Modeling of Stylistic Variation in Social Media with Stretchy Patterns

Philip Gianfortoni
Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA
pwg@cs.cmu.edu

David Adamson
Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA
dadamson@cs.cmu.edu

Carolyn P. Rosé
Carnegie Mellon University
Language Technologies Institute
Pittsburgh, PA
cprose@cs.cmu.edu

Abstract

In this paper we describe a novel feature discovery technique that can be used to model stylistic variation in sociolects. While structural features offer much in terms of expressive power over simpler features used more frequently in machine learning approaches to modeling linguistic variation, they frequently come at an excessive cost in terms of feature space size expansion. We propose a novel form of structural features referred to as "stretchy patterns" that strike a balance between expressive power and compactness in order to enable modeling stylistic variation with reasonably small datasets. As an example we focus on the problem of modeling variation related to gender in personal blogs. Our evaluation demonstrates a significant improvement over standard baselines.

1 Introduction

The contribution of this paper is a novel approach to feature induction seeking to model stylistic variation at a level that not only achieves high performance, but generalizes across domains better than alternative techniques. Building on an earlier template based approach for modeling sarcasm (Tsur et al., 2010), we investigate the use of what we have termed "stretchy" features to model stylistic variation related to sociolects, which can be thought of as a form of dialect. Specifically, we focus on the problem of gender based classification. Gender classification and age classification have both received increased attention in the social media analysis community in recent years (Goswami et al., 2009; Barbieri, 2008; Cieri et al., 2004), most likely because large data sets annotated with these variables have recently become available. Machine learning technology provides a lens with which to explore linguistic variation that complements earlier statistical techniques used by variationist sociolinguists in their work mapping out the space of dialect variation and its accompanying social interpretation (Labov, 2010a; Labov, 2010b; Eckert & Rickford, 2001). These complementary approaches share a common foundation in numerical methods, however while descriptive statistics and inferential statistics mainly serve the purpose of describing non-random differences in distributions between communities, machine learning work in the area of social media analysis asks the more challenging question of whether the differences described are big enough to enable identification of community membership by means of those differences.

In the remainder of the paper, we first introduce prior work in a variety of related areas that both demonstrates why generalizable models characterizing sociolects within social media contexts are challenging to create and motivates our novel approach. Next we describe our technical approach for inducing "stretchy patterns". We then present a series of experiments that demonstrate that our stretchy patterns provide advantages over alternative feature spaces in terms of avoiding overfitting to irrelevant content-based features as evidenced both in terms of achieving higher performance with smaller amounts of training data and in terms of generalizing better across subpopulations that share other demographic and individual difference variables.

2 Prior Work

Analysis of social media has grown in popularity over the past decade. Nevertheless, results on problems such as gender classification (Argamon et al., 2003), age classification (Argamon et al., 2007), political affiliation classification (Jiang & Argamon, 2008), and sentiment analysis (Wiebe et al., 2004) demonstrate how difficult stylistic classification tasks can be, and even more so when the generality is evaluated by testing models trained in one domain on examples from another domain. Prior work on feature engineering has attempted to address this generalization difficulty. Here we motivate our "stretchy pattern" approach to feature engineering for modeling sociolects, using gender analysis as a lens through which to understand the problem.

2.1 Variation Analysis and Gender

Since the earliest work in the area of variationist sociolinguistics, gender has been a variable of interest, which explains interesting differences in communication style that have been the topic of discussion both in academic circles (Holmes & Meyerhoff, 2003) and in the popular press (Tannen, 2001). The immense significance that has been placed on these differences, whether they are viewed as essentially linked to inherent traits, learned cultural patterns, or socially situated identities that are constructed within interaction, warrants attention to gender based differences within the scope of dialect variation. While one may view gender differences in communication from multiple angles, including topic, stance, and style, we focus specifically on linguistic style in our work.

Numerous attempts to computationally model gender based language variation have been published in the past decade (Corney et al., 2002; Argamon et al., 2003; Schler et al., 2005; Schler, 2006; Yan & Yan, 2006; Zhang et al., 2009; Mukherjee & Liu, 2010). Gender based language variation arises from multiple sources. For example, within a single corpus comprised of samples of male and female language that the two genders do not speak or write about the same topics. This has been reported to be the case with blog corpora such as the one used in this paper. Even in cases where pains have been taken to control for the distribution of topics associated with each gender within a corpus (Argamon et al., 2003), it's still not clear the extent to which that distribution is completely controlled. For example, if one is careful to have equal numbers of writing samples related to politics from males and females, it may still be the case that males and females are discussing different political issues or are addressing political issues from a different role based angle. While these differences are interesting, they do not fit within the purview of linguistic style variation.

