Modeling Prosodic Sequences with K-Means and Dirichlet Process GMMs

Andrew Rosenberg

Computer Science Department, Queens College CUNY, NYC, USA

andrew@cs.qc.cuny.edu

Abstract

In this paper we describe two unsupervised representations of prosodic sequences based on k-means and Dirichlet Process Gaussian Mixture Model (DPGMM) clustering. The clustering algorithms are used to infer an inventory of prosodic categories over automatically segmented syllables. A tri-gram model is trained over these sequences to characterize speech. We find that DPGMM clusters show a greater correspondence with manual ToBI labels than k-means clusters. However, sequence models trained on k-means clusters significantly outperform DPGMM sequences in classifying speaking style, nativeness and speakers. We also investigate the use of these sequence models in the detection of outliers regarding these three tasks. Non-parametric Bayesian techniques have the advantage of being able to learn a clustering solution and infer the number of clusters directly from data. While it is attractive to avoid specifying k before clustering, on the tasks of characterizing prosodic sequences we find that effective use of DPGMMs still requires a significant amount of parameter tuning, and performance fails to reach the level of k-means.

Index Terms: Prosodic Analysis, Speaking-style classification, Nativeness classification, Non-parametric Bayesian models

1. Introduction

Prosodic variation is used to communicate a wide range of information. Emotion and speaker state are indicated prosodically [1], contrast, topic and focus are indicated through prosodic prominence [2], and syntactic attachment can, in some cases, be disambiguated with prosodic information [3]. In addition to these specific elements of information, at a broader level, different types of speech are communicated with distinct prosody. Formal, prepared speech sounds different than spontaneous speech. Individual speakers have idiosyncrasies that are quickly recognizable. In this paper, we describe a sequential modeling approach over prosodic sequences to characterize speech in terms of speaking style, nativeness and speaker. This work builds on our previous work that explored these tasks using a sequence model over pitch points, and over hypothesized and manual ToBI labels [4].

There have been previous approaches to use clustering algorithms to identify an inventory of prosodic categories. Levow [5] used an asymmetric clustering algorithm as well as k-means, finding a high agreement with ToBI tones on broadcast news data. Ananthakrishnan et al. [6] used k-means, soft k-means and GMM clustering of acoustic information and combined this with lexical and syntactic priors to generate reliable prediction of accenting and phrasing. Somewhat less work has been done on constructing sequence models over prosodic categories. Early work by Wightman and Ostendorf used a transition model over ToBI labels to improve the performance of automatic ToBI labeling [7]. Shriberg et al. described "SNERF" or Syllable-based Non-uniform Extraction Region Feature n-grams [8]. These are quantized acoustic features drawn from syllable regions. An SVM was trained on the frequency of SNERF n-grams and applied to speaker recognition.

In this work, we represent prosodic sequences as a categorical representation of acoustic/prosodic content drawn from sequences of syllables. The syllables that we use are identified by an envelope-based pseudosyllabification algorithm (cf. Section 2). We experiment with two clustering algorithms to generate representations of the acoustic information contained in these syllables. We use the traditional k-means clustering, and a Non-parametric Bayesian approach, Dirichlet Process Gaussian Mixture Models (DPGMM). One of the limitations of k-means is the requirement that the user must specify k ahead of time. Non-parametric Bayesian clustering approaches have the advantage that the number of clusters is inferred from data.

Our hypothesis is that that DPGMMs are an effective tool to cluster prosodic inventories. There are two bases for this hypothesis. First, by being able to learn the number of clusters from data directly, the requisite tuning experiments to identify an effective number of clusters will be minimized. Second, this model should effectively model prosodic inventories. The DPGMM (as do many other Non-parametric Bayesian models) have a "rich-get-richer" property. That is, the clusters that are discovered first tend to grow quite large, relative to the population of other clusters. This leads to a highly skewed distribution of cluster sizes, often following a Zipf distribution. For representing prosodic variation, this should be an effective representation. In Standard American English, most syllables contain no prosodic marking; they are not lexically or prosodically stressed and they do not convey phrase ending indicators. Under the ToBI framework, even within prominent syllables, pitch accent categories also follow a heavily skewed distribution with approximately 80% of pitch accents being either H* or !H* [9].

