Lessons Learned in Part-of-Speech Tagging of Conversational Speech

Vladimir Eidelman[†], **Zhongqiang Huang**[†], **and Mary Harper**[†‡]
[†]Laboratory for Computational Linguistics and Information Processing
Institute for Advanced Computer Studies
University of Maryland, College Park, MD
[‡]Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD
{vlad,zhuang,mharper}@umiacs.umd.edu

Abstract

This paper examines tagging models for spontaneous English speech transcripts. We analyze the performance of state-of-the-art tagging models, either generative or discriminative, left-to-right or bidirectional, with or without latent annotations, together with the use of ToBI break indexes and several methods for segmenting the speech transcripts (i.e., conversation side, speaker turn, or human-annotated sentence). Based on these studies, we observe that: (1) bidirectional models tend to achieve better accuracy levels than left-to-right models, (2) generative models seem to perform somewhat better than discriminative models on this task, and (3) prosody improves tagging performance of models on conversation sides, but has much less impact on smaller segments. We conclude that, although the use of break indexes can indeed significantly improve performance over baseline models without them on conversation sides, tagging accuracy improves more by using smaller segments, for which the impact of the break indexes is marginal.

1 Introduction

Natural language processing technologies, such as parsing and tagging, often require reconfiguration when they are applied to challenging domains that differ significantly from newswire, e.g., blogs, twitter text (Foster, 2010), or speech. In contrast to text, conversational speech represents a significant challenge because the transcripts are not segmented into sentences. Furthermore, the transcripts are often disfluent and lack punctuation and case information. On the other hand, speech provides additional information, beyond simply the sequence of words, which could be exploited to more accurately assign each word in the transcript a part-of-speech (POS) tag. One potentially beneficial type of information is prosody (Cutler et al., 1997).

Prosody provides cues for lexical disambiguation, sentence segmentation and classification, phrase structure and attachment, discourse structure, speaker affect, etc. Prosody has been found to play an important role in speech synthesis systems (Batliner et al., 2001; Taylor and Black, 1998), as well as in speech recognition (Gallwitz et al., 2002; Hasegawa-Johnson et al., 2005; Ostendorf et al., 2003). Additionally, prosodic features such as pause length, duration of words and phones, pitch contours, energy contours, and their normalized values have been used for speech processing tasks like sentence boundary detection (Liu et al., 2005).

Linguistic encoding schemes like ToBI (Silverman et al., 1992) have also been used for sentence boundary detection (Roark et al., 2006; Harper et al., 2005), as well as for parsing (Dreyer and Shafran, 2007; Gregory et al., 2004; Kahn et al., 2005). In the ToBI scheme, aspects of prosody such as tone, prominence, and degree of juncture between words are represented symbolically. For instance, Dreyer and Shafran (2007) use three classes of automatically detected ToBI break indexes, indicating major intonational breaks with a `4`, hesitation with a `p`, and all other breaks with a `1`.

Recently, Huang and Harper (2010) found that they could effectively integrate prosodic informa-

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 821–831,
MIT, Massachusetts, USA, 9-11 October 2010. ©2010 Association for Computational Linguistics

tion in the form of this simplified three class ToBI encoding when parsing spontaneous speech by using a prosodically enriched PCFG model with latent annotations (PCFG-LA) (Matsuzaki et al., 2005; Petrov and Klein, 2007) to rescore n-best parses produced by a baseline PCFG-LA model without prosodic enrichment. However, the prosodically enriched models by themselves did not perform significantly better than the baseline PCFG-LA model without enrichment, due to the negative effect that misalignments between automatic prosodic breaks and true phrase boundaries have on the model.

This paper investigates methods for using state-of-the-art taggers on conversational speech transcriptions and the effect that prosody has on tagging accuracy. Improving POS tagging performance of speech transcriptions has implications for improving downstream applications that rely on accurate POS tags, including sentence boundary detection (Liu et al., 2005), automatic punctuation (Hillard et al., 2006), information extraction from speech, parsing, and syntactic language modeling (Heeman, 1999; Filimonov and Harper, 2009). While there have been several attempts to integrate prosodic information to improve parse accuracy of speech transcripts, to the best of our knowledge there has been little work on using this type of information for POS tagging. Furthermore, most of the parsing work has involved generative models and rescoring/reranking of hypotheses from the generative models. In this work, we will analyze several factors related to effective POS tagging of conversational speech:

- discriminative versus generative POS tagging models (Section 2)
- prosodic features in the form of simplified ToBI break indexes (Section 4)
- type of speech segmentation (Section 5)

2 Models

In order to fully evaluate the difficulties inherent in tagging conversational speech, as well as the possible benefits of prosodic information, we conducted experiments with six different POS tagging models. The models can be broadly separated into two classes: generative and discriminative. As the first of our generative models, we used a Hidden Markov Model (HMM) trigram tagger (Thede and Harper, 1999), which serves to establish a baseline and to gauge the difficulty of the task at hand. Our second model, HMM-LA, was the latent variable bigram HMM tagger of Huang et al. (2009), which achieved state-of-the-art tagging performance by introducing latent tags to weaken the stringent Markov independence assumptions that generally hinder tagging performance in generative models.