Word based features such as unigrams and bigrams are highly likely to pick up on differences in topic (Schler, 2006) and possibly perspective. Thus, in cases where linguistic style variation is specifically of interest, these features are not likely to be included in the set of features used to model the variation even if their use leads to high performance within restricted domains. Typical kinds of features that are used instead include part-of-speech (POS) n-grams (Koppel, 2002; Argamon et al., 2003), word structure features that cluster words according to endings that indicate part of speech (Zhang et al., 2009), features that indicate the distribution of word lengths within a corpus (Corney et al., 2002), usage of punctuation, and features related to usage of jargon (Schler et al., 2005). In Internet-based communication, additional features have been investigated such as usage of internet specific features including "internet speak" (e.g., lol, wtf, etc.), emoticons, and URLs (Yan & Yan, 2006). In addition to attention to feature space design issues, some work on computational modeling of gender based language variation has included the development of novel feature selection techniques, which have also had a significant impact on success (Mukherjee & Liu, 2010; Zhang, Dang, & Chen, 2009).

Of these features, the only ones that capture stylistic elements that extend beyond individual words at a time are the POS ngram features. The inclusion of these features has been motivated by their hypothesized generality, although in practice, the generality of gender prediction models has not been formally evaluated in the gender prediction literature.

2.2 Domain Adaptation in Social Media

Recent work in the area of domain adaptation (Arnold et al., 2008; Daumé III, 2007; Finkel & Manning, 2009) raises awareness of the difficulties with generality of trained models and offers insight into the reasons for the difficulty with generalization. We consider these issues specifically in connection with the problem of modeling gender based variation.

One problem, also noted by variationist sociolinguists, is that similar language variation is associated with different variables (McEnery, 2006). For example, linguistic features associated with older age are also more associated with male communication style than female communication style for people of the same age (Argamon et al., 2007). Another problem is that style is not exhibited by different words than those that serve the purpose of communicating content. Thus, there is much about style that is expressed in a topic specific way.

What exacerbates these problems in text processing approaches is that texts are typically represented with features that are at the wrong level of granularity for what is being modeled. Specifically, for practical reasons, the most common types of features used in text classification tasks are still unigrams, bigrams, and part-of-speech bigrams. While relying heavily on these relatively simple features has computational advantages in terms of keeping the feature space size manageable, which aids in efficient model learning, in combination with the complicating factors just mentioned, these text classification approaches are highly prone to over-fitting.

Specifically, when text is represented with features that operate at too fine grained of a level, features that truly model the target style are not present within the model. Thus, the trained models are not able to capture the style itself and instead capture features that merely correlate with that style within the data. Thus, in cases where the data is not independent and identically distributed (IID), and where instances that belong to different subpopulations within the non-IID data have different class value distributions, the model will tend to give weight to features that indicate the subpopulation rather than features that model the style. This may lead to models that perform well within datasets that contain the same distribution of subpopulations, but will not generalize to different subpopulations, or even datasets composed of different proportions of the same subpopulations. Models employing primarly unigrams and bigrams as features are particularly problematic in this respect.

2.3 Automatic Feature Engineering

In recent years, a variety of manual and automatic feature engineering techniques have been developed in order to construct feature spaces that are adept at capturing interesting language variation without overfitting to content based variation, with the hope of leading to more generalizable models.

POS n-grams, which have frequently been utilized in genre analysis models (Argamon et al., 2003), are a strategic balance between informativity and simplicity. They are able to estimate syntactic structure and style without modeling it directly. In an attempt to capture syntactic structure more faithfully, there has been experimentation within the area of sentiment analysis on using syntactic dependency features (Joshi & Rosé, 2009; Arora, Joshi, & Rosé, 2009). However, results have been mixed. In practice, the added richness of the features comes at a tremendous cost in terms of dramatic increases in feature space size. What has been more successful in practice is templatizing the dependency features in order to capture the same amount of structure without creating features that are so specific.