However, despite our best efforts, we find little evidence to support this hypothesis. We find that k-means yields a more effective representation of prosody for distinguishing speaking style, nativeness and speaker. Our best explanation for this result may be due to the rich-get-richer property of the DPGMM. Inspecting the cluster size distribution, we find that over 90% of syllables are contained in the two largest clusters regardless of hyperparameter settings.

2. Material

In this section, we describe 1) the corpora used in our experiments, 2) the pseudosyllabification algorithm and 3) the feature representation used.

Corpora: We model prosodic sequences to distinguish speaking style, nativeness and speakers. The four speaking-styles, READ, SPONTANEOUS, BROADCAST NEWS (BN) and DIALOG, are drawn from three corpora. The Boston Directions Corpus (BDC) [10] contains spoken material from four

speakers delivering both spontaneous elicited monologues and reading their own spontaneous monologues approximately two weeks after their original production. Each BDC file represents a unique direction giving task. In total there are 50 minutes of read and 60 minutes of spontaneous speech. Both are spoken by the same four speakers. These two subcorpora will be used to represent READ and SPONTANEOUS speech. The Boston University Radio News Corpus (BURNC) [11] is a corpus of professionally read radio news data. A 2.35 hour subset from six speakers (three female and three male) has been annotated with the full ToBI standard. Each file represents a paragraph of BN speech. This material will be used to represent the BROADCAST NEWS or BN style of speech. The Columbia Games Corpus (CGC) [12] is a collection of 12 spontaneous task-oriented dydactic conversations between native speakers of Standard American English (SAE). In each session, two subjects played a set of computer games requiring verbal communication to goals of identifying or moving images on a screen. Neither subject could see the other participant. Each file represents a transcript from a whole game. The corpus includes $\sim$320 minutes of speech. The DIALOG material is drawn from the CGC. It is valuable to note that this "dialog" material is evaluated based on one side of the dialog at a time; only one speaker is present in any training or evaluation file.

In the nativeness classification experiment, we will compare NATIVE or L1 speech drawn from the BURNC, to NON-NATIVE or L2 speech productions of the same material. The material contains two news stories drawn from BURNC: **p** – computerized parole officers – and **r** – the Safe Roads Act. The non-native material was read by 4 native Mandarin Chinese speakers, between 25 and 30 years old, with 6 to 19 years of experience with English. The non-native material has been annotated using the ToBI standard. Each file in this corpus contains a full BURNC story. While we refer to this corpus as non-native or L2 speech, it is important to acknowledge that it only represents L2 productions by native Mandarin Chinese speakers rather than "non-nativeness" defined more broadly.

Pseudosyllabification: Our intention in this work is to develop a completely unsupervised approach to modeling prosodic contours. To that end, we use a syllabification algorithm that requires no manual intervention or tuning to new material. For this segmentation, we use a pseudosyllabification algorithm described by Villing et al. [13] and implemented in AuToBI [14]. This algorithm operates by assuming that amplitude peaks are associated with syllable nuclei, and valleys are syllable boundaries. The waveform is passed through an equal loudness filter, and a low-pass filter. Onset velocities are determined by the maxima of the filtered envelope slopes. These onsets are used as candidate syllable boundaries. Boundaries are then selected based on a scoring function incorporating the velocity of the onset and spectral content at the candidate vowel peak. An additional temporal filter is applied to the candidate boundaries to suppress weak boundaries within 100ms of strong boundaries. This also serves to eliminate spurious boundaries at the start and end of an utterance. Any identified syllable region with a mean intensity below 10dB is considered silence. The implementation used in this paper is available as part of AuToBI and can be found at `http://speech.cs.qc.cuny.edu/autobi`.

The hypothesized syllabic regions form the basis of the remaining experiments in this paper. The BURNC material contains 45,598 hypothesized syllables, BDC-read 11,651, BDC-spontaneous 13,739, L2 6,043, and Games 81,698. This leads to a total data set of 164,729 hypothesized syllables.

