Classifying Factored Genres with Part-of-Speech Histograms

S. Feldman, M. Marin, J. Medero, and M. Ostendorf
Dept. of Electrical Engineering
University of Washington, Seattle, Washington 98195
{sergeyf,amarin,jmedero,ostendor}@u.washington.edu

Abstract

This work addresses the problem of genre classification of text and speech transcripts, with the goal of handling genres not seen in training. Two frameworks employing different statistics on word/POS histograms with a PCA transform are examined: a single model for each genre and a factored representation of genre. The impact of the two frameworks on the classification of training-matched and new genres is discussed. Results show that the factored models allow for a finer-grained representation of genre and can more accurately characterize genres not seen in training.

1 Introduction

With increasing quantities of text and transcribed speech available online, the ability to categorize documents based on characteristics beyond topic becomes ever more important. In particular, the genre of a document – whether it is a news report or an editorial, a speech transcript or a weblog – may be relevant for many human 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 human language technologies, such as speech recognition, parsing, 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, e.g. (Wang, 2008). Since text harvested from the web is increasingly used to address problems due to sparse training data, genre classification can be useful for sampling such text sources to obtain a better match to the target domain for offline language model training. Prior work on genre-dependent web text filtering for language modeling relied on standard search engine methods, designing queries based on frequent n-grams in the domain, e.g. (Bulyko et al., 2007). However, as the variety of genres online has grown, this method has become less reliable. This work addresses explicit genre classification, with the assumption that genre representation in the training data is incomplete.

In prior work on genre classification, an important question has been the definition of "genre." For many studies, genre has been associated with categories of text, such as research article, novel, news report, editorial, advertisement, etc. In particular, several studies use classes identified in the Brown corpus or the British National Corpus. Spoken genres, including conversation, interview, debate, and planned speech are considered in (Santini, 2004). Examples of spoken and written genres, represented in several corpora available from the Linguistics Data Consortium, are explored in (Feldman et al., 2009). Yet another study focuses on internet-specific document types, including different types of home pages (personal, public, commercial), bulletin boards, and link lists (Lim et al., 2004). A limitation of all of this work is that only a small, fixed set of different genres are explored, with performance assessed on matched data. In this paper, we assess classification results of texts that come from new genres, as well as those matching the training set.

In addressing new genres, we have two main contributions: new features and factored coding.

Proceedings of NAACL HLT 2009: Short Papers, pages 173–176,
Boulder, Colorado, June 2009.

The standard features for genre classification models include words, part-of-speech (POS) tags, and punctuation (Kessler et al., 1997; Stamatatos et al., 2000; Lee and Myaeng, 2002; Biber, 1993), but constituent-based syntactic categories have also been explored (Karlgren and Cutting, 1994). (Feldman et al., 2009) used mixed word and POS histogram mean and variance as features for genre classification. In this work, we augment those histogram statistics with higher-order ones, as well as add new word features aimed at capturing online genres. Further, we propose a factored genre model, and demonstrate its effect on genre classification of out-of-domain documents.

2 Methods

2.1 Corpora

To train our algorithm, we use eight different genres: broadcast news (bn, 671 docs), broadcast conversations (bc, 698 docs), meetings (mt, 493 docs), newswire (nw, 471 docs), conversational telephone speech (sb, 890 docs), weblogs (wl, 543 docs), Amazon reviews of books, videogames and films (az train, 218 docs), and chat data (chat, 187 docs). To test our algorithm, we add six additional genres: Amazon reviews of appliances (az test, 27 docs), Wikipedia entries (wiki, 254 docs), Wikipedia discussion entries (wiki talk, 1792 docs), European Parliament transcripts (europarl, 1423 docs), a web collection obtained from Google searches for common conversational n-grams (web, 18540 docs), and transcribed McCain and Obama speeches (speeches, 20 docs). With the exception of the chat data, Amazon reviews, and a subset of the Europarl transcripts, the training corpora are from standard published datasets. The reviews, chat, Wikipedia, and web data were all collected from websites and cleaned locally. The documents average 600-1000 words in length, except for smaller corpora like Amazon reviews, whose documents average about 200 words. For training factored models, we assume that all the documents within a corpus share the same class.