Syntactic dependency based features are able to capture more structure than POS bigrams, however, they are still limited to representing relationships between pairs of words within a text. Thus, they still leave much to be desired in terms of representation power. Experimentation with graph mining from dependency parses has also been used for generating rich feature spaces (Arora et al., 2010). However, results with these features has also been disappointing. In practice, the rich features with real predictive power end up being difficult to find amidst myriads of useless features that simply add noise to the model. One direction that has proven successful at exceeding the representational power and performance of POS bigrams with only a very modest increase in feature space size has been a genetic programming based approach to learning to build a strategic set of rich features so that the benefits of rich features can be obtained without the expense in terms of feature space expansion. Successful experiments with this technique have been conducted in the area of sentiment analysis, with terminal symbols including unigrams in one case (Mayfield & Rosé, 2010) and graph features extracted from dependency parses in another (Arora et al., 2010). Nevertheless, improvements using these strategic sets of evolved features have been very small even where statistically significant, and thus it is difficult to justify adding so much machinery for such a small improvement.

Another direction is to construct template based features that combine some aspects of POS n-grams in that they are a flat representation, and the backoff version of dependency features, in that the symbols represent sets of words, which may be POS tags, learned word classes, distribution based word classes (such as high frequency words or low frequency words), or words. Such types of features have been used alone or in combination with sophisticated feature selection techniques or bootstrapping techniques, and have been applied to problems such as detection of sarcasm (Tsur et al., 2010), detection of causal connections between events (Girju, 2010), or machine translation (Gimpel et al., 2011). Our work is most similar to this class of approaches.

3 Technical Approach: Stretchy Patterns

Other systems have managed to extract and employ patterns containing gaps with some success. For example, Gimpel (2011) uses Gibbs sampling to collect patterns containing single-word gaps, and uses them among other features in a machine translation system.

Our patterns are more like the ones described in Tsur (2010), which were applied to the task of identifying sarcasm in sentences. We predicted that a similar method would show promise in extracting broader stylistic features indicative of the author's group-aligned dialect. We have chosen the classification of an author's gender as the task to which we can apply our patterns.

3.1 Pattern-Based Features

To extract their sarcasm-detecting patterns, Tsur (2010) first defined two sets of words: High Frequency Words (HFW) and Content Words (CW). The HFW set contained all words that occurred more than 100 times per million, and the CW set contained all words in the corpus that occurred fewer than 1000 times per million. Thus, a word could be contained in the HFW set, the CW set, or both. Such patterns must begin and end with words in the HFW set, and (as in our implementation) are constrained in the number of words drawn from each set. Additionally, as a preprocessing step, in their approach they made an attempt to replace phrases belonging to several categories of domain-specific phrases, such as product and manufacturer names with a label string, which was then added to the HFW set, indicating membership. For example, given an input such as "Garmin apparently does not care much about product quality or customer support", a number of patterns would be produced, including "[company] CW does not CW much".

3.2 Stretchy Patterns

Tsur's patterns were applied as features to classify sentences as sarcastic (or not), within the domain of online product reviews. Here our implementation and application diverge from Tsur's — the blog corpus features large multi-sentence documents, and span a diverse set of topics and authors. We aim to use these patterns not to classify sentiment or subtlety, but to capture the style and structure employed by subsets of the author-population.

We define a document as an ordered list of tokens. Each token is composed of a surface-form lexeme and any additional syntactic or semantic information about the word at this position (in our case this is simply the POS tag, but other layers such as Named Entity might be included). We refer to any of the available forms of a token as a *type*.
A category is a set of word-types. Each type must belong to at least one category. All categories have a corresponding label, by which they'll be referred to within the patterns to come. *Gap* is a special category, containing all types that aren't

part of any other category. The types belonging to any defined category may also be explicitly added to the Gap category.

A *stretchy pattern* is defined as a sequence of categories, which must not begin or end with a Gap category. We designate any number of adjacent Gap instances in a pattern by the string "GAP+"[1] and every other category instance by its label. As a convention, the label of a singleton category is the name of the type contained in the category (thus "writes" would be the label of a category containing only surface form "writes" and "VBZ" would be the label of the a category containing only the POS tag "VBZ"). The overall number of Gap and non-Gap category instances comprising a pattern is restricted - following Tsur (2010), we allow no more than six tokens of either category. In the case of Gap instances, this restriction is placed on the number of underlying tokens, and not the collapsed GAP+ form.

A sequence of tokens in a document matches a pattern if there is some expansion where each token corresponds in order to the pattern's categories. A given instance of GAP+ will match between zero and six tokens, provided the total number of Gap instances in the pattern do not exceed six[2].

By way of example, two patterns follow, with two strings that match each. Tokens that match as Gaps are shown in parenthesis.