Feature Representation: To maintain a low dimensionality of feature representation for the DPGMM clustering algorithm we identify a relatively small set of acoustic features to describe the prosodic content of each syllable. We extract seven acoustic/prosodic features including 1,2) the mean range normalized intensity and delta, 3, 4) the mean z-score normalized pitch (log f0) and delta, 5) the duration, 6, 7) the duration of previous and following pauses. While this is certainly an impoverished representation of the full range of prosodic variation, it captures fundamental dimensions of variation corresponding to phrasing and prominence.

3. Clustering

We apply two clustering algorithms to generate an inventory of prosodic categories in this work.

The k-means algorithm operates by identifying k cluster centroids and iteratively assigns data points to the closest centroid and then re-estimates the centroid location as the average of all assigned points. This is a simple and efficient clustering algorithm. There are two major limitations of k-means. First the results can be sensitive to the choice of the initial centroids. Second, the user must specify a value of k before generating the clusters. In this work, we investigate values of k between 2 and 20, and at intervals of ten between 30 and 100.

Dirichlet Process Gaussian Mixture Models (DPGMM) are infinite mixture models with a Dirichlet Process prior over the cluster distribution. We use an implementation of a *stick-breaking* process for the Dirichlet Process [15]. For each sample there is a probability (governed by a Beta distribution) that a point will be generated by an existing cluster, or a new, unassigned cluster. The metaphor here is to assume a stick of length 1. The Beta distribution determines how much of the remaining stick is assigned to the next cluster. When a new cluster is needed, the beta distribution determines where to "break" the remainder of the stick. Within each cluster, the observation, here the acoustic feature vector, is modeled by a multivariate gaussian. While there is no closed form solution to fit a DPGMM to data, both Gibbs sampling and variational methods have been developed to perform inference.

In this work, we use a variational inference method based on [16] and written by Tom Haines. It is available at: `https://code.google.com/p/haines/wiki/dpgmm`. This implementation allows for two hyper-parameters which are priors over the scaling parameter β of the beta distribution. These hyperparameters are the parameters of a Gamma distribution, $Gamma(\alpha_0, \alpha_1)$. We keep α_0 fixed at 1, and select α_1 from 75, 50, 10, 1, 0.5, 0.25, 0.10, 0.05. We note, however, that the choice of prior has little impact on the number of resulting clusters; all clustering solutions are comprised of 11,11,11,18,15,13,13,13 clusters. Increasing α_1 to 90 or greater leads to singular matrices during fitting.

4. Experiments

In this section we describe three experiments. The first experiment (cf. Section 1 examines the similarity between the inferred clusters, and manual annotation of ToBI labels. The next two experiments use n-gram sequential modeling to 1) classify speaking style, nativeness and speakers (cf. Section 4.2 and 2) recognize outliers by thresholding model likelihood (cf. Section 4.3). In both experiments we use a tri-gram model with back off and Good Turing smoothing. These models are trained and evaluated using SRILM [17]. In all experiments, we use clusters learned over *all* material described in Section 2.

4.1. Consistency with ToBI labels

We evaluate the agreement between the categories learned by unsupervised means to manually annotated ToBI Labels. We construct class-cluster contingency matrices over all corpora. We evaluate the class-cluster consistency using V-Measure [18].

Since each syllable can contain *both* an accent and phrase boundary event, we calculate the correspondence between clusters and accent types, and clusters and phrase-ending types separately. Figures 1 and 1 contains the V-Measures between k-means, DPGMM clusterings and ToBI accent and phrase ending labels respectively. The DPGMM V-measure values do not vary by more that 0.001 under any hyperparameter setting.