2.2 Features and Classifier

The features used in (Feldman et al., 2009) were derived from a union of POS tags and a set of hand-picked, informative words. A similar approach is used here, including a collapsed version of the Treebank POS tag set (Marcus et al., 1993), with additions for specific words (e.g. personal pronouns and filled pause markers), compound punctuation (e.g. multiple exclamation marks), and a general emoticon tag, resulting in a total of 41 tags. Histograms are computed for a sliding window of length $w = 5$ over the tag sequence, and then statistics of each histogram bin are extracted. In the previous work, mean and standard deviation were extracted from the histogram bins. To this, we add skewness and kurtosis, which we will show are necessary for increased differentiation of unseen genres. For feature reduction, we used Principal Components Analysis and retained all PC dimensions with variance above 1% of the maximum PC variance.

Different approaches have been explored for computational modeling, including naive Bayes, linear discriminant modeling, and neural networks (Santini, 2004; Kessler et al., 1997; Stamatatos et al., 2000; Lee and Myaeng, 2002). Since (Feldman et al., 2009) found that quadratic discriminant analysis (QDA) outperforms naive Bayes, we use it here with full covariance matrices estimated by maximum likelihood, and trained on the reduced-dimension POS histogram features.

2.3 Factors

Linguistic research has tended to look at attributes of language rather than defining genre in terms of task domains. Since the number of task domains appears to be growing with new uses of the internet, we conjecture that an attribute approach is more practical for web-based text. We introduce the notion of a factored model for genre. The genre of each document can encoded as a vector of factors. Given data limits, the set of factors explored so far are:

- number of speakers/authors (1,2,3+),
- level of formality (low, medium, high),
- intended audience (personal, broadcast), and
- intent (inform, persuade).

Assuming factor independence, we train four separate QDA classifiers, one per factor. Using factors increases the richness of the space represented by the training set, in that it is possible to identify genres with factor combinations not seen in training.

3 Experiments and Discussion

3.1 Within-Domain Validation

As a preliminary step, and to ensure that the addition of skewness and kurtosis, as well as extra syntactic features, does not significantly impact the within-domain classification accuracy, we performed experiments with both the features in (Feldman et al., 2009) and our expanded features. For this, we split the training data 75/25 into training/test sets, and repeated the random split 50 times. We ran the experiments for both the original genre classification problem and the individual factors. We found that the addition of new moments and features decreased performance by less than 1% on average. We hypothesize that this small deterioration in performance is likely due to overtuning to the original training set.

3.2 New Features with Unseen Genres

To assess the use of our new features (added punctuation and emoticons) and the higher-order moments, we classified the web data with different processing configurations. In addition to the eight training genres, we introduced an "undetermined" genre or class for documents with a uniform posterior probability across all genres, which occurs when there is a large mismatch to all training genres. The distribution of labels is shown in Figure 1. While we do not have hand-labeled categories for this data, we thought it highly unlikely that the vast majority is bn, as predicted by the models using only mean and variance moments.

To validate our hypothesis that the spread of labels was more appropriate for the data, we randomly selected 100 documents and hand-labeled these using the eight classes plus "undetermined." The undetermined class was used for new genres (play scripts, lectures, newsgroups, congressional records). We found that it was difficult to annotate the data, since many samples had characteristics of more than one genre; this finding motivates the factor representation. The main difference between the various feature extraction configurations was in the detection of the undetermined cases. For the subset of undetermined documents that we labeled (34), none were detected using only 2 moments, but 35-40% were detected with the higher-order moments. Of the false detections, roughly 25-30% were associated with documents with characteristics of multiple classes. The effect of adding more detailed punctuation and emoticons to the tag set was not significant.

It should be noted that the web collection was based on queries designed to extract BC-style text, yet only 3 of 100 hand-labeled samples were in that category, none of which were accurately classified. Roughly 16 of the 100 documents are labeled as very informal and another 55 include some informal text or are moderately informal. This finding, combined with the observation that many documents reflect a mix of genres, suggests that a factored representation of genre (with formality as one "factor") may be more useful than explicit modeling of genres.

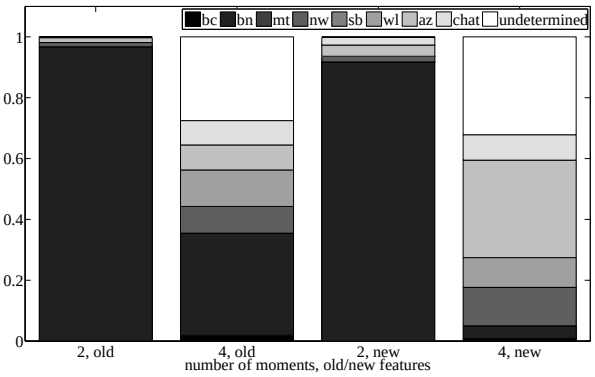

Figure 1: Fraction of web data classified as each genre.

