PART-OF-SPEECH HISTOGRAMS FOR GENRE CLASSIFICATION OF TEXT

S. Feldman, M. A. Marin, M. Ostendorf, and M. R. Gupta

Dept. of Electrical Engineering
University of Washington, Seattle, Washington 98195
{sergeyf,amarin}@u.washington.edu, {mo,gupta}@ee.washington.edu

ABSTRACT

This work addresses the problem of classifying the genre of text, which is useful for a variety of language processing problems. We propose statistics of POS histograms as classification features, coupled with a quadratic discriminant classifier. In experiments on six different text and speech genres, we demonstrate enhanced performance compared to standard techniques using word frequency count features and POS trigram features. Experiments on genres that were not seen in training show intuitive overlaps with the training classes.

Index Terms— genre, web-filtering, text classification

1. INTRODUCTION

With increasing quantities of text and transcribed speech available online, it becomes of interest to search for documents based on characteristics beyond just topic. In particular, the genre of the document – whether it is a news report or an editorial, a speech transcript or a weblog – may be relevant for many tasks. For example, one might want to find "speeches on ethanol," or "weblog entries on Fannie Mae, sorted by most formal first." Genre classification is also of growing importance for natural language processing tasks, such as parsing, word sense disambiguation and translation, because of the potentially large differences in language associated with genre. Researchers find that genre-dependent models lead to improved performance on these tasks. One area where genre is known to be particularly important is in language modeling for speech recognition. Mismatching the genre of training and test data (e.g. using Wall Street Journal news for training a language model designed to recognize conversational telephone speech) can yield an order of magnitude worse perplexity scores and hurt recognition performance if the added data is given the same weight as the genre-matched training data [1]. At the same time, research on language modeling has consistently shown that having more data is the most important factor in performance. As a result, several researchers have investigated methods for searching the web for (roughly) genre-matched text to use in language model training [2, 3, 4, 5].

In prior work on genre classification, an important question is the definition of "genre." For many studies, genre is associated with purpose of the text, such as research article, novel, news report, editorial, advertisement, instructions, etc. In particular, several studies use those classes identified in the Brown corpus or British National Corpus. One study also considers spoken genres, including conversation, interview, debate and planned speech [6]. Another focuses on internet-specific document types, including different types of home pages (personal, public, commercial), bulletin boards, link lists, etc. [7]. In the work described here, we consider classes of text that include both written text and speech transcripts, formal and informal styles, spontaneous and pre-planned speech, and within-group vs. broadcast contexts. While we do not claim to cover the space of possibilities, we have tried to include texts that have some similarities as well as significant differences in style, but the focus is on the text and not on the web page type.

The standard features for text classification tasks in general are words, and both words and part-of-speech (POS) tags have been used with success in genre classification. Punctuation is shown to be useful [8, 9]. Kessler *et al.* [8] also found counts of particular syntactic categories to improve genre classification, and other knowledge-based features are described in [10]. In this paper, we explore the use of POS features (including punctuation) and a few added function words and interjections, since such labels are topic-independent. However, unlike prior work that looks at unigram or trigram counts, we propose the use of POS histogram statistics, which indirectly provide syntactic information without the cost of parsing.

Most prior work on genre detection has relied on either naive Bayes models [11, 6] or linear discriminant classifiers [10, 9], though [8] used neural networks. We include a naive Bayes baseline, but achieve best performance with a quadratic discriminant classifier, motivated by distributional analysis.

In the sections to follow, we describe the features used in more detail, with experiments on a six-genre data set, followed by analysis of results on new genres. The results show significant improvement in performance over word-based and POS trigram baselines. In a companion paper, experiments apply our results to the problem of filtering web text for lan-

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

guage modeling [12].

2. GENRE CLASSIFICATION FEATURES

The variation in word usage associated with topic dynamics in language is well-known and very powerful. It explains the success of bag-of-words models in information retrieval, and is the reason that simple cache n-gram models are among the most important techniques for reducing language model perplexity [13]. However, genre is also known to have a significant effect. Biber [14] provides statistics showing differences in the frequency of use of different POS or syntactic structures for fiction vs. exposition, focusing on ambiguous cases. For example, for the *-ed* forms of verbs such as *remembered*, the passive form is more common in exposition while the past tense is more common in fiction. For function words such as *until, before,* and *as*, subordination is more common in fiction, while the preposition usage is more common in exposition. Others have shown part-of-speech differences associated with different types of conversational speech, news text and email [1, 15]. Not surprisingly, filled pauses and pronouns are more frequent in spoken language than in written language; long noun phrases are more common in written language. Of course, the differences are much more complex than can be characterized by POS sequences, as evidenced by the relative lack of success in using a POS n-gram to "select" more conversational news broadcasts [1]. While prior work has shown usefulness of word-based features [8, 9, 11], and we augment our POS features with a few key words here, we focus on POS to keep the base dimensionality lower and allow the use of higher order statistics rather than simple counts.

