THE (NON)UTILITY OF LINGUISTIC FEATURES FOR PREDICTING PROMINENCE IN SPONTANEOUS SPEECH

Jason M. Brenier, Ani Nenkova, Anubha Kothari, Laura Whitton, David Beaver, Dan Jurafsky

Department of Linguistics, Stanford University

ABSTRACT

Conversational speech is characterized by prosodic variability which makes pitch accent prediction for this genre especially difficult. The linguistic literature points out that complex features such as information status, contrast and animacy help predict pitch accent placement. In this paper, we use a corpus annotated for such features to determine if they improve prominence prediction over traditional shallow features such as frequency and part-of-speech, or over new ones that we introduce. We demonstrate that while correlated with prominence, complex linguistic features do not improve prediction accuracy. Furthermore, the performance of our classifier is quite close to the ceiling defined by variability in human accent placement. An oracle experiment demonstrates, though, that at least some accuracy improvement is still possible.

Index Terms— prosody, prominence, givenness, contrast, animacy

1. INTRODUCTION

Predicting prosodic characteristics of conversational speech like the *prominence* or *pitch accent* status of a word is important for text-to-speech in dialog systems. Early investigations of automatic prediction used robust features such as a word's part-of-speech (POS) [1], or its unigram and bigram probability [2, 3]. The success of these predictors is surprising given that linguistic theories explain prominence very differently, in terms of whether a word is given, used contrastively, or in focus. At the same time, results for prominence prediction using new statistical techniques and shallow features have hovered around 76% for conversational speech and 85% for read speech. Several questions have remained unanswered: Will more complex linguistic features improve accuracy? Will any new features at all help improve results, or have we achieved a performance ceiling due to variability in prominence? How much do more elaborate features improve upon the classic POS and predictability measures?

In this paper, we address the above questions, using a richly annotated corpus manually labeled for information status (IS), contrast and animacy. We also introduce new robust features: *lead value*, *verb specificity*, and *accent ratio* and examine the performance ceiling resulting from variability.

Thanks to M. Ostendorf, J. Carletta, S. Calhoun, M. Steedman, D. R. Ladd, N. Mayo, M. Nissim, for providing annotations, to Edinburgh-Stanford Link and ONR (MURI award N000140510388) for generous support.

2. DATA AND FEATURES

Our experiments use 12 Switchboard conversations, 14,555 tokens in total. Each word was manually labeled for presence or absence of pitch accent [4]. The data was also annotated for information status, contrast and concrete/non-concrete distinctions, features that linguistic literature suggests are predictive of prominence [5, 6].

INFORMATION STATUS Details of the IS (or givenness) annotation can be found in [7]. First mentions of entities were marked as *new* and subsequent mentions as *old*. Entities that were not previously mentioned, but that were generally known or semantically related to other entities in the preceding context were marked as *med*iated.

CONTRAST Contrast captures the reason an item has been highlighted relative to other items of the same type. Several possible reasons were marked: *correction* for when the speaker intends to correct or clarify a previous word or phrase; *contrastive* when the word is directly differentiated from a previous topical or semantically-related word; *subset* when it refers to a member of a more general set mentioned in the surrounding context; *adverbial* when a focus-sensitive adverb such as "only" or "even" is associated with the word being annotated; *answer* when the word completes a question by the other speaker; *background* when the word is not intended to be salient and *nonapplic* for filler phrases such as "in fact", "I mean", etc. A complete description of the annotation guidelines can be found in [8].

ANIMACY Each noun and pronoun is labeled for the animacy of its referent [9]. The categories include *concrete*, *nonconcrete*, *human*, *org*anizations, *place*, and *time*.

In addition to the above theoretically motivated features, we used several automatically derivable word measures.

PART-OF-SPEECH Two such features were used, the full Penn Treebank tagset (POS), and a collapsed tagset (BroadPOS) with six broad categories (nouns, verbs, function words, pronouns, adjectives and adverbs).

UNIGRAM AND BIGRAM PROBABILITY These features are defined as $log(p_w)$ and $log(p_{w_i}|p_{w_{i-1}})$, respectively, and their values were calculated from the Fisher corpus. High probability words are less likely to be prominent.

TF.IDF This measure captures how central a word is for a

particular conversation. It is a function of the frequency of occurrence of the word in the conversation (n_w), the number of conversations that contain the word in a background corpus (k) and the number of all conversations in the background corpus (N). Formally, *TF.IDF1* $= n_w \times log(\frac{N}{k})$. We also used a variant, *TF.IDF2*, computed by normalizing *TF.IDF1* by the number of occurrences of the most frequent word in the conversation. Words with high *TF.IDF* values are important in the conversation and are more likely to be prominent.

STOPWORD This is a binary feature indicating if the word appears in a high-frequency stopword list from the Bow toolkit [10]. The list spans both function and content word classes, though numerals and some nouns and verbs were removed.

