

Modeling Latent Biographic Attributes in Conversational Genres

Nikesh Garera and David Yarowsky

Department of Computer Science, Johns Hopkins University

Human Language Technology Center of Excellence

Baltimore MD, USA

{ngarera,yarowsky}@cs.jhu.edu

Abstract

This paper presents and evaluates several original techniques for the latent classification of biographic attributes such as gender, age and native language, in diverse genres (conversation transcripts, email) and languages (Arabic, English). First, we present a novel partner-sensitive model for extracting biographic attributes in conversations, given the differences in lexical usage and discourse style such as observed between same-gender and mixed-gender conversations. Then, we explore a rich variety of novel sociolinguistic and discourse-based features, including mean utterance length, passive/active usage, percentage domination of the conversation, speaking rate and filler word usage. Cumulatively up to 20% error reduction is achieved relative to the standard Boullis and Ostendorf (2005) algorithm for classifying individual conversations on Switchboard, and accuracy for gender detection on the Switchboard corpus (aggregate) and Gulf Arabic corpus exceeds 95%.

1 Introduction

Speaker attributes such as gender, age, dialect, native language and educational level may be (a) stated overtly in metadata, (b) derivable indirectly from metadata such as a speaker's phone number or userid, or (c) derivable from acoustic properties of the speaker, including pitch and f0 contours (Bocklet et al., 2008). In contrast, the goal of this paper is to model and classify such speaker attributes from only the latent information found in textual transcripts. In particular, we are interested in modeling and classifying biographic at-

tributes such as gender and age based on lexical and discourse factors including lexical choice, mean utterance length, patterns of participation in the conversation and filler word usage. Furthermore, a speaker's lexical choice and discourse style may differ substantially depending on the gender/age/etc. of the speaker's interlocutor, and hence improvements may be achieved via dyadic modeling or stacked classifiers.

There has been substantial work in the sociolinguistics literature investigating discourse style differences due to speaker properties such as gender (Coates, 1997; Eckert, McConnell-Ginet, 2003). Analyzing such differences is not only interesting from the sociolinguistic and psycholinguistic point of view of language understanding, but also from an engineering perspective, given the goal of predicting latent author/speaker attributes in various practical applications such as user authentication, call routing, user and population profiling on social networking websites such as facebook, and gender/age conditioned language models for machine translation and speech recognition. While most of the prior work in sociolinguistics has been approached from a non-computational perspective, Koppel et al. (2002) employed the use of a linear model for gender classification with manually assigned weights for a set of linguistically interesting words as features, focusing on a small development corpus. Another computational study for gender classification using approximately 30 weblog entries was done by Herring and Paolillo (2006), making use of a logistic regression model to study the effect of different features.

While small-scale sociolinguistic studies on monologues have shed some light on important features, we focus on modeling attributes from spoken conversations, building upon the work of

Boulis and Ostendorf (2005) and show how gender and other attributes can be accurately predicted based on the following original contributions:

1. **Modeling Partner Effect:** A speaker may adapt his or her conversation style depending on the partner and we show how conditioning on the predicted partner class using a stacked model can provide further performance gains in gender classification.
2. **Sociolinguistic features:** The paper explores a rich set of lexical and non-lexical features motivated by the sociolinguistic literature for gender classification, and show how they can effectively augment the standard ngram-based model of Boulis and Ostendorf (2005).
3. **Application to Arabic Language:** We also report results for Arabic language and show that the ngram model gives reasonably high accuracy for Arabic as well. Furthermore, we also get consistent performance gains due to partner effect and sociolinguistic features, as observed in English.
4. **Application to Email Genre:** We show how the models explored in this paper extend to email genre, showing the wide applicability of general text-based features.
5. **Application to new attributes:** We show how the lexical model of Boulis and Ostendorf (2005) can be extended to Age and Native vs. Non-native prediction, with further improvements gained from our partner-sensitive models and novel sociolinguistic features.