For the third model, we implemented a bidirectional variant of the HMM-LA (HMM-LA-Bidir) that combines evidence from two HMM-LA taggers, one trained left-to-right and the other right-to-left. For decoding, we use a product model (Petrov, 2010). The intuition is that the context information from the left and the right of the current position is complementary for predicting the current tag and thus, the combination should serve to improve performance over the HMM-LA tagger.

Since prior work on parsing speech with prosody has relied on generative models, it was necessary to modify equations of the model in order to incorporate the prosodic information, and then perform rescoring in order to achieve gains. However, it is far simpler to directly integrate prosody as features into the model by using a discriminative approach. Hence, we also investigate several log-linear models, which allow us to easily include an arbitrary number and varying kinds of possibly overlapping and non-independent features.

First, we implemented a Conditional Random Field (CRF) tagger, which is an attractive choice due to its ability to learn the globally optimal labeling for a sequence and proven excellent performance on sequence labeling tasks (Lafferty et al., 2001). In contrast to an HMM which optimizes the joint likelihood of the word sequence and tags, a CRF optimizes the conditional likelihood, given by:

$$p_\lambda(t|w) = \frac{\exp \sum_j \lambda_j F_j(t, w)}{\sum_t \exp \sum_j \lambda_j F_j(t, w)} \qquad (1)$$

where the λ's are the parameters of the model to estimate and F indicates the feature functions used. The denominator in (1) is $Z_\lambda(x)$, the normalization factor, with:

$$F_j(t, w) = \sum_i f_j(t, w, i)$$

Class	Model Name	Latent Variable	Bidirectional	N-best-Extraction	Markov Order
Generative	Trigram HMM			√	2nd
	HMM-LA	√		√	1st
	HMM-LA-Bidir	√	√		1st
Discriminative	Stanford Bidir		√		2nd
	Stanford Left5				2nd
	CRF				2nd

Table 1: Description of tagging models

The objective we need to maximize then becomes :

$$\mathcal{L} = \sum_n \left[\sum_j \lambda_j F_j(t_n, w_n) - \log Z_\lambda(x_n) \right] - \frac{\|\lambda\|^2}{2\sigma^2}$$

where we use a spherical Gaussian prior to prevent overfitting of the model (Chen and Rosenfeld, 1999) and the wide-spread quasi-Newtonian L-BFGS method to optimize the model parameters (Liu and Nocedal, 1989). Decoding is performed with the Viterbi algorithm.

We also evaluate state-of-the-art Maximum Entropy taggers: the Stanford Left5 tagger (Toutanova and Manning, 2000) and the Stanford bidirectional tagger (Toutanova et al., 2003), with the former using only left context and the latter bidirectional dependencies.

Table 1 summarizes the major differences between the models along several dimensions: (1) generative versus discriminative, (2) directionality of decoding, (3) the presence or absence of latent annotations, (4) the availability of n-best extraction, and (5) the model order.

In order to assess the quality of our models, we evaluate them on the section 23 test set of the standard newswire WSJ tagging task after training all models on sections 0-22. Results appear in Table 2. Clearly, all the models have high accuracy on newswire data, but the Stanford bidirectional tagger significantly outperforms the other models with the exception of the HMM-LA-Bidir model on this task.[1]

Model	Accuracy
Trigram HMM	96.58
HMM-LA	97.05
HMM-LA-Bidir	97.16
Stanford Bidir	97.28
Stanford Left5	97.07
CRF	96.81

Table 2: Tagging accuracy on WSJ

3 Experimental Setup

In the rest of this paper, we evaluate the tagging models described in Section 2 on conversational speech. We chose to utilize the Penn Switchboard (Godfrey et al., 1992) and Fisher treebanks (Harper et al., 2005; Bies et al., 2006) because they provide gold standard tags for conversational speech and we have access to corresponding automatically generated ToBI break indexes provided by (Dreyer and Shafran, 2007; Harper et al., 2005)[2].

We utilized the Fisher dev1 and dev2 sets containing 16,519 sentences (112,717 words) as the primary training data and the entire Penn Switchboard treebank containing 110,504 sentences (837,863 words) as an additional training source[3]. The treebanks were preprocessed as follows: the tags of auxiliary verbs were replaced with the AUX tag, empty nodes

[1] Statistically significant improvements are calculated using the sign test ($p < 0.05$).

[2] A small fraction of words in the Switchboard treebank do not align with the break indexes because they were produced based on a later refinement of the transcripts used to produce the treebank. For these cases, we heuristically added break *1* to words in the middle of a sentence and *4* to words that end a sentence.