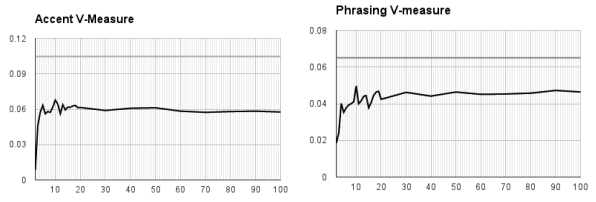


Figure 1: *V-measure of clusters and ToBI labels varying k in k-means (in black) and constant DPGMM value (in gray).*

We find that clusters learned with DPGMM are much more consistent with manual annotation of ToBI labels than k-means derived clusters are with 0.104 for accenting and 0.065 for phrasing. We note that k-means with 10 clusters yields the highest consistency with ToBI labels with 0.0678 for accenting and 0.050 for phrasing These agreements are quite low; a perfect correspondence would lead to a V-Measure of 1.0.

We attribute much of this agreement to the correspondence of the large DPGMM cluster with the similarly large unaccented and non-phrase boundary classes. The k-means clustering generate multiple partitions of both of these groups.

4.2. Classification

In this section, we evaluate the use of the inferred prosodic clusters to distinguish speaking-style, nativeness and speakers. In each of these, we train a sequence model independently for each of the label. The evaluation assigns the class whose model yields the smallest perplexity against a test sequence. We first evaluate performance using 500 samples of sequences of 100 syllables randomly drawn from test utterances. To determine how many syllables are required for reliable classification, we also evaluate performance with shorter sequences. We also compare the performance of the clusters to n-gram models over manual ToBI labels. In each experiment we identify the optimal hyperparameter setting using cross-validation and report the results on the test set.

In some cases 90% of data points are assigned to the same cluster by DPGMM clustering. Thus we also experiment with removing this large cluster from the mixture model (DPGMM'). We find that for speaking style and nativeness classification that this modification improves performance, suggesting that the rich-get-richer quality of the clustering may be working against us in these cases.

Speaking Style: We evaluate the use of sequential modeling over clusters for the classification of four speaking styles: READ, SPONTANEOUS, BN and DIALOG. From each corpus, we identify a single male speaker for testing: h2 for the BDC material, m1b from BURNC and 101 from CGC. No training material is spoken by the evaluation speaker. Table 1 contains the results of these 4-way classification experiments. We find that sequence modeling over k-means clusters outperforms the other three approaches by nearly 20%. We notice that omitting the dominant cluster from DPGMM modeling provides a small gain to classification performance, while ToBI labels yield only modest performance on this task.

	k-means (14)	DPGMM (75)	DPGMM' (0.25)	ToBI
Accuracy	86.6%	66.0%	67.4%	68.0%

Table 1: *Speaking-style classification Accuracy*

Nativeness: In this experiment, we examine the use of sequential modeling for the classification of speech as spoken by a native or non-native speaker. The training and evaluation material in this study is smaller than in the speaking style experiments. We use only two stories from the BURNC for training and evaluation, **p** and **r**. The evaluation material includes a single speaker – speaker m1b from the BURNC and s03 from the Mandarin Chinese material. Table 2 contains the results of these binary classification experiments. Again we find that k-means clustering yields the best classification results. However, all approaches are able to perform quite well. The impact of removing the dominant cluster from DPGMM modeling improves classification performance by 7.4%. This performance is achieved with a high value of k.

	k-means (40)	DPGMM (10)	DPGMM' (75)	ToBI
Accuracy	98.4%	90.8%	98.2%	94.2%

Table 2: *Nativeness classification Accuracy*

Speaker: In this experiment, we examine the use of sequential modeling for speaker identification. The training and evaluation material is drawn from the four BDC speakers. We collapse over read and spontaneous speech here, and use the subjects first six responses for training and 7, 8 and 9 for testing. Table 3 contains the results of these 4-way classification experiments. Again we find that the best classification performance is obtained by modeling k-means clusters. Neither DPGMM variation generates competitive performance. Again, we find that ToBI labels are unable to differences in prosodic sequences as effectively as k-means clusters.