2 Related Work

Much attention has been devoted in the sociolinguistics literature to detection of age, gender, social class, religion, education, etc. from conversational discourse and monologues starting as early as the 1950s, making use of morphological features such as the choice between the *-ing* and the *-in* variants of the present participle ending of the verb (Fisher, 1958), and phonological features such as the pronunciation of the “*r*” sound in words such as *far*, *four*, *cards*, etc. (Labov, 1966). Gender differences has been one of the primary areas of sociolinguistic research, including work such as Coates (1998) and Eckert and McConnell-Ginet (2003). There has also been some work in developing computational models based on linguistically interesting clues suggested

by the sociolinguistic literature for detecting gender on formal written texts (Singh, 2001; Koppel et al., 2002; Herring and Paolillo, 2006) but it has been primarily focused on using a small number of manually selected features, and on a small number of formal written texts. Another relevant line of work has been on the blog domain, using a bag of words feature set to discriminate age and gender (Schler et al., 2006; Burger and Henderson, 2006; Nowson and Oberlander, 2006).

Conversational speech presents a challenging domain due to the interaction of genders, recognition errors and sudden topic shifts. While prosodic features have been shown to be useful in gender/age classification (e.g. Shafran et al., 2003), their work makes use of speech transcripts along the lines of Boulis and Ostendorf (2005) in order to build a general model that can be applied to electronic conversations as well. While Boulis and Ostendorf (2005) observe that the gender of the partner can have a substantial effect on their classifier accuracy, given that same-gender conversations are easier to classify than mixed-gender classifications, they don’t utilize this observation in their work. In Section 5.3, we show how the predicted gender/age etc. of the partner/interlocutor can be used to improve overall performance via both dyadic modeling and classifier stacking. Boulis and Ostendorf (2005) have also constrained themselves to lexical n-gram features, while we show improvements via the incorporation of non-lexical features such as the percentage domination of the conversation, degree of passive usage, usage of subordinate clauses, speaker rate, usage profiles for filler words (e.g. “umm”), mean-utterance length, and other such properties.

We also report performance gains of our models for a new genre (email) and a new language (Arabic), indicating the robustness of the models explored in this paper. Finally, we also explore and evaluate original model performance on additional latent speaker attributes including age and native vs. non-native English speaking status.

3 Corpus Details

Consistent with Boulis and Ostendorf (2005), we utilized the Fisher telephone conversation corpus (Cieri et al., 2004) and we also evaluated performance on the standard Switchboard conversational corpus (Godfrey et al., 1992), both collected and annotated by the Linguistic Data Consortium. In both cases, we utilized the provided metadata

(including true speaker gender, age, native language, etc.) as only class labels for both training and evaluation, but never as features in the classification. The primary task we employed was identical to Boulis and Ostendorf (2005), namely the classification of gender, etc. of each speaker in an isolated conversation, but we also evaluate performance when classifying speaker attributes given the combination of multiple conversations in which the speaker has participated. The Fisher corpus contains a total of 11971 speakers and each speaker participated in 1-3 conversations, resulting in a total of 23398 *conversation sides* (i.e. the transcript of a single speaker in a single conversation). We followed the preprocessing steps and experimental setup of Boulis and Ostendorf (2005) as closely as possible given the details presented in their paper, although some details such as the exact training/test partition were not currently obtainable from either the paper or personal communication. This resulted in a training set of 9000 speakers with 17587 conversation sides and a test set of 1000 speakers with 2008 conversation sides. The Switchboard corpus was much smaller and consisted of 543 speakers, with 443 speakers used for training and 100 speakers used for testing, resulting in a total of 4062 conversation sides for training and 808 conversation sides for testing.

4 Modeling Gender via Ngram features (Boulis and Ostendorf, 2005)

As our reference algorithm, we used the current state-of-the-art system developed by Boulis and Ostendorf (2005) using unigram and bigram features in a SVM framework. We reimplemented this model as our reference for gender classification, further details of which are given below:

4.1 Training Vectors

For each conversation side, a training example was created using unigram and bigram features with tf-idf weighting, as done in standard text classification approaches. However, stopwords were retained in the feature set as various sociolinguistic studies have shown that use of some of the stopwords, for instance, pronouns and determiners, are correlated with age and gender. Also, only the ngrams with frequency greater than 5 were retained in the feature set following Boulis and Ostendorf (2005). This resulted in a total of 227,450 features for the Fisher corpus and 57,914 features for the Switchboard corpus.