[3] Preliminary experiments evaluating the effect of training data size on performance indicated using the additional Switchboard data leads to more accurate models, and so we use the combined training set.

and function tags were removed, words were downcased, punctuation was deleted, and the words and their tags were extracted. Because the Fisher treebank was developed using the lessons learned when developing Switchboard, we chose to use its eval portion for development (the first 1,020 tagged sentences containing 7,184 words) and evaluation (the remaining 3,917 sentences with 29,173 words).

We utilize the development set differently for the generative and discriminative models. Since the EM algorithm used for estimating the parameters in the latent variable models introduces a lot of variability, we train five models with a different seed and then choose the best one based on dev set performance. For the discriminative models, we tuned their respective regularization parameters on the dev set. All results reported in the rest of this paper are on the test set.

4 Integration of Prosodic Information

In this work, we use three classes of automatically generated ToBI break indexes to represent prosodic information (Kahn et al., 2005; Dreyer and Shafran, 2007; Huang and Harper, 2010): `4`, `1`, and `p`. Consider the following speech transcription example, which is enriched with ToBI break indexes in parentheses and tags: `i(1)/PRP did(1)/VBD n't(1)/RB you(1)/PRP know(4)/VBP i(1)/PRP did(1)/AUX n't(1)/RB...` The speaker begins an utterance, and then restarts the utterance. The automatically predicted break `4` associated with `know` in the utterance compellingly indicates an intonational phrase boundary and could provide useful information for tagging if we can model it appropriately.

To integrate prosody into our generative models, we utilize the method from (Dreyer and Shafran, 2007) to add prosodic breaks. As Figure 1 shows, ToBI breaks provide a secondary sequence of observations that is parallel to the sequence of words that comprise the sentence. Each break b_i in the secondary sequence is generated by the same tag t_i as that which generates the corresponding word w_i, and so it is conditionally independent of its corresponding word given the tag:

$$P(w, b|t) = P(w|t)P(b|t)$$

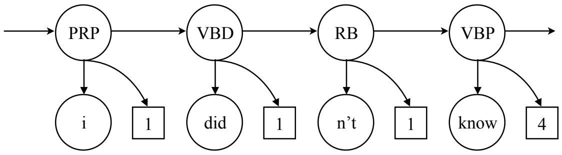


Figure 1: Parallel generation of words and breaks for the HMM models

The HMM-LA taggers are then able to split tags to capture implicit higher order interactions among the sequence of tags, words, and breaks.

The discriminative models are able to utilize prosodic features directly, enabling the use of contextual interactions with other features to further improve tagging accuracy. Specifically, in addition to the standard set of features used in the tagging literature, we use the feature templates presented in Table 3, where each feature associates the break b_i, word w_i, or some combination of the two with the current tag t_i[4].

Break and/or word values	Tag value
b_i=B	t_i = T
b_i=B & b_{i-1}=C	t_i = T
w_i=W & b_i=B	t_i = T
w_{i+1}=W & b_i=B	t_i = T
w_{i+2}=W & b_i=B	t_i = T
w_{i-1}=W & b_i=B	t_i = T
w_{i-2}=W & b_i=B	t_i = T
w_i=W & b_i=B & b_{i-1}=C	t_i = T

Table 3: Prosodic feature templates

5 Experiments

5.1 Conversation side segmentation

When working with raw speech transcripts, we initially have a long stream of unpunctuated words, which is called a conversation side. As the average length of conversation side segments in our data is approximately 630 words, it poses quite a challenging tagging task. Thus, we hypothesize that it is on these large segments that we should achieve the most

[4] We modified the Stanford taggers to handle these prosodic features.

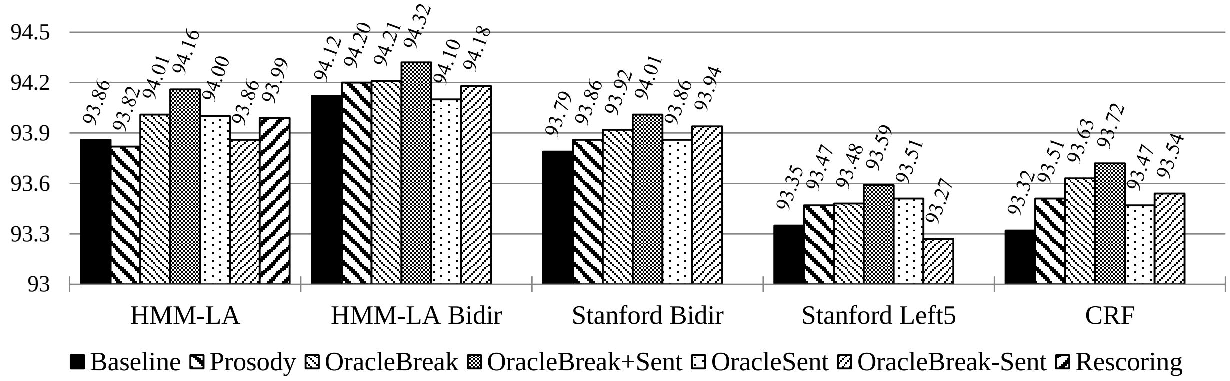


Figure 2: Tagging accuracy on conversation sides

improvement from the addition of prosodic information.