[*cc*] (GAP+) [*adj*] [*adj*]
"and (some clients were) *kinda popular...*"
"*from* (our) *own general* election..."

for (GAP+) [*third-pron*] (GAP+) [*end*] [*first-pron*]
"ready *for* () *them* (to end) . *I* am..."
"*for* (murdering) *his* (prose) . *i* want…"

Although the matched sequences vary in length and content, the stretchy patterns preserve information about the proximity and ordering of particular words and categories. They focus on the relationship between key (non-Gap) words, and allow a wide array of sequences to be matched by a single pattern in a way that traditional word-class n-grams would not.

Our "stretchy pattern" formalism strictly subsumes Tsur's approach in terms of representational power. In particular, we could generate the same patterns described in Tsur (2010) by creating a singleton surface form category for each word in Tsur's HFW and then creating a category called [CW] that contains all of the words in the Tsur CW set, in addition to the domain-specific product/manufacturer categories Tsur employed.

Label	Category Members
adj	JJ, JJR, JJS
cc	CC, IN
md	MD
end	<period>, <comma>, <question>, <exclamation>
first-pron	I, me, my, mine, im, I'm
second-pron	you, your, youre, you're, yours, y'all
third-pron	he, him
emotional	feel, hurt, lonely, love
time	hour, hours, late, min, minute, minutes, months, schedule, seconds, time, years,
male_curse	fucking, fuck, jesus, cunt, fucker
female_curse	god, bloody, pig, hell, bitch, pissed, assed, shit

Table 1. Word Categories

3.3 Word Categories

With the aim of capturing general usage patterns, and motivated by the results of corpus linguists and discourse analysts, a handful token categories were defined, after the fashion of the LIWC categories as discussed in Gill (2009). Tokens belonging to categories may be replaced with their category label as patterns are extracted from each document. As a token might belong to multiple categories, the same token sequence may generate, and therefore match multiple patterns.

Words from a list of 800 common prepositions, conjunctions, adjectives, and adverbs were included as singleton surface-form categories. Determiners in particular are absent from this list (and from the POS categories that follow), as their absence or presence in a noun phrase is one of the primary variations the stretchy gaps of our patterns were intended to smooth over.

A handful of POS categories were selected, reflecting previous research and predictions about gender differences in language usage. For example, to capture the "hedging" discussed in Holmes (2003) as more common in female speech, the modal tag MD was included as a singleton category. A category comprising the coordinating

[1] *This is actually an extractor parameter, but we collapse all adjacent gaps for all our experiments.*

[2] *The restrictions on gaps are extractor parameters, but we picked zero to six gaps for our experiments.*

conjunction and preposition tags (CC, IN) was included to highlight transitions in complicated or nested multi-part sentences.

Additionally, where previous results suggested variation within a category based on gender (e.g. swearing, as in McEnery (2006)), two categories were added, with the words most discriminative for each gender. However, even those words most favored by male authors might appear in contexts where males would never use them - it is our hope that by embedding these otherwise-distinguishing features within the structure afforded by gap patterns we can extract more meaningful patterns that more accurately and expressively capture the style of each gender.

3.4 Extraction and Filtering

Patterns are extracted from the training set, using a sliding window over the token stream to generate all allowable combinations of category-gap sequences within the window. This generates an exponential number of patterns - we initially filter this huge set based on each pattern's accuracy and coverage as a standalone classifier, discarding those with less than a minimum precision or number of instances within the training set. In the experiments that follow, these thresholds were set to a minimum of 60% per-feature precision, and at least 15 document-level hits.

4 Evaluation

We have motivated the design of our stretchy patterns by the desire to balance expressive power and compactness. The evidence of our success should be demonstrated along two dimensions: first, that these compact features allow our models to achieve a higher performance when trained on small datasets and second, that models trained with our stretchy patterns generalize better between domains. Thus, in this section, we present two evaluations of our approach in comparison to three baseline approaches.

4.1 Dataset

We chose to use the Blog Authorship Corpus for our evaluation, which has been used in earlier work related to gender classification (Schler 2006), and which is available for web download[3]. Each instance contains a series of personal blog entries from a single author. For each blog, we have metadata indicating the gender, age, occupation, and astrological sign of the author. From this corpus, for each experiment, we randomly selected a subset in which we have balanced the distribution of gender and occupation. In particular, we selected 10 of the most common occupations in the dataset, specifically Science, Law, Non-Profit, Internet, Engineering, Media, Arts, Education, Technology, and Student. We randomly select the same number of blogs from each of these occupations, and within occupation based sets, we maintain an even distribution of male and female authors. We treat the occupation variable as a proxy for topic since bloggers typically make reference to their work in their posts. We make use of this proxy for topic in our evaluation of domain generality below.