We hypothesized that genres can be differentiated by the richness of POS used in phrases. To that end, we designed POS histogram statistic classification features as follows for each document,

Step 1: POS tagging Tag the document word sequence, resulting in the ℓ-length sequence p, based on a set of K POS or specific word tags.

Step 2: Sliding Window POS Histograms Let w be the length of a sliding window. For $j \in \{1, \ldots, \ell - w + 1\}$ calculate the histogram $h_j \in R^K$ of $\{p_j, \ldots, p_{j+w-1}\}$.

Step 3: Histogram Statistics Let $\mathcal{H} = \{h_1, \ldots, h_{\ell-w+1}\}$, and let $\mu(\mathcal{H}) \in R^K$ and $\sigma(\mathcal{H}) \in R^K$ be the mean and standard deviation of $\mathcal{H}$, respectively. Call $[\mu(\mathcal{H})\ \sigma(\mathcal{H})]^T$ the unnormalized feature vector.

Step 4: Normalize Standard-normalize the unnormalized feature vector according to the mean and variance computed from the training documents.

Step 5: Feature Reduction Perform principal components analysis on the normalized training features, then project the normalized feature vector onto these principal components and retain the top-ranked components as the feature vector.

Parameters of this approach include the choice of tag set, the sliding window size w, and the principal component threshold for feature reduction. We used a moving window width of $w = 5$ as a balance between capturing complex phrases and the short meter of conversational speech. Also, preliminary results indicated that small variations in w did not affect classification accuracy. Note that for a single window the histogram will be sparse, but the statistics of multiple histograms will not be. The POS tags are a modified version of the Penn Treebank set [16], where we have collapsed plural and singular forms, included four different punctuation markers, and added a few words that tend to be indicative of conversational or informal speech (such as "yeah", "ok", and "uh") for a total of $K = 36$ POS tags. For the feature reduction, we retain all PC dimensions with variance above 1% of the maximum PC variance.

3. EXPERIMENTS

3.1. Experimental details

We compared the three discussed feature extraction methods: top 10000 word-frequency counts, top 1000 POS trigram counts, and the proposed POS histogram statistics on a problem of distinguishing six genre classes useful for language modeling: broadcast news (bn, 671 docs), broadcast conversations (bc, 698 docs), meetings (mt, 493 docs), news wire (nw, 471 docs), switchboard (sb, 890 docs), and weblogs (wl, 543 docs). The majority of the documents are 600 to 1000 elements in length. We used the Ratnaparkhi maximum entropy tagger [17], and the same 36 POS tags for both the POS histogram and POS trigram features. In order to have an estimate of the variance of the classifier, the data set for each class was randomly split 75/25 into training/test sets, and the random split was repeated 50 times. Typically, the reduced dimension of the POS histogram feature space is around 30 (out of a maximum of 72), after performing PCA. The classifier used is the standard generative classifier quadratic discriminant analysis (QDA) with full covariance matrices estimated by maximum likelihood, and trained on the reduced-dimension POS histogram features. In addition, to ensure that the improvement in classification is due to the features, we also implemented a naive Bayes classifier (diagonalized QDA) with the POS histogram features. The QDA approach is motivated by the distributional characteristics observed in Figure 1, which is an example PCA projection from one particular training/test split onto the top three ranked components.

We implemented two baselines for comparison. One approach was modeled after Stamatatos et al. [9], which used

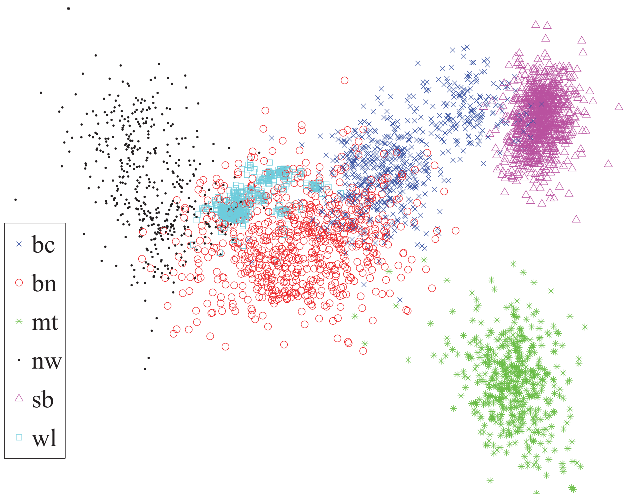

Fig. 1. Two-dimensional PCA projection of the top three ranked principal components for the POS histogram features.