In fact, as the baseline results in Figure 2 show, the accuracies achieved on this task are much lower than those on the newswire task. The trigram HMM tagger accuracy drops to 92.43%, while all the other models fall to within the range of 93.3%-94.12%, a significant departure from the 96-97.3% range on newswire sentences. Note that the Stanford bidirectional and HMM-LA tagger perform very similarly, although the HMM-LA-Bidir tagger performs significantly better than both. In contrast to the newswire task on which the Stanford bidirectional tagger performed the best, on this genre, it is slightly worse than the HMM-LA tagger, albeit the difference is not statistically significant.

With the direct integration of prosody into the generative models (see Figure 2), there is a slight but statistically insignificant shift in performance. However, integrating prosody directly into the discriminative models leads to significant improvements in the CRF and Stanford Left5 taggers. The gain in the Stanford bidirectional tagger is not statistically significant, however, which suggests that the left-to-right models benefit more from the addition of prosody than bidirectional models.

5.2 Human-annotated sentences

Given the lack-luster performance of the tagging models on conversation side segments, even with the direct addition of prosody, we chose to determine the performance levels that could be achieved on this task using human-annotated sentences, which we will refer to as sentence segmentation. Figure 3 reports the baseline tagging accuracy on sentence segments, and we see significant improvements across all models. The HMM Trigram tagger performance increases to 93.00%, while the increase in accuracy for the other models ranges from around 0.2-0.3%. The HMM-LA taggers once again achieve the best performance, with the Stanford bidirectional close behind. Although the addition of prosody has very little impact on either the generative or discriminative models when applied to sentences, the baseline tagging models (i.e., not prosodically enriched) significantly outperform all of the prosodically enriched models operating on conversation sides.

At this point, it would be apt to suggest using automatic sentence boundary detection to create shorter segments. Table 4 presents the results of using baseline models without prosodic enrichment trained on the human-annotated sentences to tag automatically segmented speech[5]. As can be seen, the results are quite similar to the conversation side segmentation performances, and thus significantly lower than when tagging human-annotated sentences. A caveat to consider here is that we break the standard assumption that the training and test set be drawn from the same distribution, since the training data is human-annotated and the test is automatically segmented. However, it can be quite challenging to create a corpus to train on that represents the biases of the systems that perform automatic sentence segmentation. Instead, we will examine an-

[5] We used the Baseline Structural Metadata System described in Harper et al. (2005) to predict sentence boundaries.

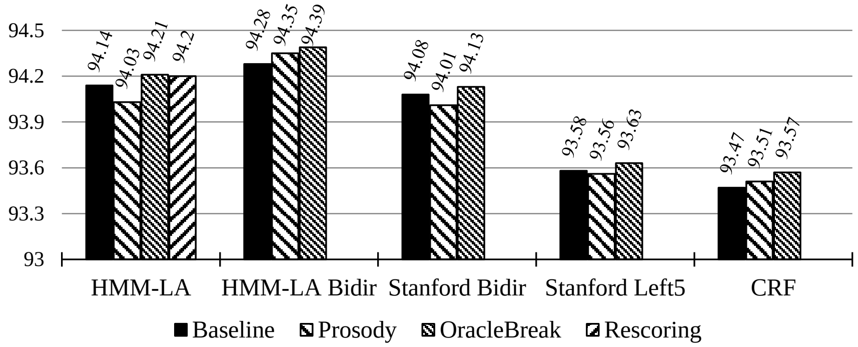


Figure 3: Tagging accuracy on human-annotated segments

other segmentation method to shorten the segments automatically, i.e., by training and testing on speaker turns, which preserves the train-test match, in Section 5.5.

Model	Accuracy
HMM-LA	93.95
HMM-LA-Bidir	94.07
Stanford Bidir	93.77
Stanford Left5	93.35
CRF	93.29

Table 4: Baseline tagging accuracy on automatically detected sentence boundaries

5.3 Oracle Break Insertion

As we believe one of the major roles that prosodic cues serve for tagging conversation sides is as a proxy for sentence boundaries, perhaps the efficacy of the prosodic breaks can, at least partially, be attributed to errors in the automatically induced break indexes themselves, as they can misalign with syntactic phrase boundaries, as discussed in Huang and Harper (2010). This may degrade the performance of our models more than the improvement achieved from correctly placed breaks. Hence, we conduct a series of experiments in which we systematically eliminate noisy phrase and disfluency breaks and show that under these improved conditions, prosodically enriched models can indeed be more effective.