4.2 Baseline Approaches

We can find in the literature a variety of approaches to modeling gender based linguistic variation, as outlined in our prior work discussion above. If our purpose was to demonstrate that our stretchy patterns beat the state-of-the-art at the predictive task of gender classification, it would be essential to implement one of these approaches as our baseline. However, our purpose here is instead to address two more specific research questions instead, and for that we argue that we can learn something from comparing with three more simplistic baselines, which differ only in terms of feature extraction. The three baseline models we tested included a unigram model, a unigram+bigram model, and a Part-of-Speech bigram model. For part-of-speech tagging we used the Stanford part-of-speech tagger[4] (Toutanova et al., 2003).

Our three baseline feature spaces have been very commonly used in the language technologies community for a variety of social media analysis tasks, the most common of which in recent years has been sentiment analysis. While these feature spaces are simple, they have remained surprisingly strong baseline approaches when testing is done within domain, and with large enough training sets.

[3] http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm

[4] http://nlp.stanford.edu/software/tagger.shtml

However, these relatively weak, low level features are notorious for low performance when datasets are too small and for low generalizability when evaluated in a cross-domain setting. Because of this, we expect to see our baseline approaches perform well when both training and testing data match in terms of topic distribution and when we use our largest amount of training data. However, we expect performance to degrade as training data set size decreases as well as when we test in a cross-domain setting. We expect to see degradation also with our proposed stretchy patterns. However, we will consider our claims to have been supported if we see less degradation with our stretchy patterns than with the baseline approaches.

We did minimal preprocessing on the textual data prior to feature extraction for all approaches. Specifically, all numbers in the text were replaced with a <number> symbol. Punctuation was separated from words and treated as a separate symbol. All tokens were downcased so that we generalize across capitalization options. In all cases, we use a support vector machine approach to training the model, using the SMO implementation found in Weka (Witten & Frank, 2005), using a linear polynomial kernel and default settings. For each model, we first use a Chi-Squared filter for attribute selection over the training data, retaining only the top 3,000 features prior to training.

4.3 Study 1: Learning on Small Datasets

The purpose of Study 1 was to test the claim that our stretchy patterns achieve higher performance when we train using a small amount of data. For this evaluation, we constructed a test set of 3,000 instances that we use consistently across training configurations. Specifically, we selected 300 blogs from each of the 10 occupations listed above such that 150 of them were from male authors and 150 from female authors. We constructed also a set of training sets of size 300, 800, 1500, 2000, and 3000 randomly selected blogs respectively, in which we maintain the same occupation and gender distribution as in the test set. To compensate for sampling eccentricities, two samples of each training size were extracted, and their results averaged for each experiment. In all cases, from each blog, we randomly selected one blog entry that was at least 100 words long. For each baseline approach as well as the stretchy feature approach, we build a model using each training set, which we then test using the common test set. Thus, for each approach, we can examine how performance increases as amount of training data increases, and we can compare this growth curve between approaches.

Training Set Size	Unigram	Unigram + Bigram	POS Bigram	Stretchy Patterns
300	49.9 (-.002)	49.85(-.002)	51.6 (.032)	48.65(-.027)
800	51.65(.029)	50.15 (.003)	50.55 (.014)	53.15 (.072)
1500	48.6 (-.028)	49.98 (0)	48.63 (-.028)	**53.95 (.066)**
2000	50.55(.011)	51.7 (.034)	51.82 (.063)	**53.98 (.079)**
3000	49.48(-.010)	50.8 (.016)	49.88 (.0025)	**59.05 (.181)**

Table 2 Classification accuracy for varying data sizes (with kappa in parentheses)

The dramatic mediocrity of the baselines' performance highlights the difficulty of the selected data set, confirming the sense that most of what these n-gram models pick up is not truly gender-specific usage, but shadows of the distribution of topics (here, occupations) between the genders. At all sizes except the smallest (where no approach is significantly better than random), our approach outperforms the baselines. At size 800, this difference is marginal ($p < .1$), and at the larger sizes, it is a significant increase ($p < .05$).