Female		Male	
Fisher Corpus			
husband	-0.0291	my wife	0.0366
my husband	-0.0281	wife	0.0328
oh	-0.0210	uh	0.0284
laughter	-0.0186	ah	0.0248
have	-0.0169	er	0.0222
mhm	-0.0169	i i	0.0201
so	-0.0163	hey	0.0199
because	-0.0160	you doing	0.0169
and	-0.0155	all right	0.0169
i know	-0.0152	man	0.0160
hi	-0.0147	pretty	0.0156
um	-0.0141	i see	0.0141
boyfriend	-0.0134	yeah i	0.0125
oh my	-0.0124	my girlfriend	0.0114
i have	-0.0119	thats thats	0.0109
but	-0.0118	mike	0.0109
children	-0.0115	guy	0.0109
goodness	-0.0114	is that	0.0108
yes	-0.0106	basically	0.0106
uh huh	-0.0105	shit	0.0102
Switchboard Corpus			
oh	-0.0122	wife	0.0078
laughter	-0.0088	my wife	0.0077
my husband	-0.0077	uh	0.0072
husband	-0.0072	i i	0.0053
have	-0.0069	actually	0.0051
uhhuh	-0.0068	sort of	0.0041
and i	-0.0050	yeah i	0.0041
feel	-0.0048	got	0.0039
umhum	-0.0048	a	0.0038
i know	-0.0047	sort	0.0037
really	-0.0046	yep	0.0036
women	-0.0043	the	0.0036
um	-0.0042	stuff	0.0035
would	-0.0039	yeah	0.0034
children	-0.0038	pretty	0.0033
too	-0.0036	that that	0.0032
but	-0.0035	guess	0.0031
and	-0.0034	as	0.0029
wonderful	-0.0032	is	0.0028
yeah yeah	-0.0031	i guess	0.0028

Table 1: Top 20 ngram features for gender, ranked by the weights assigned by the linear SVM model

4.2 Model

After extracting the ngrams, a SVM model was trained via the SVM^{light} toolkit (Joachims, 1999) using the linear kernel with the default toolkit settings. Table 1 shows the most discriminative ngrams for gender based on the weights assigned by the linear SVM model. It is interesting that some of the gender-correlated words proposed by sociolinguistics are also found by this empirical approach, including the frequent use of “oh” by females and also obvious indicators of gender such as “my wife” or “my husband”, etc. Also, named entity “Mike” shows up as a discriminative unigram, this maybe due to the self-introduction at the beginning of the conversations and “Mike” being a common male name. For compatibility with Boulis and Ostendorf (2005), no special pre-

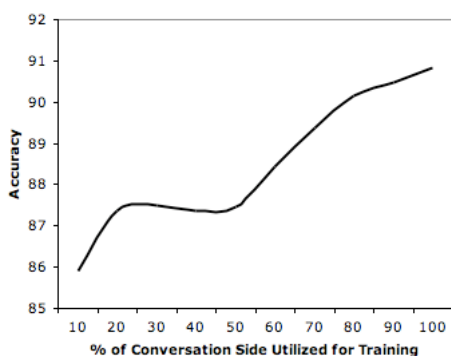


Figure 1: The effect of varying the amount of each conversation side utilized for training, based on the utilized % of each conversation (starting from their beginning).

processing for names is performed, and they are treated as just any other unigrams or bigrams¹.

Furthermore, the ngram-based approach scales well with varying the amount of conversation utilized in training the model as shown in Figure 1.

The “Boulis and Ostendorf, 05” rows in Table 3 show the performance of this reimplemented algorithm on both the Fisher (90.84%) and Switchboard (90.22%) corpora, under the identical training and test conditions used elsewhere in our paper for direct comparison with subsequent results².

5 Effect of Partner’s Gender

Our original contribution in this section is the successful modeling of speaker properties (e.g. gender/age) based on the prior and joint modeling of the partner speaker’s gender/age in the same discourse. The motivation here is that people tend to use stronger gender-specific, age-specific or dialect-specific word/phrase usage and discourse properties when speaking with someone of a similar gender/age/dialect than when speaking with someone of a different gender/age/dialect, when they may adapt a more neutral speaking style. Also, discourse properties such as relative use of the passive and percentage of the conversation dominated may vary depending on the gender or age relationship with the speaking partner. We employ several varieties of classifier stacking and joint modeling to be effectively sensitive to these differences. To illustrate the significance of

¹A natural extension of this work, however, would be to do explicit extraction of self introductions and then do table-lookup-based gender classification, although we did not do so for consistency with the reference algorithm.