To investigate to what extent noisy breaks are impeding the possible improvements from prosodically enriched models, we replaced all `4` and `p` breaks in the training and evaluation sets that did not align to the correct phrase boundaries as indicated by the treebank with break `1` for both the conversation sides and human-annotated sentences. The results from using Oracle Breaks on conversation sides can be seen in Figure 2. All models except Stanford Left5 and HMM-LA-Bidir significantly improve in accuracy when trained and tested on the Oracle Break modified data. On human-annotated sentences, Figure 3 shows improvements in accuracies across all models, however, they are statistically insignificant.

To further analyze why prosodically enriched models achieve more improvement on conversation sides than on sentences, we conducted three more Oracle experiments on conversation sides. For the first, OracleBreak-Sent, we further modified the data such that all breaks corresponding to a sentence ending in the human-annotated segments were converted to break `1`, thus effectively only leaving inside sentence phrasal boundaries. This modification results in a significant drop in performance, as can be seen in Figure 2.

For the second, OracleSent, we converted all the breaks corresponding to a sentence end in the human-annotated segmentations to break `4`, and all the others to break `1`, thus effectively only leaving sentence boundary breaks. This performed largely on par with OracleBreak, suggesting that the phrase-aligned prosodic breaks seem to be a stand-in for sentence boundaries.

Finally, in the last condition, OracleBreak+Sent, we modified the OracleBreak data such that all breaks corresponding to a sentence ending in the human-annotated sentences were converted to break

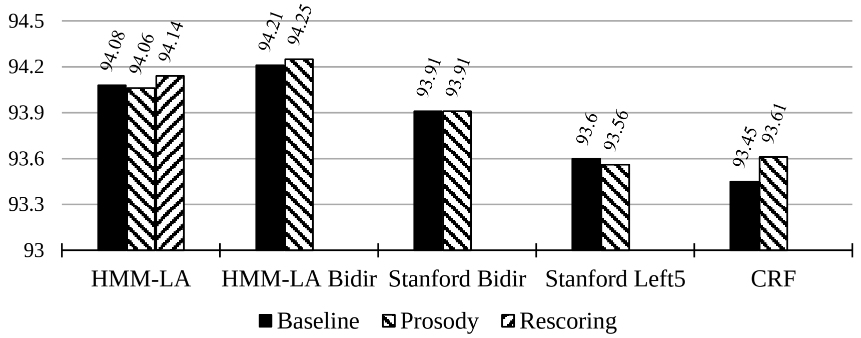


Figure 4: Tagging accuracy on speaker turns

4 (essentially combining OracleBreak and OracleSent). As Figure 2 indicates, this modification results in the best tagging accuracies for all the models. All models were able to match or even improve upon the baseline accuracies achieved on the human segmented data. This suggests that when we have breaks that align with phrasal and sentence boundaries, prosodically enriched models are highly effective.

5.4 N-best Rescoring

Based on the findings in the previous section and the findings of (Huang and Harper, 2010), we next apply a rescoring strategy in which the search space of the prosodically enriched generative models is restricted to the n-best list generated from the baseline model (without prosodic enrichment). In this manner, the prosodically enriched model can avoid poor tag sequences produced due to the misaligned break indexes. As Figure 2 shows, using the baseline conversation side model to produce an n-best list for the prosodically enriched model to rescore results in significant improvements in performance for the HMM-LA model, similar to the parsing results of (Huang and Harper, 2010). The size of the n-best list directly impacts performance, as reducing to $n = 1$ is akin to tagging with the baseline model, and increasing $n \rightarrow \infty$ amounts to tagging with the prosodically enriched model. We experimented with a number of different sizes for n and chose the best one using the dev set. Figure 3 presents the results for this method applied to human-annotated sentences, where it produces only marginal improvements[6].