	% correct	% std
QDA with POS histograms	98.42	0.40
naive Bayes with POS histograms	96.38	0.49
naive Bayes with word frequencies	95.19	0.52
naive Bayes with POS trigrams	89.31	0.85

Table 1. Classification Performance.

	bc	bn	mt	nw	sb	wl
bc	97.4	2.6	0	0	0	0
bn	0.5	99.5	0	0	0	0
mt	0	0	100	0	0	0
nw	0	0.4	0	99.6	0	0
sb	0.2	0	0.4	0	99.4	0
wl	0	4.5	0	1.2	0	94.3

Table 2. Confusion matrix for POS histogram featurees, with QDA.

	bc	bn	mt	nw	sb	wl
bc	98.4	0.2	0	0	0.9	0.5
bn	0.2	93.8	0	1.4	0	4.7
mt	0	0	99.9	0	0.1	0
nw	1.0	3.4	0	81.9	0	13.7
sb	0.4	0	0.1	0	99.5	0
wl	2.1	3.0	0.2	1.2	0.3	93.1

Table 3. Confusion matrix for word frequency features, with naive Bayes.

	bc	bn	mt	nw	sb	wl
bc	91.3	3.9	0.3	0.1	0.5	3.9
bn	4.4	78.1	0	3.1	0	14.3
mt	0.1	0	99.2	0.7	0	0
nw	0	0.7	0	83.9	0	15.4
sb	0	0	0	0	100.0	0
wl	2.4	10.3	0.2	7.8	0	79.3

Table 4. Confusion matrix for POS trigram features, with naive Bayes.

counts of the most frequently occurring words, and then classified with naive Bayes. We compare to this method, though our comparison differs in the use of information gain rather than word frequency for selecting the terms to retain, in order to have a more fair comparison to our own approach. POS n-grams are another standard classification feature for this problem, and thus we also compare to Santini's approach, which used top-ranked POS trigrams followed by a naive Bayes classifier [6]. For the POS trigrams and word frequency comparisons, we pruned all but the top 1000 and 10000 dimensions using the information gain method, and both were implemented with the Rainbow toolkit [18].

3.2. Results on Training Genres

Classification results are in Table 1, averaged over the 50 random training/test splits. The confusion matrices are shown in Tables 2, 3, and 4. The (ith,jth) entry is the percent of documents from class i classified as class j, on average. We omit the confusion matrix for naive Bayes with POS histograms due to space constraints. Even though the word frequency features yield higher accuracy on the broadcast conversations, news wire becomes confused with weblogs, and broadcast news performance worsens. The POS trigram features have particular difficulties classifying weblogs and broadcast news correctly.

3.3. Classifying New Genres

As an experiment of the POS histogram genre classifier's ability to handle new genres, we extracted POS histogram features for 10 speech transcripts each from Barack Obama and John McCain, and plotted the features' location in the same PCA projection as mentioned previously; the plot is shown, zoomed in, in Figure 2. In the 31 dimensional classification space, all ten speeches (shown as purple and brown dots) were classified as the broadcast news class, as one would expect given the classes used in training. The visual separability of the Obama and McCain speeches leads us to hypothesize that the POS histogram features may be used for distinguishing finer-grain genres, but training with text from more genres would probably be required for such an analysis.

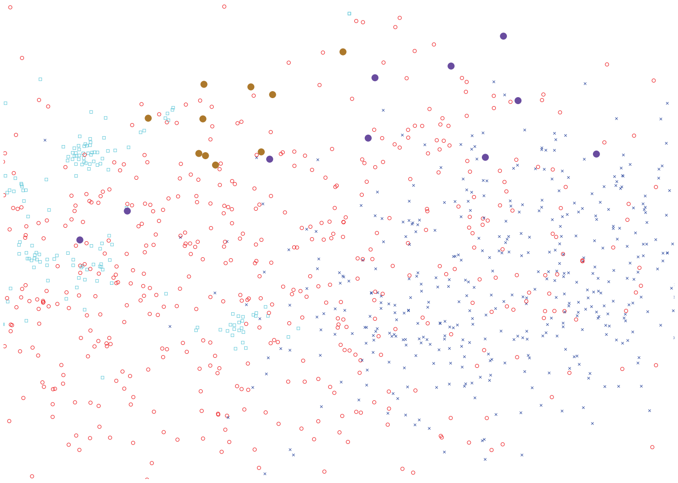

Fig. 2. Zoomed-in two-dimensional PCA projection of the top three ranked principal components illustrating Obama (purple dots) and McCain (brown dots) speeches; the cyan squares are weblogs, the red dots are broadcast news, and the blue crosses are broadcast conversations.