²The modest differences with their reported results may be due to unreported details such as the exact training/test splits or SVM parameterizations, so for the purposes of assessing the *relative* gain of our subsequent enhancements we base all reported experiments on the internally-consistent configurations as (re-)implemented here.

Fisher Corpus	
Same gender conversations	94.01
Mixed gender conversations	84.06
Switchboard Corpus	
Same gender conversations	93.22
Mixed gender conversations	86.84

Table 2: Difference in Gender classification accuracy between mixed gender and same gender conversations using the reference algorithm

Classifying speaker’s and partner’s gender simultaneously	
Male-Male	84.80
Female-Female	81.96
Male-Female	15.58
Female-Male	27.46

Table 3: Performance for 4-way classification of the entire conversation into (mm, ff, mf, fm) classes using the reference algorithm on Switchboard corpus.

the “partner effect”, Table 2 shows the difference in the standard algorithm performance between same-gender conversations (when gender-specific style flourishes) and mixed-gender conversations (where more neutral styles are harder to classify). Table 3 shows the classwise performance of classifying the entire conversation into four possible categories. We can see that the mixed-gender cases are also significantly harder to classify on a conversation level granularity.

5.1 Oracle Experiment

To assess the potential gains from full exploitation of partner-sensitive modeling, we first report the result from an oracle experiment, where we assume we know whether the conversation is homogeneous (same gender) or heterogeneous (different gender). In order to effectively utilize this information, we classify both the test conversation side and the partner side, and if the classifier is more confident about the partner side then we choose the gender of the test conversation side based on the heterogeneous/homogeneous information. The overall accuracy improves to 96.46% on the Fisher corpus using this oracle (from 90.84%), leading us to the experiment where the oracle is replaced with a non-oracle SVM model trained on a subset of training data such that all test conversation sides (of the speaker and the partner) are excluded from the training set.

5.2 Replacing Oracle by a Homogeneous vs Heterogenous Classifier

Given the substantial improvement using the Oracle information, we initially trained another bi-

nary classifier for classifying the conversation as mixed or single-gender. It turns out that this task is much harder than the single-side gender classification, task and achieved only a low accuracy value of 68.35% on the Fisher corpus. Intuitively, the homogeneous vs. heterogeneous partition results in a much harder classification task because the two diverse classes of male-male and female-female conversations are grouped into one class (“homogeneous”) resulting in linearly inseparable classes³. This subsequently lead us to create two different classifiers for conversations, namely, male-male vs rest and female-female vs rest⁴ used in a classifier combination framework as follows:

5.3 Modeling partner via conditional model and whole-conversation model

The following classifiers were trained and each of their scores was used as a feature in a meta SVM classifier:

1. Male-Male vs Rest: Classifying the entire conversation (using test speaker and partner’s sides) as male-male or other⁵.
2. Female-Female vs Rest: Classifying the entire conversation (using test speaker and partner’s sides) as female-female or other.
3. Conditional model of gender given most likely partner’s gender: Two separate classifiers were trained for classifying the gender of a given conversation side, one where the partner is male and other where the partner is female. Given a test conversation side, we first choose the most likely gender of the partner’s conversation side using the ngram-based model⁶ and then choose the gender of the test conversation side using the appropriate conditional model.
4. Ngram model as explained in Section 4.

The row labeled “+ Partner Model” in Table 4 shows the performance gain obtained via this meta-classifier incorporating conversation type and partner-conditioned models.

³Even non-linear kernels were not able to find a good classification boundary

⁴We also explored training a 3-way classifier, male-male, female-female, mixed and the results were similar to that of the binarized setup

⁵For classifying the conversations as male-male vs rest or female-female vs rest, all the conversations with either the speaker or the partner present in any of the test conversations were eliminated from the training set, thus creating a disjoint training and test conversation partitions.

⁶All the partner conversation sides of test speakers were removed from the training data and the ngram-based model was retrained on the remaining subset.