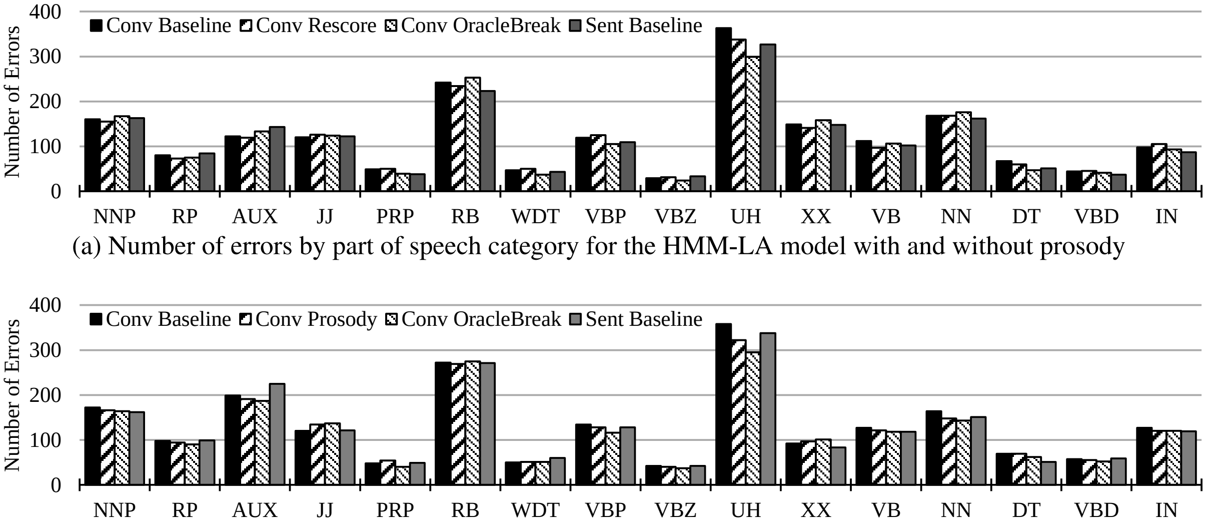
5.5 Speaker turn segmentation

The results presented thus far indicate that if we have access to close to perfect break indexes, we can use them effectively, but this is not likely to be true in practice. We have also observed that tagging accuracy on shorter conversation sides is greater than longer conversation sides, suggesting that post-processing the conversation sides to produce shorter segments would be desirable.

We thus devised a scheme by which we could automatically extract shorter speaker turn segments from conversation sides. For this study, speaker turns, which effectively indicate speaker alternations, were obtained by using the metadata in the treebank to split the sentences into chunks based on speaker change. Every time a speaker begins talking after the other speaker was talking, we start a new segment for that speaker. In practice, this would need to be done based on audio cues and automatic transcriptions, so these results represent an upper bound.

Figure 4 presents tagging results on speaker turn segments. For most models, the difference in accuracy achieved on these segments and that of human-annotated sentences is statistically insignificant. The only exception is the Stanford bidirectional tagger,

[6] Rescoring using the CRF model was also performed, but led to a performance degradation. We believe this is due to the fact that the prosodically enriched CRF model was able to directly use the break index information, and so restricting it to the baseline CRF model search space limits the performance to that of the baseline model.



(a) Number of errors by part of speech category for the HMM-LA model with and without prosody

(b) Number of errors by part of speech category for the CRF model with and without prosody

Figure 5: Error reduction for prosodically enriched HMM-LA (a) and CRF (b) models

which performs worse on these slightly longer segments. With the addition of break indexes, we see marginal changes in most of the models; only the CRF tagger receives a significant boost. Thus, models achieve performance gains from tagging shorter segments, but at the cost of limited usefulness of the prosodic breaks. Overall, speaker turn segmentation is an attractive compromise between the original conversation sides and human-annotated sentences.

6 Discussion

Across the different models, we have found that taggers applied to shorter segments, either sentences or speaker turns, do not tend to benefit significantly from prosodic enrichment, in contrast to conversation sides. To analyze this further we broke down the results by part of speech for the two models for which break indexes improved performance the most: the CRF and HMM-LA rescoring models, which achieved an overall error reduction of 2.8% and 2.1%, respectively. We present those categories that obtained the greatest benefit from prosody in Figure 5 (a) and (b). For both models, the UH category had a dramatic improvement from the addition of prosody, achieving up to a 10% reduction in error.

For the CRF model, other categories that saw impressive error reductions were NN and VB, with 10% and 5%, respectively. Table 5 lists the prosodic features that received the highest weight in the CRF model. These are quite intuitive, as they seem to represent places where the prosody indicates sentence or clausal boundaries. For the HMM-LA model, the VB and DT tags had major reductions in error of 13% and 10%, respectively. For almost all categories, the number of errors is reduced by the addition of breaks, and further reduced by using the OracleBreak processing described above.

Weight	Feature
2.2212	w_i=um & b_i=4 & t=UH
1.9464	w_i=uh & b_i=4 & t=UH
1.7965	w_i=yes & b_i=4 & t=UH
1.7751	w_i=and & b_i=4 & t=CC
1.7554	w_i=so & b_i=4 & t=RB
1.7373	w_i=but & b_i=4 & t=CC

Table 5: Top break 4 prosody features in CRF prosody model

To determine more precisely the effect that the segment size has on tagging accuracy, we extracted the oracle tag sequences from the HMM-LA and CRF baseline and prosodically enriched models across conversation sides, sentences, and speaker turn segments. As the plot in Figure 6 shows, as we increase the n-best list size to 500, the oracle accuracy of the models trained on sentences in-

