Phrase Boundary Assignment from Text in Multiple Domains

Andrew Rosenberg[1]*, Raul Fernandez*[2]*, Bhuvana Ramabhadran*[2]

[1]Computer Science Department, Queens College (CUNY), New York, USA
[2]IBM TJ Watson Research Lab, Yorktown Heights, New York, USA
andrew@cs.qc.cuny.edu, fernanra@us.ibm.com, bhuvana@us.ibm.com

Abstract

Detecting and modeling proper phrasing from an input text string is an important aspect when producing synthesis that sounds intelligible and natural. Knowledge of proper phrase structure influences, e.g., the placement and length of pauses, and the realization of phrase-final boundary contours, both of which can have an effect in a listener's percepts ranging from naturalness to semantic interpretation. In this work, we look at modeling the occurrence, and types, of phrase breaks from purely textual features, paying close attention to how the performance of the systems generalizes in- and out-of-domain for corpora of various types (such as broadcast news, spontaneous speech, and synthesis databases), and as a function of various subsets of syntactical and lexical features investigated.

Index Terms: Prosody Modeling, Prosodic Assignment, Speech Synthesis

1. Introduction

Intonational phrasing is used in speech communication to break utterances into meaningful subunits that convey syntactic, semantic and other organizational functions [1]. In particular, differences in intonation at the end of a phrase can be used to signal phenomena such as declarative statements, questions, and turn-taking behavior [2]. There has been a substantive amount of research that investigates the automatic labeling of intonational phrase (IP) boundaries, relying on acoustic and/or linguistic properties of the input [3, 4, 5, 6]. In this work we limit ourselves to investigating the problem of phrase boundary assignment from textual features alone because we are interested in the implications of this line of work for generating phrasing within text-to-speech (TTS) systems, a typical scenario where the restriction to text-only features applies. In particular, we propose a series of experiments in applying models trained on material from one domain and applying them to another. This work seeks to address a question that we believe remains unexplored in the literature: how generalizable the prediction models trained for a specific purpose are, and what degree of performance can be expected when applying them to unexpected, possibly out-of-domain, synthesis tasks. This task seems particularly relevant, as there has been growing interest in extending TTS systems to generate synthesis across different domains, with varying qualifications of spontaneity or expressiveness.

The ToBI standard [7] describes IP boundaries with a break index of '4'. Each IP boundary is also an intermediate phrase boundary, and therefore, the prosodic behavior at IP boundaries is marked by both phrase accents and boundary tones. Rather than cleave the influence of these two tones, we instead choose to model phrase accent/boundary tone (PABT) pairs. The ToBI standard for American English includes five valid PABT pairs: L-L%, L-H%, H-L%, !H-L%, H-H%. We describe experiments toward 1) the detection of IP boundaries, and 2) a six-way classification task, where each word boundary is classified as one of five PABT pairs or NOBOUNDARY. We lay out our experimental approach in Section 2. We present and discuss the results in Section 3. In Section 4, we conclude and discuss future work.

2. Approach

Corpus	L-L%	L-H%	H-L%	!H-L%	H-H%
BDC	0.49	0.36	0.10	0.04	0.10
BURNC	0.61	0.35	0.03	0.00	0.00
Games	0.35	0.15	0.30	0.04	0.16
TTS-F	0.58	0.29	0.09	0.02	0.02
SWB	0.43	0.26	0.26	0.01	0.04

Table 1: *Distribution of PABT pairs in each corpus*

2.1. Corpora

We experiment with five corpora representing a diverse set of domains, distinct speaking styles, and prosodically labeled by a distinct set of annotators using the ToBI standard. A summary of the distribution of PABT pairs can be found in Table 1.