4. CONCLUSIONS AND OPEN QUESTIONS

Our preliminary studies suggest that the results are robust to changes in the different parameters involved in the feature extraction, including the feature reduction threshold and using higher-order POS histogram statistics. One research direction is to investigate the impact of additional non-POS word features in the histogram.

In Figure 1 it is clear that even the top-variance two-dimensional feature space is not well-covered by these six classes. For example, there is a clear gap in the feature space between meetings and broadcast news or broadcast conversation that we hypothesize would be filled by spontaneous but formal speech such as lectures. It is not clear how many training classes or which ones would be needed to effectively cover the entire feature space, and differentiating subclasses of these six classes could be helpful depending on the application. We hypothesize that in general a better approach might be to cast genre classification as a multi-task learning task, replacing singular genre classes with multiple factors such as purpose, formality, spontaneous vs. pre-planned, size of audience, etc. Such a factorization would allow one to characterize genre in a more generalizable way, so as to perform classification for documents that do not easily conform to a genre class, such as those treated in this paper. It may also help with characterizing genres that include variations, such as formal and informal styles of written reviews.

5. REFERENCES

[1] R. Iyer and M. Ostendorf, “Relevance weighting for combining multi-domain data for n-gram language modeling,” *Computer Speech and Language*, vol. 13, no. 3, pp. 267–282, 1999.

[2] I. Bulyko, M. Ostendorf, and A. Stolcke, “Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures,” in *Proc. HLT/NAACL*, 2003, pp. 7–9.

[3] I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and Ö. Cetin, “Web resources for language modeling in conversational speech recognition,” *ACM Trans. on Speech and Language Processing*, vol. 5, no. 1, pp. 1–25, 2007.

[4] A. Sethy, P. Georgiou, and S. Narayanan, “Building topic-specific language models from webdata using competitive models,” in *Proc. Interspeech*, 2005, pp. 1293–1296.

[5] R. Sarikaya, A. Gravano, and Y. Gao, “Rapid language model development using external resources for new spoken dialog domains,” in *Proc. ICASSP*, 2005, vol. I, pp. 573–576.

[6] M. Santini, “A shallow approach to syntactic feature extraction for genre classification,” *CLUK 7: The UK special-interest group for computational linguistics*, 2004.

[7] C. S. Lim, K. J. Lee, and G. C. Kim, “Automatic genre detection of web documents,” in *IJCNLP*, 2004.

[8] B. Kessler, G. Numberg, and H. Schütze, “Automatic detection of text genre,” in *ACL-35*, 1997, pp. 32–38.

[9] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, “Text genre detection using common word frequencies,” in *COLING*, 2000, pp. 808–814.

[10] J. Karlgren and D. Cutting, “Recognizing text genres with simple metrics using discriminant analysis,” in *Proc. Computational Linguistics*, 1994, pp. 1071–1075.

[11] Y.-B. Lee and S. H. Myaeng, “Text genre classification with genre-revealing and subject-revealing features,” in *ACM SIGIR*, 2002, pp. 145–150.

[12] M. A. Marin, S. Feldman, M. Ostendorf, and M. Gupta, “Filtering web text to match target genres,” in *Proc. ICASSP*, 2009.

[13] J. Goodman, “A bit of progress in language modeling,” *Computer Speech and Language*, vol. 15, no. 4, pp. 403–434, 2001.

[14] D. Biber, “Using register-diversified corpora for general language studies,” *Computational Linguistics*, vol. 19, no. 2, pp. 219–242, 1993.

[15] S. Schwarm, I. Bulyko, and M. Ostendorf, “Adaptive language modeling with varied sources to cover new vocabulary items,” *IEEE Trans. Speech and Audio*, vol. 12, no. 3, pp. 334–342, 2004.

[16] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, “Building a large annotated corpus of English: the Penn Treebank,” *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[17] A. Ratnaparkhi, “A maximum entropy part-of-speech tagger,” in *Proc. Empirical Methods in Natural Language Processing Conference*, 1996, pp. 133–141.

[18] A. K. McCallum, “BOW: A toolkit for statistical language modeling, text retrieval, classification and clustering,” http://www.cs.cmu.edu/ mccallum/bow, 1996.