4.4 Study 2: Evaluation of Domain Generality

For our evaluation of domain generality, we randomly selected 200 blogs from each of the 10 most common occupations in the corpus, 100 of which were by male authors and 100 by female authors. As in the evaluation above, from each blog, we randomly selected one blog entry that was at least 100 words long. In order to test domain generality, we perform a leave-one-occupation-out cross validation experiment, which we refer to as a Cross Domain evaluation setting. In this setting, on each fold, we always test on blogs from an occupation that was not represented within the training data. Thus, indicators of gender that are specific to an occupation will not generalize from training to test.

Table 3 displays the results from the comparison of our stretchy feature approach with each of the baseline approaches. On average, stretchy patterns generalized better to new domains than the other approaches. The stretchy feature approach beat the baseline approaches in a statistically significant way (p < .05).

Occupation	Unigram	Unigram + Bigram	POS Bigram	Stretchy Patterns
Engineering	49.5 (-.01)	53 (.06)	49 (-.02)	50.5 (.01)
Education	49 (-.02)	52 (.04)	54.5 (.09)	51 (.02)
Internet	55.5 (.11)	47.5 (-.05)	55.5 (.11)	56.5 (.13)
Law	51.5 (.03)	46.5 (-.07)	46.5 (-.07)	50.5 (.01)
Non-Profit	50 (0)	54 (.08)	49 (-.02)	51. (.02)
Technology	50 (0)	53.5 (.07)	50 (0)	51.5 (.03)
Arts	48 (-.04)	46.5 (-.07)	51 (.02)	55.4 (.11)
Media	53 (.06)	50 (0)	45 (-.1)	51.5 (.02)
Science	52 (.04)	48 (-.04)	40.5 (-.19)	59.5 (.19)
Student	51 (.02)	46 (-.09)	55 (.10)	62 (.24)
Average	50.95 (.002)	49.7 (-.007)	49.6 (.01)	**53.94 (.08)**
Random CV	61.05 (.22)	59.65 (.19)	57.95 (.16)	**62.8 (.26)**

Table 3 Accuracy from leave-one-occupation-out cross-validation (with kappa in parentheses)

For random cross-validation, our approach performed marginally better than the unigram baseline, and again significantly exceeds the performance of the other two baselines. Note that for all approaches, there is a significant drop in performance from Random CV to the cross-domain setting, showing that all approaches, including ours, suffer from domain specificity to some extent. However, while all of the baselines drop down to essentially random performance in the cross-domain setting, and stretchy patterns remain significantly higher than random, we show that our approach has more domain generality, although it still leaves room for improvement on that score.

5 Qualitative Analysis of Results

Here we present a qualitative analysis of the sorts of patterns extracted by our method. Although we cannot draw broad conclusions from a qualitative investigation of such a small amount of data, we did observe some interesting trends.

As our features do not so much capture syntactic structure as the loose proximity and order of classes of words, we'll say less about structure and more about what sort of words show up in each others' neighborhood. In particular, a huge proportion of the top-ranked patterns feature instances of the [*end*] and [*first-pron*] categories, suggesting that much of the gender distinction captured by our patterns is to be found around sentence boundaries and self-references. It's believable and encouraging that "the way I talk about myself" is an important element in distinguishing style between genders.

The Chi-squared ranking of the stretchy patterns gives us a hint as to the predictive effect of each as a feature. In the discussion and examples that follow, we'll draw from the highest-ranked features, and refer to the weights' signs to label each pattern as "male" or "female".

In these features the discourse analyst or dialectician can find fodder for their favorite framework, or support for popularly held views on gender and language. For example, we find that about twice as many of the patterns containing either [*third-pron*] or [*second-pron*] in the neighborhood of [*first-pron*] are weighted toward female, supporting earlier findings that women are more concerned with considering interpersonal relationships in their discourse than are men, as in Kite (2002). For example,

[*first-pron*] (GAP+) [*third-pron*]
"*i* (have time for) *them*"

Supporting the notion that distinctively female language is "deviant," and viewed as a divergence from a male baseline, as discussed in Eckert & McConnell-Ginet (1992), we note that more of the top-ranked patterns are weighted toward female. This might suggest that the "female" style is less standard and therefore harder to detect. Additionally, we only find adjacent [*end*] markers, capturing repeated punctuation, in our female-weighted patterns. For instance,

[*adj*] (GAP+) [*end*] (GAP+) [*end*] [*end*]
"*new* (songs) *!* (:-) see yas) . ."