Boston Directions Corpus: The Boston Directions Corpus (BDC) [8] is made up of elicited monologues by four non-professional speakers. The speakers were asked to perform a series of direction-giving tasks. Their spontaneous speech was transcribed, and disfluencies removed. Two weeks later, the subjects read the transcripts of their own material. The full BDC contains spontaneous and read speech. To avoid any bias introduced by the repetition of lexical content, we use only the read portion of the BDC in this work. The BDC-read material includes 10,831 words and an average IP length of 7.74 words.

Boston University Radio News Corpus: The Boston Radio News Corpus (BURNC) [9] includes prosodically annotated speech produced by six professional broadcast news speakers. Each speaker read the same set of stories in a radio news style. As with the BDC, we attempt to limit any bias introduced by exact repetition of lexical content and use only the material from one speaker, f2b. The f2b section of the BURNC contains the most transcribed and prosodically annotated material with 12,608 words and an average IP length of 4.50 words.

Columbia Games Corpus: The Columbia Games Corpus (CGC) [2] is a collection of 12 spontaneous dyadic conversations between native speakers of US English. In each session, two subjects played computer games requiring verbal communication to identify or move images on a screen. Here we use spoken material from the OBJECTS game, in which both players were presented with a screen containing a set of icons. On one player's screen, one object was blinking. The other player's task

was to move the object on their screen to the location it appeared on the describing player's screen. The conversations have been orthographically transcribed without punctuation. In order to enable parsing and the extraction of punctuation features, we introduce pseudo-sentence boundaries based on pausing information whenever an inter-word silence is greater than 500ms. This material includes speech from 13 speakers, and contains 35,236 words, with an average IP length of 3.86 words.

TTS-F: The TTS-F data is a proprietary set containing speech from a single, professional female speaker of US English, carefully recorded in laboratory conditions as training data for TTS systems. The material reflects a variety of topics of interest (e.g., weather, navigation, etc.) as well as a subcorpus optimized for triphone phonetic coverage. The TTS-F corpus contains 60,061 words with an average IP length of 3.54 words,

Switchboard The Switchboard Corpus (SWB) [10] is a collection of two-party spontaneous dialogs recorded from telephone speech. A subset of the corpus has been prosodically annotated using the NXT-format [11]. The portion of the corpus we used includes 75 conversations comprising 94,578 words. The average IP length is 7.48 words. We note that in the NXT annotations, approximately 66.4% of all IP boundaries do not have annotated phrase accents and boundary tones. The distribution that appears in Table 1 is based on only annotated tokens. In the binary task, we consider these unlabeled boundaries to be positive instances. In the N-way classification use an UNK label during the modeling in order to ensure that the sequential modeling isn't biased by many missing phrase boundaries. However, during the evaluation, we omit these tokens from calculation of average recall and accuracy to avoid systematic increase or decrease of performance.

2.2. Features

The feature set contains predictors that attempt to summarize syntactical structure and lexical constituency of the input string. We use the Stanford Parser [12] to tag the sentences with POS labels and to obtain a parse tree. The POS tag is used as a feature directly, and the parse tree is used to extract the following three syntactic features roughly measuring the syntactic distance between pairs of adjacent words, as follows: Let N be the deepest node in the parse tree that dominates a pair of adjacent words (terminals) w_k and w_{k+1}, and let d_k and d_{k+1} be their respective distances to N (i.e., number of intermediate nodes between N and the terminals). From these we define the following syntax-tree features: $s1_k = \min(d_k, d_{k+1})$, $s2_k = \max(d_k, d_{k+1})$, and $s3_k = d_k + d_{k+1}$.

The degree of coupling between words is also independently measured using bi-gram forward and reverse language models (LMs) to evaluate, respectively, the following features that have already been explored in the context of pitch-accent labeling [13]: $lm1_k = p(w_k|w_{k-1})$ and $lm2_k = p(w_k|w_{k+1})$. These LMs are trained from 8 different corpora (independent of those already described) containing approximately 86,000 word types and also reflecting various styles (broadcast news, transcribed conversational speech, closed-captions, etc.). Bi-gram models are first estimated with Kneser-Ney smoothing for each of the 8 sub-corpora, and one final (forward and backward) LM built by interpolating the 8 LMs, with weights chosen to minimize perplexity on the full LM training set.