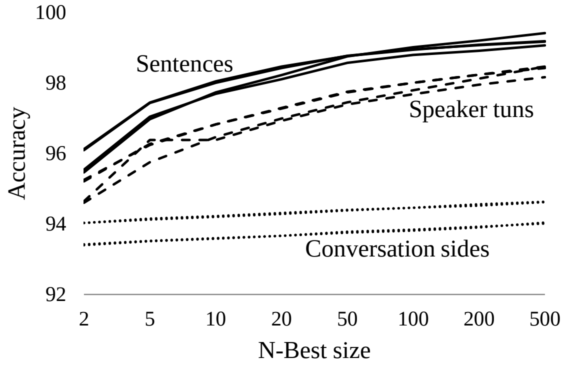


Figure 6: Oracle comparison: solid lines for sentences, dashed lines for speaker turns, and dotted lines for conversation sides

creases rapidly to 99%; whereas, the oracle accuracy of models on conversation sides grow slowly to between 94% and 95%. The speaker turn trained models, however, behave closely to those using sentences, climbing rapidly to accuracies of around 98%. This difference is directly attributable to the length of the segments. As can be seen in Table 6, the speaker turn segments are more comparable in length to sentences.

	Train	**Eval**
Conv	627.87 ± 281.57	502.98 ± 151.22
Sent	7.52± 7.86	7.45 ± 8.29
Speaker	15.60± 29.66	15.27± 21.01

Table 6: Length statistics of different data segmentations

Next, we return to the large performance degradation when tagging speech rather than newswire text to examine the major differences among the models. Using two of our best performing models, the Stanford bidirectional and HMM-LA, in Figure 7 we present the categories for which performance degradation was the greatest when comparing performance of a tagger trained on WSJ to a tagger trained on spoken sentences and conversation sides. The performance decrease is quite similar across both models, with the greatest degradation on the NNP, RP, VBN, and RBS categories.

Unsurprisingly, both the discriminative and generative bidirectional models achieve the most impressive results. However, the generative HMM-LA and HMM-LA-Bidir models achieved the best results across all three segmentations, and the best overall result, of 94.35%, on prosodically enriched sentence-segmented data. Since the Stanford bidirectional model incorporates all of the features that produced its state-of-the-art performance on WSJ, we believe the fact that the HMM-LA outperforms it, despite the discriminative model's more expressive feature set, is indicative of the HMM-LA's ability to more effectively adapt to novel domains during training. Another challenge for the discriminative models is the need for regularization tuning, requiring additional time and effort to train several models and select the most appropriate parameter each time the domain changes. Whereas for the HMM-LA models, although we also train several models, they can be combined into a product model, such as that described by Petrov (2010), in order to further improve performance.

Since the prosodic breaks are noisier features than the others incorporated in the discriminative models, it may be useful to set their regularization parameter separately from the rest of the features, however, we have not explored this alternative. Our experiments used human transcriptions of the conversational speech; however, realistically our models would be applied to speech recognition transcripts. In such a case, word error will introduce noise in addition to the prosodic breaks. In future work, we will evaluate the use of break indexes for tagging when there is lexical error. We would also apply the n-best rescoring method to exploit break indexes in the HMM-LA bidirectional model, as this would likely produce further improvements.

7 Conclusion

In this work, we have evaluated factors that are important for developing accurate tagging models for speech. Given that prosodic breaks were effective knowledge sources for parsing, an important goal of this work was to evaluate their impact on various tagging model configurations. Specifically, we have examined the use of prosodic information for tagging conversational speech with several different discriminative and generative models across three different speech transcript segmentations. Our find-

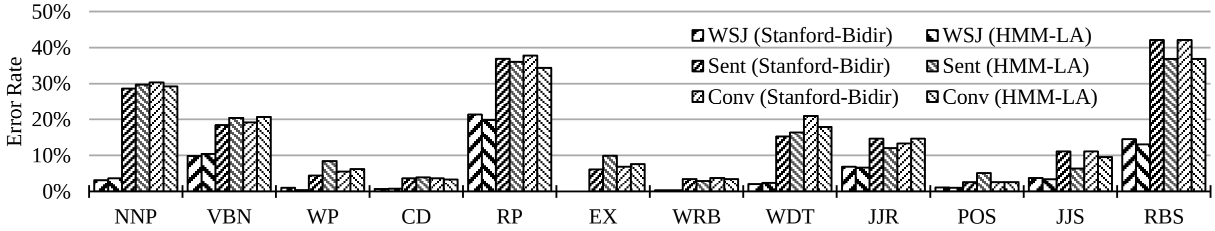


Figure 7: Comparison of error rates between the Standford Bidir and HMM-LA models trained on WSJ, sentences, and conversation sides

ings suggest that generative models with latent annotations achieve the best performance in this challenging domain. In terms of transcript segmentation, if sentences are available, it is preferable to use them. In the case that no such annotation is available, then using automatic sentence boundary detection does not serve as an appropriate replacement, but if automatic speaker turn segments can be obtained, then this is a good alternative, despite the fact that prosodic enrichment is less effective.