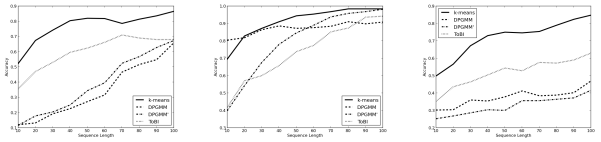
	k-means (70)	DPGMM (0.05)	DPGMM' (0.5)	ToBI
Accuracy	84.8%	47.0%	41.6%	63.0%

Table 3: *Speaker classification Accuracy*

Impact of Sequence Length: To determine the influence of sequence length on the classification performance, we evaluate each of these classifiers with 500 randomly selected sequences of length 10-100 syllables. A syllable has an average duration of 200ms, this can be interpreted as sequences between $\sim$2 and 20 seconds. We plot the results of k-means, DPGMM, DPGMM' and ToBI labels on the three tasks in Figure 2. We find that on the Speaker and Speaking Style tasks, that the k-means performance consistently dominates the other three modeling approaches. On classifying nativeness, we see that k-means performs better with shorter sequences – with the exception of the performance of DPGMM with 10 syllable sequences. However DPGMM performance does not improve to the same extent with longer sequences.

4.3. Outlier detection

In this set of experiments, we evaluate the ability of each representation of prosodic contours to reliably capture the characteristics of either the speaking style, speaker or nativeness setting



(a) Speaking Style (b) Nativeness (c) Speaker

Figure 2: *Classification Results on varied-length sequences*

in isolation. We construct these experiments as outlier detection tasks, where we train a single model on a single class. We then determine the ability of each representation to accurately identify members of the training classes by thresholding the perplexity under this model. As in Section 4.2, we evaluate using 500 samples of sequences of 100 syllables drawn from random start points in test utterances. For each task, the best AUC and corresponding EER is reported across parameters of k-means and DPGMM and DPGMM'. We also report performance using sequences of manually annotated ToBI labels. We then present the performance of the best performing models with evaluation sequences between 10 and 100 syllables.

Speaking Style: We train a sequence model on Games Speakers 101-106. Table 4 reports the AUC of detecting the remaining Games speakers from BDC and BURNC material. Here, we find that k-means with k=2 is an effective detector of

	k-means (2)	DPGMM (0.1)	DPGMM' (0.05)	ToBI
AUC	.902	.185	.807	.776
EER	.192	.784	.278	.292

Table 4: *Speaking-style detection performance.*

speaking style, more effectively modeling prosodic sequences for this task than ToBI labels. However, the *best* DPGMM model performs outrageously poorly. Yet, if we omit the largest DPGMM cluster from the modeling (DPGMM'), we find that performance increases substantially, though not to the levels of k-means but better than ToBI labels.

Nativeness: We train a sequence model on the material from BURNC speakers f2b, m1b and m2b, using only the **p** and **r** stories. Table 5 reports the performance detecting native speech from non-native outliers using the L2 material and the 3 BURNC speakers. Here we find that all models are able to dis-

	k-means (13)	DPGMM (.5)	DPGMM' (0.05)	ToBI
AUC	.972	.986	.976	0.952
EER	.078	.064	.098	0.105

Table 5: *Nativeness detection performance.*

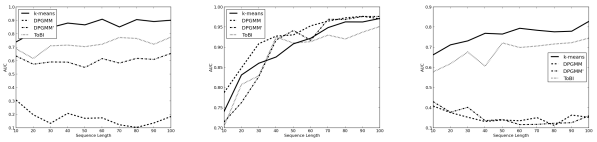
tinguish nativeness effectively based on sequences of syllables. The removal of the majority cluster does not help the DPGMM clustering as much in this task as elsewhere; this may be due to relatively high performance on this task and the method used to identify test sequences. Moreover, removal of the largest class does not dramatically impact performance.