We also propose a word-level score related to the likelihood that a lexical token is the last word in an IP. This *Ratio* feature was first proposed by [14] as an *Accent Ratio* feature to model pitch accentability, and generalized to other prosodic phenomena in [13]. *Ratio* is a memory-based feature that reflects prior knowledge about how likely a word is to be associated with a particular prosodic phenomenon, by collecting statistics from a labeled corpus and retaining this corpus-based prior whenever it is statistically significant, and is given by:

$$Ratio(e,w) = \begin{cases} \frac{k_e}{n_w} & \text{if } B(k_e, n_w; p(e)) \leq 0.05 \\ p(e) & \text{otherwise,} \end{cases} \quad (1)$$

where n_w is the number of times a word w appears in the corpus, k_e the number of times that it is associated with a prosodic event e, $p(e)$ is the rate at which the event occurs in the corpus, and $B(k_e, n_w; p(e))$ is a binomial distribution with parameter $p = p(e)$. This feature equals the fraction of "successes" whenever there is sufficient evidence in a corpus to establish how likely a word co-occurs with an event, as diagnosed by a binomial distribution. When not, the feature reflects uncertainty and is set to $p(e)$. We calculate $Ratio$ for the presence and absence of an IP boundary as well as for PABT types.

We also include as a feature the type of punctuation, if any, that follows a word based on the following inventory: comma, colon, double quote, exclamation point, semicolon, period, question mark and no-punctuation. Note that hyphens, parentheses and brackets were not included in the classification of punctuation. In the spontaneous corpora, Games, and SWB, these were frequently used to indicate disfluencies in the transcripts rather than punctuation. Finally, we also consider 5 broad lexical class membership features via binary flags that indicate whether a given word is a member of the following classes: 1) auxiliary verbs, 2) conjunctions, 3) function words, 4) wh-words, 5) adpositions.

2.3. Modeling

A Conditional Random Field (CRF) [15] is an undirected graph with nodes $\mathbf{x}$ and $\mathbf{y}$, corresponding, respectively, to an observation sequence and a sequence to be inferred, which directly encodes the conditional distribution $p(\mathbf{y}|\mathbf{x})$ using a log-linear model. A linear-chain CRF, in particular, is a structure that assumes that, conditioned on $\mathbf{x}$, the dependencies of the elements of $\mathbf{y}$ form a Markov chain. In this paper, we make use of this assumption and restrict ourselves to first-order chains with a the conditional distribution of the form:

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(y_t, y_{t-1}, \mathbf{x})\right)}{\sum_y \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(y_t, y_{t-1}, \mathbf{x})\right)} \quad (2)$$

where each $f_k(y_t, y_t - 1, \mathbf{x})$ is the binary feature:

$$f_k(y_t, y_{t-1}, \mathbf{x}) = \delta(y_t = i)\delta(y_{t-1} = j)att^l(\mathbf{x}), \quad (3)$$

and $att^l(\mathbf{x})$ stands for a boolean function that signals when some property l of the input sequence is true. For implementation, we have made use of the CRF++ toolkit to estimate the parameters λ_k of the model and decode the input sequences [16]. We consider 6 different types of templates to generate the f_k indicators of Eq. 3 based on evaluating the following single and pairwise attributes around the token at time t: (i) $att(x_{t-1})$, (ii) $att(x_t)$, (iii) $att(x_{t+1})$, (iv) $att(x_{t-1}) \wedge att(x_t)$, (v) $att(x_t) \wedge att(x_{t+1})$, and (vi) $att(x_{t-1}) \wedge att(x_{t+1})$. The $att(\cdot)$ function is evaluated for each of the basic raw measures already described, and therefore generates a number of features f_k equal to the cardinality of the symbol table of its arguments (or, in the case of the pairwise attributes, their product). This can lead to a