This divergence from the standard sentence form, while more common overall in informal electronic communications, does occur more frequently among female authors in the data. Further analysis of the data suggests that emoticons like *:-)* would have formed a useful category for our patterns, as they occur roughly twice as often in female posts, and often in the context of end-of-sentence punctuation.

We provide a rough numerical overview of the features extracted during the random cross-validation experiment. Samples of high-ranking stretchy patterns appear in Tables 4 and 5. Note that sequences may match more than one pattern, and that GAP+ expansions can have zero length.

[***first-pron***]	
(female)	and **i** have time for...
(female)	a freshman , **my** brother is...
(male)	and **i** overcame my fear ...
[*end*] (GAP+) [*first-pron*]	
(female)	no ! ! **! (i just guess) i**...
(female)	all year **. (. .) i** am so...
(male)	the internet **. () i** ask only...
[*end*] (GAP+) [*end*] and	
(female)	positives **. (gotta stay positive) . and** hey...
(female)	at the park ..**(. sitting at teh bench alone ..). and** walking down on my memory line...
(male)	sunflower **. (she has a few photo galleries ..). and** i would like...
like **(GAP+) [*first-pron*]**	
(female)	well **like (anywho . . . I got) my** picture back…
(female)	it’s times **like (these that I miss) my** friends...
(male)	with something **like (that in the air ,) i** don't...

Table 4. Female Patterns.

[*adj*] (GAP+) [*end*] (GAP+) [*first-pron*]	
(male)	her own **digital (camera) . (what enlightens) me** is...
(male)	a **few (photo galleries ..) . (and) i** would...
(female)	money **all (year) . (..) i** am so much...
[*first-pron*] (GAP+) [*end*]	
(male)	again . **i** (ate so well today , too) **.** lots of ...
(male)	movie **i ('d already seen once before) .**
(female)	a junior and **i (have the top locker) .** lol
[*end*] (GAP+) [*first-pron*] (GAP+) [*cc*]	
(male)	food **! () i ('m so hooked) on this delicious...**
(male)	galleries **.(.. and) i (would like) for** you to...
(female)	alot better **.** () **i (have a locker right) above**...
so **(GAP+) [*end*]**	
(male)	was it ? **so (cousins , stay posted) .** remember...
(male)	experience you've gained **so (far) .** if...
(female)	, its been **so (damn crappy out) .** ok bye

Table 5. Male Patterns.

Although our patterns capture much more than the unigram frequencies of categories, a glance at such among the extracted patterns will prove enlightening. Of the 3000 patterns considered, 1407 were weighted to some degree toward male, and 1593 toward female. Overall, female patterns include more of our chosen categories than their male counterparts. Many of these imbalances matched our initial predictions, in particular the greater number of female patterns with [*first-pron*] (772 vs. 497), [*second-pron*] (47 vs 27), [*third-pron*] (286 vs. 203), and [*end*] (851 vs. 618), [emotion] (36 vs. 20).

Contrary to our expectations, [*md*] appeared only slightly more frequently in female patterns (73 vs. 66), and [*time*] appeared in only a few male patterns (22 female vs. 7 male) - of these time-patterns, most of the matching segments included the word “time” itself, instead of any other time-related words. No patterns containing the divided *curse* categories were among the top-ranked features.

6 Conclusions and Current Directions

In this paper we described a novel template based feature creation approach that we refer to as stretchy patterns. We have evaluated our approach two ways, once to show that with this approach we are able to achieve higher performance than baseline approaches when small amounts of training data are used, and one in which we demonstrated that we are able to achieve better performance in a cross domain evaluation setting.

While the results of our experiments have shown promising results, we acknowledge that we have scratched the surface of the problem we are investigating. First, our comparison was limited to just a couple of strategically selected baselines. However, there have been many variations in the literature on gender classification specifically, and genre analysis more generally, that we could have included in our evaluations, and that would likely offer additional insights. For example, we have tested our approach against POS bigrams, but we have not utilized longer POS sequences, which have been used in the literature on gender classification with mixed results. In practice, longer POS sequences have only been more valuable than POS bigrams when sophisticated feature selection techniques have been used (Mukherje & Liu, 2010). Attention may also be directed to the selection or generation of word categories better suited to stretchy patterns. Alternative approaches to selecting or clustering these features should also be explored.

7 Acknowledgements

This research was funded by ONR grant N000141110221 and NSF DRL-0835426.

References

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. (2003). Gender, genre, and writing style in formal written texts, *Text*, 23(3), pp 321-346.