Speaker: We train a sequence model on one half of the material from BURNC speaker f2b. Table 6 contains an ROC curve of detecting outliers from the remaining 5 BURNC speakers and other half of the f2b material. We observe again that k-means generates the best performing results for this detection task outperforming manual ToBI labels. Neither DPGMM configuration is effective at distinguishing speakers.

	k-means (60)	DPGMM (50)	DPGMM' (0.25)	ToBI
AUC	.829	.352	.360	.745
EER	.226	.604	.593	.316

Table 6: *Speaker detection performance.*

Impact of Sequence Length: As in Section 4.2, to determine how much spoken material is required to achieve these results, we evaluate this outlier detection approach using sequences from 10-100 syllables. We use the parameterizations that yielded the best performance as reported in this Section. Figure 3 contains a plot of the performance of each of the four modeling approaches on each of the three tasks while varying the evaluation sequence length. On all tasks, the k-means



(a) Speaking Style (b) Nativeness (c) Speaker

Figure 3: *Detection Results on varied-length sequences*

performance decreases by approximately 0.2 AUC when reducing the sequence length to 10. This is consistent across modeling style for nativeness recognition. On speaking style, the DPGMM' and ToBI decrease, but more slowly. The limited performance if DPGMM and DPGMM' on the speaker identification task is clearly seen here.

5. Conclusions and Future Work

We hypothesize that DPGMM clustering is an effective approach to modeling the prosodic content of a syllable. Our expectation was to use k-means as a comparison to this, more sophisticated and more flexible, model. However, we learned 1) that DPGMM clustering does not generate informative clusterings of prosodic content for distinguishing speaking style, nativeness or speaker, 2) while non-parametric methods are free of the requirement of specifying a number of clusters *a priori*, the resulting clusters is still dependent on hyperparameters which require tuning and 3) perhaps most surprisingly, that k-means derived clusters can be effectively used to model prosodic sequences, consistently outperforming ToBI labeling on all tasks.

DPGMM clustering may be limited by the dominance of a single cluster. We had anticipated that the smaller clusters would effectively capture unique qualities of speaking style, nativeness or speakers. Despite exploring a wide range of hyperparameter settings, we find that the largest two clusters dominate, covering more than 90% of all data points.

The acoustic/prosodic features we use in these experiments are rather limited, and yet still effective. In future work, we will explore clustering over a larger and more diverse set of features. While DPGMM has been shown to be an effective density estimation technique, we find that it is not well suited for clustering, we plan to explore a hybrid approach between k-means and DPGMM to limit the impact of the "rich-get-richer" property by using DPGMM mixture components to seed k-means, or using k-means solution to initialize the DPGMM mixture model.

6. Acknowledgements

This work was partially supported by DARPA FA8750-13-2-0041 - Deep Exploration and Filtering of Text (DEFT) Program.

7. References

[1] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and long-term features for emotion recognition," in *Interspeech*, 2009.

[2] N. Hedberg, "The prosody of contrastive topic and focus in spoken english," in *Workshop on information structure in context*, 2003.

[3] J. Pynte and B. Prieur, "Prosodic breaks and attachment decisions in sentence parsing," *Language and Cognitive Processes*, vol. 11, pp. 165–191, 1996.

[4] A. Rosenberg, "Symbolic and direct sequential modeling of prosody for classification of speaking-style and nativeness," in *Interspeech*, 2011.

[5] G.-A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *HLT-NAACL*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 224–231.

[6] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," in *ICSLP*, 2006.

[7] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processig*, vol. 2, no. 4, October 1994.

[8] E. Shriberg, L. Ferrer, and S. Kajarekar, "Svm modeling of snerfgrams for speaker recognition," in *Proc. ICSLP, South Korea*, 2004.

[9] 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 International Conference on Spoken Language Processing*, vol. 2, 1992, pp. 12–16.

[10] 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.

[11] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep. ECS-95-001, March 1995.

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

[13] R. Villing, J. Timoney, T. Ward, and J. Costello, "Automatic blind syllable segmentation for continuous speech," in *ISSC*, vol. 2004. IEEE, 2004, pp. 41–46.

[14] A. Rosenberg, "Autobi – a tool for automatic tobi annotation," in *Interspeech*, 2010.

[15] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[16] D. M. Blei and M. I. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[17] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *ICSLP*, 2002.

[18] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *EMNLP*, 2007.