tabulated in Table 1. This shows that the restricted domain achieves 4,7% higher accuracy than the unrestricted one in the case of the CART framework, while this difference is reduced to 3% upon the application of C4.5. In order to measure the statistical correlation between the predicted and actual values we also derived the kappa statistic metric, which proved to be higher than 0.75 in all cases, generally regarded as a good statistic correlation.

Table 1: Total accuracy (A) and Kappa statistic (K) of the prosodic phrase breaks models.

Methods	**Generic**		**Museum**	
	A	K	A	K
CART	83.8%	75.6%	87.7%	81.1%
C4.5	86.0%	79.2%	88.6%	82.6%

5. Conclusions

In order to evaluate the acquired knowledge from domain restrictions in the task of prosodic phrase break prediction, we utilized CART and C4.5 decision trees trained on top of shallow textual feature sequences from a limited and a generic domain corpus. As expected, museum domain models gave higher prediction scores for all PPB classes as breaks are described by simpler "rules" due to the restrictions of the domain. The prediction of the *b2* class proved to be the most difficult to predict in the generic case. However, the introduction of the C4.5 algorithm though did not showed any significantly improvement in other cases, it increased *b2* prediction by 50%. Also, C4.5 seemed to decrease the performance difference between limited and generic models from 4,7% to 3%.

6. References

[1] Bolinger, D., "Intonation and its Uses: Melody in Grammar and Discourse", London, UK, Edward Arnold, 1989.

[2] Anderson, M., Pierrehumbert, J., and 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, pp. 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. and Kouroupetroglou G, "Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora", in IEICE Transactions of Information and Systems, 2005 (to appear).

[17] Calder, J., Melengoglou, A. C., Callaway, C., Not, E., Pianesi, F., Androutsopoulos, I., Spyropoulos, C., Xydas, G., Kouroupetroglou, G., and Roussou, M., "Multilingual Personalised Information Objects", In Multimodal Intelligent Information Presentation, "Text, Speech and Language Technology" Series, Oliviero Stock and Massimo Zancanaro (editors), Springer, vol. 27, pp. 177-201, 2005.

[18] Carletta, J., "Assessing Agreement on Classification Tasks: The Kappa Statistic", Computational Linguistics, vol. 22, no. 2, pp. 249-254, 1996.