Corpus	All		Ratio		LM		Punc		Parse		Word Class	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
BDC	93.1	.705	88.5	.471	86.9	.241	91.1	.518	92.8	.682	89.5	.492
BURNC	90.5	.781	79.2	.352	79.6	.532	87.2	.606	89.3	.750	81.2	.555
Games	89.4	.766	76.6	.264	75.7	.403	88.3	.714	88.8	.737	79.3	.435
TTS-F	89.1	.795	84.8	.679	81.7	.658	79.4	.431	86.6	.751	80.9	.654
SWB	85.7	.427	82.1	.064	81.8	.000	85.7	.425	85.3	.396	82.0	.041

Table 2: *Within-Corpus Binary Phrase Prediction Feature Analysis. Results are reported using Accuracy and F_1-measure*

Corpus	All		Ratio		LM		Punc		Parse		Word Class	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
BDC	93.2	.734	88.1	.205	86.2	.316	92.2	.573	92.3	.711	89.1	.458
BURNC	88.9	.706	79.5	.291	80.7	.356	87.2	.609	87.5	.660	81.2	.346
Games	82.9	.558	74.2	.058	75.2	.249	83.7	.556	83.1	.581	75.7	.219
TTS-F	80.7	.491	78.8	.405	78.3	.402	78.2	.384	80.2	.471	79.9	.471
SWB	84.5	.520	83.5	.328	81.7	.319	88.2	.458	83.7	.501	84.4	.310

Table 3: *Across-Corpus Binary Phrase Prediction Feature Analysis. Results are reported using F_1-measure*

large number of features with few tokens observed in the training set, so to avoid overfitting, any generated feature with fewer than 10 occurrences is dropped from consideration at training time. Overfitting is also controlled by a regularization parameter $C = 0.01$ on the size of the parameter vector.

Since implicit in the above formulation is the fact that all variables are discrete, we need to quantize the LM and *Ratio* features. Inspection of the generated LM features revealed a fairly log-normal distribution, which motivated using z-score normalization over the log-likelihood values. This transformation generates a zero mean and unit standard deviation distribution, from which we quantize the following z-score intervals: $z \leq -1 \rightarrow$ VERYLOW; $-1 < z \leq 0 \rightarrow$ LOW; $0 < z \leq 1 \rightarrow$ HIGH and $z > 1 \rightarrow$ VERYHIGH. Due to the log-normal distribution this yields approximately one third of the data in the LOW and HIGH bins, and one sixth in each of the VERYLOW and VERYHIGH bins. We note that this normalization does not require any event labels. We find the distribution of *Ratio* feature values to be remarkably multi-modal, possibly due to the fact that many words fail the binomial test in Eq. 1 indicating that there is not sufficient evidence to trust the value is significantly different from the event prior. Because of this, rather than using a parametric quantization approach, we divide the *Ratio* scores into four bins using the following thresholds [0, 0.005, 0.05, 0.5, 1]. We find each of the first two bins contain approximately one third of the data while the other two contain approximately one sixth of the data each.

In Sec. 3, we report the results of two types of experiments. We explore the performance of CRF labeling in a within-corpus setting, where we use half of the material from each corpus for training and evaluate on the remaining half. We also investigate the potential of applying a model trained on material from one domain, to material in a different domain. Here we use a cross-corpus evaluation technique, where the training set uses the data from four corpora and is evaluated on the fifth. Since the *Ratio* scores are derived from the labels of the training data, particular care is taken to ensure these features never use information from any test fold. Therefore, in the within-corpus experiments, we use *Ratio* features that have been trained using the four *other* corpora to generate features for the training and test sets. In each across-corpus experiment, one corpus is used as the evaluation data. Therefore no *Ratio* feature is derived from the test corpus. The *Ratio* features used in the test data are based on the four training corpora while those extracted for each corpus in the training data are calculated using the corpus itself.