ACCENT RATIO This feature measures a word's preference for a prominence class. Originally, it was an estimate of how likely it is for a word to be accented in a corpus, without taking statistical significance into account [11]. Our version of this feature incorporates the significance of the preference, assuming a default value of 0.5 for those words for which there is insufficient evidence of preference. More specifically,

$$AccentRatio(w) = \begin{cases} \frac{k}{n} & \texttt{if } B(k,n,0.5) \leq 0.05 \\ 0.5 & \texttt{otherwise} \end{cases}$$

where k is the number of times word w appeared accented in the corpus, n is the total number of times the word w appeared, and $B(k, n, 0.5)$ is the probability (under a binomial distribution) that k successes occur out of n trials if the probability of success and failure is equal. Simply put, the accent ratio of a word is equal to the estimated probability of the word being accented if this probability is significantly different from 0.5, and equal to 0.5 otherwise. For example, *AccentRatio(you)=0.3407*, *AccentRatio(education)=0.8666*, and *AccentRatio(probably)=0.5*. A larger subcorpus of 50 Switchboard conversations, annotated for pitch accent, was used to compute k and n for each word. We also have a categorical variant of the feature, AR_CAT with values *acc* if *AccentRatio* $<$ *0.3*, *notacc* if *AccentRatio* $>$ *0.7* and *unk* else.

Finally, two features were borrowed from text summarization [12], where they are used to identify globally important words. Such words are likely to be prominent, so we assessed the usefulness of these features for pitch accent prediction. We directly used Schiffman *et al.*'s dictionaries.

LEAD WORD VALUE In news, the most important information occurs in the first paragraph of an article. We used a large dictionary containing the ratio between the number of times a word occurred in the first paragraph of an article and the number of times the word occurred elsewhere in an article.

VERB SPECIFICITY Some verbs are very informative and tend to appear with only a small set of subject arguments. For example, the verb *arrest* suggests a police activity, while verbs such as *like, have, is* are light in content. The verb specificity measure indicates the degree of mutual information between a verb and a given nominal subject.

	med	**new**	**old**
NN	1189	420	255
PRO	64	0	1495

Table 1. Information status for nouns and pronouns

3. FEATURE SELECTION

Table 2 shows how powerful each feature is for predicting pitch accent. The best feature is accent ratio, followed by unigram and POS. Contrast, information status, animacy, verb specificity and lead word value, while still carrying some predictive information, are far less powerful predictors, probably because their values are not defined for *all* words. Leave-one-out feature selection was performed to identify the best subset of features. Four features were selected: accent ratio (both real and categorical), stopword, and contrast.

In the next section, we discuss classifier performance. Here, we analyze the relationship between the three complex linguistic features and accent. The relationship is significant in all three cases; table 3 shows the actual distribution of the accent classes for each category.

Information status and accent. It has been assumed that *new* information is generally accented, while *old* information is not. But as we saw in table 2, IS has a relatively low information gain compared to other features. Tables 1 and 3, which show the distribution between IS categories and POS and accent classes, shed some light on why this might be the case. First, IS annotations apply only to nouns and pronouns. Because the encoded items comprise only a small fraction of the dataset, IS may be insufficient for improving the overall classifier. Moreover, about half of the nouns in our corpus are assigned *med*iated information status, which is an intermediary category for which existing theories of accent placement make no claims. Second, IS also plays a role in determining the appropriate form of reference to an entity: full indefinite/definite noun phrase, or pronoun. Indeed, 85% of the *old* entities were realized as pronouns. This form of reference distinction masks the link between information status and accent, because pronouns are so frequently unaccented.

Even when constrained exclusively to nouns (table 3), the relationship between IS and accent is significant (chi-square test p-value is 0.0005). However, it is only the relationship between *new* and accent that is significant.[1] *New* nouns are accented more often than the average for the class, but since most nouns are accented anyway, features that indicate when a noun can be unaccented are more critical for the classifier. This pattern probably contributes to the low impact of IS for predicting prominence.

Contrast and accent. Contrast was the only linguistic feature selected during feature selection. Unlike information status, this feature has good applicability, with more than

[1] The *med* and *old* classes are accented at about the same rate as the overall average for the noun class, which is 65%.