3.3 Unseen Genre Factor Results

We trained a set of models for each factor and obtained posterior estimates for unseen classes. Figure 2 shows the class of out-of-domain documents for the formality factor, using 3 categories of formality: low (conversational, unprofessional), medium (casual but coherent), high (formal). We have not hand-labeled individual documents in all of these sets, but the resulting class proportions match our intuition for these genres. The Wikipedia data is labeled as highly formal, and most web data is labeled as medium. Examining the 100 hand-labeled web documents, we find that adding the higher-order moments improves classifier accuracy from 23% to 55%. The effect of the added tag set features was once again not significant.

Figure 3 shows the class of out-of-domain documents for the factor indicating number of speakers/authors. This factor appears difficult to detect.

We hypothesize that there is an unaccounted-for dependence on audience. When there is a listener, speakers may use the term "you," as in conversations and internet chat. An interesting observation is that the ten Obama speeches all appear to exhibit this behavior. McCain speeches, on the other hand, display some variation, and about a third are (correctly) characterized as single speaker.

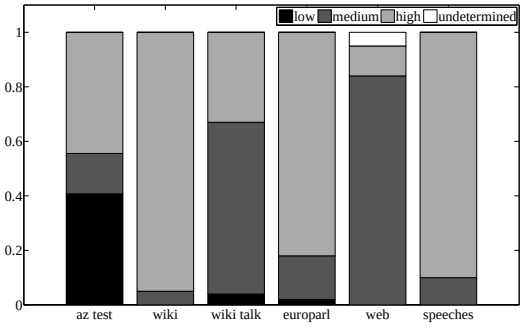

Figure 2: Test corpora classification, formality.

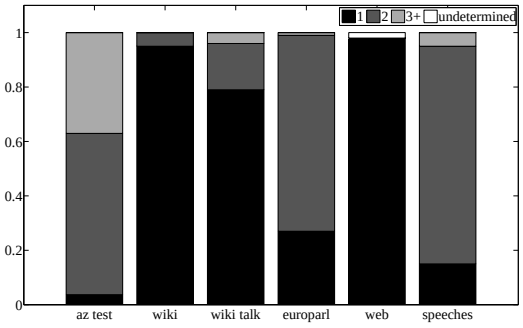

Figure 3: Test corpora classification, number of speakers.

The audience factor results are very skewed towards broadcast data, but this matches our intuition, and the scarcity of data meant for private consumption, so they are not included. However, further study is needed, since 3-dimensional projections of the training data suggest a Gaussian mixture (or other more complex model) may fit better.

The intent factor results are also mixed. The classifier labels most of the Wikipedia, europarl, web, and speeches data as "report," and most reviews as "persuade." While the "report" category fits Wikipedia, it is not clear that europarl should also be classified as "report," since parliamentary proceedings are notoriously argumentative. With this factor, the noise inherent in using genre-level labels is significant. It is not always clear what is reportage and what is persuasion, and we expect some genres (e.g. reviews) to be a mixture of both.

4 Summary

We have introduced new features that are more robust for handling domains unseen in training, and presented a factored genre framework that allows for a finer-grained representation of genre. Many open questions remain, including which other factors can or cannot be captured by our current feature set and classifier, and whether noisy label learning methods could address the problem of uncertainty in the labels for particular features and genres.

References

D. Biber. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–242.

I. Bulyko et al. 2007. Web resources for language modeling in conversational speech recognition. *ACM Trans. on Speech and Language Processing*, 5(1):1–25.

S. Feldman et al. 2009. Part-of-speech histograms for genre classification of text. *Proc. ICASSP*.

W. Wang. 2008. Weakly supervised training for parsing mandarin broadcast transcripts. *Proc. Interspeech.*

J. Karlgren and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. *Proc. Computational Linguistics*, pages 1071–1075.

B. Kessler, G. Numberg, and H. Schütze. 1997. Automatic detection of text genre. *ACL-35*, pages 32–38.

Y.-B. Lee and S. H. Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. *ACM SIGIR*, pages 145–150.

C. S. Lim, K. J. Lee, and G. C. Kim. 2004. Automatic genre detection of web documents. *IJCNLP*.

M. P. Marcus et al. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M. Santini. 2004. A shallow approach to syntactic feature extraction for genre classification. *CLUK 7: UK special-interest group for computational linguistics*.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2000. Text genre detection using common word frequencies. *Proc. Computational Linguistics*, pages 808–814.