3. Results and Discussion

In this section, we describe the results of four experiments. First, we present the results of binary classification experiments where the CRF is predicting the presence of a phrase boundary following the current word. The results of the within-corpus evaluation are reported in Table 2, while the cross-corpus results are reported in Table 3. These tables also include the results of classification experiments using a single feature set, one of i) Ratio features (6 features), ii) lexical coherence features (2 features), iii) punctuation (1 feature), iv) parsing derived features (4 features) and v) word classes (5 features).

We find that the within-corpus performance on BDC, BURNC, Games and TTS-F are approximately equivalent. The detection F-measure falls between 72.4 and 79.2 and the accuracy is between 89% and 93.5%. However, the Switchboard results are notably lower with the highest F-measure of 41.7% using only the punctuation feature. It is unclear if the poor within-corpus detection performance on Switchboard is due to the conversational style, speaker idiosyncrasies or inconsistent annotation of phrase boundaries and punctuation. We find, however, that when we use a model trained on the remaining corpora, that performance increases. This suggests that there are significant conversation-to-conversation differences in SWB, and that with a diverse set of training data these can be accommodated.

Examining the relative strengths of the feature sets, we find that punctuation is one of the stronger indicators of prosodic phrasing. This is unsurprising; sentence boundaries are natural places to introduce IPs. We also find that in many instances this information can be fruitfully incorporated into the parse tree derived features. Both within- and across-corpus, the parse features perform nearly as well or better than the punctuation features. The exception to this is the cross-corpus evaluation on the BURNC material. It is likely that the news data that makes up the BURNC has a more complicated syntactic structure than the spontaneously derived material of the BDC, Games and SWB corpora and the TTS source material of the TTS-F data.

In the cross-corpus paradigm, performance drops on all corpora except SWB. The reduction on BDC and BURNC is notably smaller than the reduction on TTS-F and Games, suggesting that news and monologues are more universally modeled by other domains, while the interactive dialog of Games and the TTS material are less generalizable. We find that punctuation features are fairly robust to domain changes while the other categories are particularly sensitive to domain differences. This is

Corpus	All		Ratio		LM		Punc		Parse		Word Class	
	Acc	AR	Acc	AR	Acc	AR	Acc	AR	Acc	AR	Acc	AR
BDC	90.2	.319	87.0	.196	86.6	.180	89.7	.277	90.0	.302	87.0	.223
BURNC	86.2	.307	78.9	.192	78.6	.216	85.5	.261	85.6	.300	80.2	.226
Games	78.6	.287	75.3	.174	75.3	.167	78.2	.259	78.6	.270	75.3	.167
TTS-F	81.7	.332	79.5	.258	77.4	.241	78.5	.281	80.5	.293	79.0	.245
SWB	91.0	.167	91.0	.167	91.0	.167	91.0	.167	91.0	.167	91.0	.167

Table 4: *Within-Corpus N-way Phrase Prediction Feature Analysis. Results are reported using Accuracy and Average Recall*

Corpus	All		Ratio		LM		Punc		Parse		Word Class	
	Acc	AR	Acc	AR	Acc	AR	Acc	AR	Acc	AR	Acc	AR
BDC	88.3	.218	87.5	.184	87.4	.176	87.4	.176	89.0	.248	87.4	.176
BURNC	82.6	.293	79.1	.184	79.2	.182	81.8	.275	81.9	.224	79.2	.182
Games	75.3	.201	74.6	.170	74.9	.168	74.9	.167	75.2	.178	74.7	.168
TTS-F	76.8	.214	77.4	.219	76.3	.208	76.5	.210	77.3	.220	76.9	.214
SWB	82.8	.228	89.0	.185	93.8	.173	88.0	.199	83.3	.217	91.6	.184