FEATURE	AccentRatio	Unigram	AR_CAT	POS	TF.IDF2	TF.IDF1	BroadPOS
INFOGAIN	0.2420	0.20235	0.17352	0.14241	0.12028	0.11525	0.10370
FEATURE	Stopword	Contrast	Bigram	InfoStatus	Animacy	VerbSpec	LeadWordVal
INFOGAIN	0.0970	0.09183	0.07791	0.01869	0.00875	0.00509	0.0041

Table 2. Information gain for different features

INFO STATUS	med	new	old					
ACCENTED	752 (*63%*)	307 (*73%*)	156 (*61%*)					
UNACCENTED	437	113	99					
CONTRAST	adverbial	answer	background	contrastive	correction	nonapplic	other	subset
ACCENTED	106 (*70%*)	20 (*80%*)	1942 (*34%*)	531 (*72%*)	19 (*76%*)	285 (*31%*)	440 (*75%*)	448 (*72%*)
UNACCENTED	45	5	3693	208	6	639	145	177
ANIMACY	animal	concrete	human	nonconc	org	place	time	
ACCENTED	11 (*31%*)	192 (*51%*)	774 (*37%*)	713 (*41%*)	86 (*47%*)	53 (*61%*)	78 (*48%*)	
UNACCENTED	24	187	1340	1024	96	34	83	

Table 3. Accent class distribution for (noun) information status, contrast, and animacy subclasses

half of the data marked for contrast and 15% (2150 tokens) marked with non-trivial contrast types excluding *background* and *nonapplic*. Moreover, contrast applies to tokens from every broad POS. As expected, contrast items are likely to be prominent. In the entire dataset, 34% of tokens are accented, which is very close to the percentage of accented items tagged as belonging to the neutral *background* or *nonapplic* categories. In the *adverbial*, *answer*, *contrastive*, *correction*, *subset* and *other* contrast categories, the percentage of accented items is between 70% and 80%, indicating a significant preference for prominence.

Animacy and accent. The concrete vs. non-concrete distinction was highly correlated with accent. Words referring to concrete entities are much more likely to be accented than those referring to non-concrete ones, with more than half of the words marked as *concrete* bearing accent. References to humans and organizations also have a slight preference for accent. As with IS, however, the form of expression masks the animacy feature, because humans are often referred to by pronouns and these are normally unaccented.

4. CLASSIFIER PERFORMANCE AND ERRORS

We trained a decision tree classifier[2], with two main goals *(i)* to identify how much new features help improve the performance above the classic unigram and POS, and *(ii)* to see if oracle features can help improve over previously reported results, showing that a performance ceiling has not yet been reached. Table 4 shows the results for six classifiers. UNIGRAM+POS (binned) uses binned versions of unigram and POS as in [13], with unigrams binned into five equal classes and POS collapsed into four classes: nouns, verbs, function words and other (adjectives+adverbs). UNIGRAM+POS uses the same two features, but without binning. FS.CONTEXT uses the four features from feature selection (FS), plus contrast, stopword, BroadPOS and categorical accent ratio of the preceding and following words. FS.CONTEXT-ORACLE has, in addition, the real accent class for the preceding and following words. AUTOMATIC.CONTEXT extends AUTOMATIC with context that does not include contrast.

FEATURES	ACCURACY
UNIGRAM+POS (*binned*)	73.47%
UNIGRAM+POS	75.09%
ALL (*features from table 2*)	76.08%
AUTOMATIC (*all but IS, contrast, animacy*)	76.17%
FS (*AccentRatio, AR_CAT, contrast, stopword*)	76.45%
AUTOMATIC.CONTEXT	77.07%
FS.CONTEXT	77.31%
FS.CONTEXT+ORACLE	79.66%

Table 4. Ten-fold cross-validation accuracy of classifiers with different feature subsets

The hand-labeled linguistic features do not improve performance over the automatic ones, as the accuracy of the ALL and AUTOMATIC classifiers (table 4) is almost identical. The UNIGRAM+POS classifier performs very reasonably, though accuracy can be enhanced by new features (ALL and AUTOMATIC).[3] However, the improvement is only about 1%, and a classifier that does not use unigram or POS at all can have the same or better performance (FS). These facts suggest that 76% is potentially very close to the performance ceiling, and that different features capture very similar properties of words and are correlated and mutually substitutable.

The most useful new features are those capturing contextual information, leading to an additional 1% boost. The best result for non-oracle features is that of FS.CONTEXT: 77.31% accuracy and 2.22% better than UNIGRAM+POS. The oracle-enhanced classifier, FS.CONTEXT+ORACLE, has

[2]Using the J49 implementation in WEKA.

[3]Binning leads to worse results than using the original unigram and POS.

the best accuracy (79.66%), which shows that at least some improvement should still be possible. We now turn to error analysis of the ALL classifier in order to identify possible areas of improvement. Space does not allow for a complete report here, but we briefly summarize three main sources of error:

Underaccenting of low-content words, particularly in pronouns (*I, you, they*), high frequency verbs (*know, think, have, do*), and negatives (*don't, no*), such as the missed accent on *don't* in `they DON'T wanna THINK about it`.[4] The ten most frequently misclassified words account for 10% of the total error. This suggests that it would be most fruitful to study the individual words that account for the majority of these types of error.