Argamon, S., Koppel, M., Pennebaker, J., & Schler, J. (2007). Mining the blogosphere: age, gender, and the varieties of self-expression. *First Monday* 12(9).

Arnold, A. (2009). Exploiting Domain And Task Regularities For Robust Named Entity Recognition. PhD thesis, Carnegie Mellon University, 2009.

Arora, S., Joshi, M., Rosé, C. P. (2009). Identifying Types of Claims in Online Customer Reviews, *Proceedings of the North American Chapter of the Association for Computational Linguistics.*

Arora, S., Mayfield, E., Rosé, C. P., & Nyberg, E. (2010). Sentiment Classification using Automatically Extracted Subgraph Features, *Proceedings of the NAACL HLT Workshop on Emotion in Text.*

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics* 12(1), pp 58-88.

Cieri, C., Miller, D., & Walker, K. (2004). The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* pp 69-71.

Corney, M., de Vel, O., Anderson, A., Mohay, G. (2002). Gender-preferential text mining of e-mail discourse, in the Proceedings of the 18th Annual Computer Security Applications Conference.

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 256-263.

Eckert, P. & Rickford, J. (2001). *Style and Sociolinguistic Variation*, Cambridge: University of Cambridge Press.

Eckert, P. & McConnell-Ginet, S. (1992). Think Practically and Look Locally: Language and Gender as Community- Based Practice. In the *Annual Review of Anthropology,* Vol. 21, pages 461-490.

Finkel, J. & Manning, C. (2009). Hierarchical Bayesian Domain Adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Gill, A., Nowson, S. & Oberlander, J. (2009). What Are They Blogging About? Personality, Topic and Motivation in Blogs. In *Proceedings of the Third International ICWSM Conference.*

Gimpel, K., Smith, N. A. (2011). Unsupervised Feature Induction for Modeling Long-Distance Dependencies in Machine Translation, *Forthcoming.*

Girju, R. (2010). Towards Social Causality: An Analysis of Interpersonal Relationships in Online Blogs and Forums, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.*

Goswami, S., Sarkar, S. & Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In *Proceedings of the Third International ICWSM Conference.*

Holmes, J. & Meyerhoff, M. (2003). *The Handbook of Language and Gender*, Blackwell Publishing.

Jiang, M. & Argamon, S. (2008). Political leaning categorization by exploring subjectivities in political blogs. In *Proceedings of the 4th International Conference on Data Mining*, pages 647-653.

Joshi, M. & Rosé, C. P. (2009). Generalizing Dependency Features for Opinion Mining, *Proceedings of the Association for Computational Linguistics.*

Kite, M. (2002) Gender Stereotypes, in the *Encyclopedia of Women and Gender: Sex Similarities and DIfferences*, Volume 1, Academic Press.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

Labov, W. (2010a). *Principles of Linguistic Change: Internal Factors (Volume 1)*, Wiley-Blackwell.

Labov, W. (2010b). *Principles of Linguistic Change: Social Factors (Volume 2)*, Wiley-Blackwell.

Mayfield, E. & Rosé, C. P. (2010). Using Feature Construction to Avoid Large Feature Spaces in Text Classification, in *Proceedings of the Genetic and Evolutionary Computation Conference.*

McEnery, T. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present,* Routledge.

Mukherjee, A. & Liu, B. (2010). Improved Gender Classification of Blog Authors, Proceedings of EMNLP 2010.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J. (2005). Effects of Age and Gender on Blogging, Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Schler, J. (2006). Effects of Age and Gender on Blogging. *Artificial Intelligence*, *86*, 82-84.

Tannen, D. (2001). *You Just Don't Understand: Women and Men in Conversation*, First Quill.

Tsur, O., Davidov, D., & Rappoport, A. (2010). ICWSM – A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Wiebe, J., Bruce, R., Martin, M., Wilson, T., & Ball, M. (2004). Learning Subjective Language, *Computational Linguistics*, 30(3).

Witten, I. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, second edition, Elsevier, San Francisco.

Yan, X., & Yan, L. (2006). Gender classification of weblog authors. *AAAI Spring Symposium Series Computational Approaches to Analyzing Weblogs* (p. 228–230).

Zhang, Y., Dang, Y., Chen, H. (2009). Gender Difference Analysis of Political Web Forums : An Experiment on International Islamic Women's Forum, Proceedings of the 2009 IEEE international conference on Intelligence and security informatics, pp 61-64.