Table 5: *Across-Corpus N-way Phrase Prediction Feature Analysis. Results are reported using Accuracy and Average Recall*

expected with the Ratio features, as they model specific lexical tokens; it is less predictable that the disjunction measures (LM and Parse based) and the word class features are sensitive to domain difference. This suggests that differences in performance style and the domain lead to significant differences in the lexical and syntactic influences on phrasing. We evaluate the 6-way phrase ending classification using Average Recall, calculated as $AR = \frac{1}{K}\sum_{k=1}^{K}\frac{c_k}{N_k}$ where K is the number of classes, c_k is the number of correctly identified instances of class k, and N_k is the number of instances of class k. This measure equals 1 when every point is classified correctly, and equals $\frac{1}{K}$, here $1/6$ = .167, under a majority class or random baseline. This measure is also called Balanced Error Rate [17]. The 6-way classification results are fairly low, with much of the results arising from correctly identifying L-L% tokens. We find that the relative importance of the feature sets are consistent between the binary and 6-way classification task. One notable difference is that the 6-way classification across-corpus performance is worse on the Games corpus than any other corpus. This can be explained by the difference in PABT distribution reported in Table 1. The dialog in the Games corpus shows a much more varied use of phrase ending intonation than the other four corpora.

4. Conclusions and Future Work

In this paper, we report experiments on IP boundary detection and PABT classification using lexical/syntactic features. We pay particular attention to the task of training a prosodic assignment model on material from one domain and speaking style and applying it to another. This work addresses the question of how generalizable the relationship between lexical and syntactic qualities and phrasing behavior is. We find that while phrases can be reliably detected in a within-corpus setting, performance significantly drops in a cross-corpus evaluation. This indicates that there are significant differences with how syntax and phrasing interact based on domain, speaking style, and labelers. Performance in the prediction of phrase ending intonation is similarly sensitive to domain and speaking style differences. While within-corpus performance is modest, the cross-corpus performance is, in some cases, only slightly over baseline. In the future, we will investigate more robust correlates of phrasing and phrase-ending intonation and investigate applying domain transfer techniques to this task.

5. References

[1] D. Bolinger, *Intonation and Its Uses: Melody in Grammar and Discourse*. Stanford University Press, 1989.

[2] A. Gravano, "Turn taking and affirmative cue words in task-oriented dialog," Ph.D. dissertation, Columbia University, 2009.

[3] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, 1996.

[4] M. Q. Wang and J. Hirschberg, "Predicting intonational boundaries automatically from text: the ATIS domain," in *HLT '91: Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA: ACL, 1991, pp. 378–383.

[5] J. Hirschberg and O. Rambow, "Learning prosodic features using a tree representation," in *Eurospeech*, 2001.

[6] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.

[7] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody," in *Proc. of the 1992 ICSLP*, vol. 2, 1992, pp. 12–16.

[8] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[9] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Boston University, Tech. Rep. ECS-95-001, March 1995.

[10] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. ICASSP-92*, vol. 1, pp. 517–520 vol.1, Mar 1992.

[11] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Lang. Resour. Eval.*, vol. 44, no. 4, pp. 387–419, Dec. 2010.

[12] "The Stanford Parser: A statistical parser," http://nlp.stanford.edu/software/lex-parser.shtml.

[13] R. Fernandez and B. Ramabhadran, "Driscriminative training and unsupervised adaptation for labeling prosodic events with limited training data," in *Interspeech*, 2010, pp. 1429–1432.

[14] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, and D. Jurafsky, "To memorize or to predict: Prominence labeling in conversational speech," in *Proc. HLT-ACL*, Rochester, NY, April 2007, pp. 9–16.

[15] C. Sutton and A. McCallum, *Introduction to Statistical Relational Learning*. MIT Press, 2006, ch. An Introduction to Conditional Random Fields for Relational Learning.

[16] T. Kudo. (2009) CRF++: Yet another CRF toolkit. [Online]. Available: http://code.google.com/p/crfpp/

[17] I. Read and S. Cox, "Automatic pitch accent prediction for text-to-speech synthesis," in *Interspeech*, 2007, pp. 482–485.