Incorrect accent in premodified nouns, including noun-noun compounds, known to be problematic [3], as well as adjectival premodification. Such constructions contain 35% of all nouns and 43% of all noun errors. This discrepancy signals that they are a primary area for improvement. For example, both words in 'high school' are predicted to be accented in the phrase `when I was in HIGH school was the BEGINNING of VIETNAM`.

Variability The extreme difficulty of automatic accent prediction is shown by the fact that the same phrase can be accented in different ways by the same speaker, e.g. "dinner parties" below:

- `WHAT would YOU have at a dinner party?`
- `i've cut DOWN on HAVING dinner PARTIES`
- *How do you serve them?* `I think for a DINNER party I..., I don't know.`

The implications of this phenomenon are noteworthy and are discussed in detail in the next section.

5. VARIABILITY AND HUMAN PERFORMANCE

The agreement among different speakers in the assignment of accent determines an upper bound for the usefulness of accuracy as a measure of classifier performance. Since one cannot estimate the degree of variability in spontaneous speech, we derived an approximation using read speech. Part of the Boston University Radio News Corpus includes the same news text read by six people and annotated for pitch accent. There were 2,112 tokens in all. We computed the accuracy of one speaker's accent placement against those of another used as a gold-standard for all 15 possible pairings of the six speakers. The average agreement was 82%, with minimum and maximum of 79% and 85%, respectively. Since variability is likely to be even higher in spontaneous speech than in read news, these numbers suggest that our current performance of 77% accuracy is not too far below human performance. However, the results from the oracle classifier and our error analysis suggest possibilities for some improvement.

[4]Capitalization indicates actual accent.

6. DISCUSSION

Our experiments suggest that hand-labeled IS, contrast and animacy features, while correlated with prominence, do not improve accent prediction over the use of robust automatic features. Thus, developing automatic approximations of IS and contrast, while interesting tasks by themselves, are unlikely to be helpful for pitch accent prediction.

Error analysis suggests that performance can be improved by directing future efforts toward premodified noun phrases and classifiers for individual words that are frequently misclassified and account for a large percentage of overall errors.

Providing context features to the classifier, along with our new powerful accent ratio feature, leads to 1% net improvement, suggesting the use of sequence labeling techniques such as Markov models or conditional random fields. [13].

Finally, the natural variability in human speech puts a performance ceiling on accuracy measures. Current results are probably slighly below this ceiling, as indicated from the better performance of a classifier using oracle features. Thus, new forms of evaluation, such as listening experiments, or computing the number of utterances predicted with unacceptable accent patterns, might be more helpful in comparing the performance of automatic classifiers.

7. REFERENCES

[1] J. Hirschberg, "Pitch accent in context: Predicting intonational prominence from text," *Artificial Intelligence*, vol. 63, 1993.

[2] S. Pan and K. McKeown, "Word informativeness and automatic pitch accent modeling," *Proceedings of EMNLP/VLC99*, 1999.

[3] S. Pan and J. Hirschberg, "Modeling local context for pitch accent prediction," *Proceedings of ACL*, pp. 233–240, 2000.

[4] M. Ostendorf, I. Shafran, S. Shattuck-Hufnagel, L. Carmichael, and W. Byrne, "A prosodically labeled database of spontaneous speech," *Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, pp. 119–121, 2001.

[5] D.L. Bolinger, "Contrastive Accent and Contrastive Stress," *Language*, vol. 37, no. 1, pp. 83–96, 1961.

[6] W. Chafe, "Givenness, contrastiveness, definiteness, subjects, topics, and point of view," *Subject and Topic*, pp. 25–55, 1976.

[7] M. Nissim, S. Dingare, J. Carletta, and M. Steedman, "An annotation scheme for information status in dialogue," in *LREC 2004*, 2004.

[8] S. Calhoun, M. Nissim, M. Steedman, and J.M. Brenier, "A framework for annotating information structure in discourse," *Pie in the Sky: Proceedings of the workshop, ACL*, pp. 45–52, 2005.

[9] A. Zaenen, J. Carletta, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitina, M.C. O'Connor, and T. Wasow, "Animacy Encoding in English: why and how," *ACL Workshop on Discourse Annotation*, 2004.

[10] A. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," http://www.cs.cmu.edu/ mccallum/bow, 1996.

[11] J. Yuan, J.M. Brenier, and D. Jurafsky, "Pitch Accent Prediction: Effects of Genre and Speaker," *Proceedings of Interspeech*, 2005.

[12] B. Schiffman, A. Nenkova, and K. McKeown, "Experiments in multidocument summarization," in *Proceedings of HLT*, 2002.

[13] M. Gregory and Y. Altun, "Using conditional random fields to predict pitch accent in conversational speech," in *Proceedings of ACL*, 2004.