Our investigation also shows that in the event that conversation sides must be used, prosodic enrichment of the discriminative and generative models produces significant improvements in tagging accuracy (by direct integration of prosody features for the former and by restricting the search space and rescoring with the latter). For tagging, the most important role of the break indexes appears to be as a stand in for sentence boundaries. The oracle break experiments suggest that if the accuracy of the automatically induced break indexes can be improved, then the prosodically enriched models will perform as well, or even better, than their human-annotated sentence counterparts.

8 Acknowledgments

This research was supported in part by NSF IIS-0703859 and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-001. Any opinions, findings, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agency or the institutions where the work was completed.

References

Anton Batliner, Bernd Möbius, Gregor Möhler, Antje Schweitzer, and Elmar Nöth. 2001. Prosodic models, automatic speech understanding, and speech synthesis: toward the common ground. In *Eurospeech*.

Ann Bies, Stephanie Strassel, Haejoong Lee, Kazuaki Maeda, Seth Kulick, Yang Liu, Mary Harper, and Matthew Lease. 2006. Linguistic resources for speech parsing. In *LREC*.

Stanley F. Chen and Ronald Rosenfeld. 1999. A Gaussian prior for smoothing maximum entropy models. Technical report, Technical Report CMU-CS-99-108, Carnegie Mellon University.

Anne Cutler, Delphine Dahan, and Wilma v an Donselaar. 1997. Prosody in comprehension of spoken language: A literature review. *Language and Speech*.

Markus Dreyer and Izhak Shafran. 2007. Exploiting prosody for PCFGs with latent annotations. In *Interspeech*.

Denis Filimonov and Mary Harper. 2009. A joint language model with fine-grain syntactic tags. In *EMNLP*.

Jennifer Foster. 2010. “cba to check the spelling”: Investigating parser performance on discussion forum posts. In *NAACL-HLT*.

Florian Gallwitz, Heinrich Niemann, Elmar Nöth, and Volker Warnke. 2002. Integrated recognition of words and prosodic phrase boundaries. *Speech Communication*.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP*.

Michelle L. Gregory, Mark Johnson, and Eugene Charniak. 2004. Sentence-internal prosody does not help parsing the way punctuation does. In *NAACL*.

Mary P. Harper, Bonnie J. Dorr, John Hale, Brian Roark, Izhak Shafran, Matthew Lease, Yang Liu, Matthew Snover, Lisa Yung, Anna Krasnyanskaya, and Robin Stewart. 2005. 2005 Johns Hopkins Summer Workshop Final Report on Parsing and Spoken Structural

Event Detection. Technical report, Johns Hopkins University.

Mark Hasegawa-Johnson, Ken Chen, Jennifer Cole, Sarah Borys, Sung suk Kim, Aaron Cohen, Tong Zhang, Jeung yoon Choi, Heejin Kim, Taejin Yoon, and Ra Chavarria. 2005. Simultaneous recognition of words and prosody in the boston university radio speech corpus. speech communication. *Speech Communication*.

Peter A. Heeman. 1999. POS tags and decision trees for language modeling. In *EMNLP*.

Dustin Hillard, Zhongqiang Huang, Heng Ji, Ralph Grishman, Dilek Hakkani-Tur, Mary Harper, Mari Ostendorf, and Wen Wang. 2006. Impact of automatic comma prediction on POS/name tagging of speech. In *ICASSP*.

Zhongqiang Huang and Mary Harper. 2010. Appropriately handled prosodic breaks help PCFG parsing. In *NAACL*.

Zhongqiang Huang, Vladimir Eidelman, and Mary Harper. 2009. Improving a simple bigram hmm part-of-speech tagger by latent annotation and self-training. In *NAACL-HLT*.

Jeremy G. Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, and Mari Ostendorf. 2005. Effective use of prosody in parsing conversational speech. In *EMNLP-HLT*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

D. C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*.

Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using conditional random fields for sentence boundary detection in speech. In *ACL*.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Probabilistic CFG with latent annotations. In *ACL*.

Mari Ostendorf, Izhak Shafran, and Rebecca Bates. 2003. Prosody models for conversational speech recognition. In *Plenary Meeting and Symposium on Prosody and Speech Processing*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*.

Slav Petrov. 2010. Products of random latent variable grammars. In *HLT-NAACL*.

Brian Roark, Yang Liu, Mary Harper, Robin Stewart, Matthew Lease, Matthew Snover, Izhak Shafran, Bonnie Dorr, John Hale, Anna Krasnyanskaya, and Lisa Yung. 2006. Reranking for sentence boundary detection in conversational speech. In *ICASSP*.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirshberg. 1992. ToBI: A standard for labeling English prosody. In *ICSLP*.

Paul Taylor and Alan W. Black. 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*.

Scott M. Thede and Mary P. Harper. 1999. A second-order hidden markov model for part-of-speech tagging. In *ACL*.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *EMNLP*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.