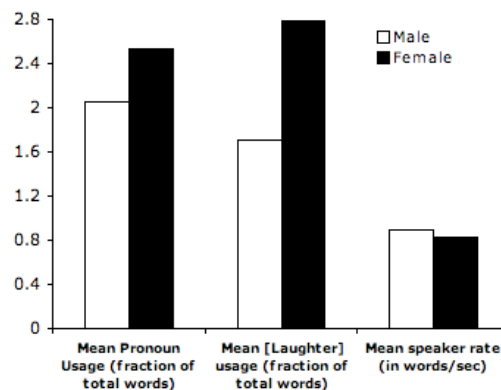


Figure 2: Empirical differences in sociolinguistic features for Gender on the Switchboard corpus

6 Incorporating Sociolinguistic Features

The sociolinguistic literature has shown gender differences for speakers due to features such as speaking rate, pronoun usage and filler word usage. While ngram features are able to reasonably predict speaker gender due to their high detail and coverage and the overall importance of lexical choice in gender differences while speaking, the sociolinguistics literature suggests that other non-lexical features can further help improve performance, and more importantly, advance our understanding of gender differences in discourse. Thus, on top of the standard Boulis and Ostendorf (2005) model, we also investigated the following features motivated by the sociolinguistic literature on gender differences in discourse (Macaulay, 2005):

1. % of conversation spoken: We measured the speaker’s fraction of conversation spoken via three features extracted from the transcripts: % of words, utterances and time.
2. Speaker rate: Some studies have shown that males speak faster than females (Yuan et al., 2006) as can also be observed in Figure 2 showing empirical data obtained from Switchboard corpus. The speaker rate was measured in words/sec., using starting and ending time-stamps for the discourse.
3. % of pronoun usage: Macaulay (2005) argues that females tend to use more third-person male/female pronouns (he, she, him, her and his) as compared to males.
4. % of back-channel responses such as “(laughter)” and “(lipsmacks)”.
5. % of passive usage: Passives were detected by extracting a list of past-participle verbs from Penn Treebank and using occurrences of “form of ”to be” + past participle”.

6. % of short utterances (≤ 3 words).
7. % of modal auxiliaries, subordinate clauses.
8. % of “mm” tokens such as “mhm”, “um”, “uh-huh”, “uh”, “hm”, “hmm”, etc.
9. Type-token ratio
10. Mean inter-utterance time: Avg. time taken between utterances of the same speaker.
11. % of “yeah” occurrences.
12. % of WH-question words.
13. % Mean word and utterance length.

The above classes resulted in a total of 16 sociolinguistic features which were added based on feature ablation studies as features in the meta SVM classifier along with the 4 features as explained previously in Section 5.3.

The rows in Table 4 labeled “+ (any sociolinguistic feature)” show the performance gain using the respective features described in this section. Each row indicates an additive effect in the feature ablation, showing the result of adding the current sociolinguistic feature with the set of features mentioned in the rows above.

7 Gender Classification Results

Table 4 combines the results of the experiments reported in the previous sections, assessed on both the Fisher and Switchboard corpora for gender classification. The evaluation measure was the standard classifier accuracy, that is, the fraction of test conversation sides whose gender was correctly predicted. Baseline performance (always guessing female) yields 57.47% and 51.6% on Fisher and Switchboard respectively. As noted before, the standard reference algorithm is Boulis and Ostendorf (2005), and all cited relative error reductions are based on this established standard, as implemented in this paper. Also, as a second reference, performance is also cited for the popular “Gender Genie”, an online gender-detector⁷, based on the manually weighted word-level sociolinguistic features discussed in Argamon et al. (2003). The additional table rows are described in Sections 4-6, and cumulatively yield substantial improvements over the Boulis and Ostendorf (2005) standard.

7.1 Aggregating results over per-speaker via consensus voting

While Table 4 shows results for classifying the gender of the speaker on a per conversation basis (to be consistent and enable fair comparison

⁷<http://bookblog.net/gender/genie.php>

Model	Acc.	Error Reduc.	
Fisher Corpus (57.5% of sides are female)			
Gender Genie	55.63	-384%	
Ngram (Boulis & Ostendorf, 05)	90.84	<i>Ref.</i>	
+ Partner Model	91.28	4.80%	
+ % of “yeah”	91.33		
+ % of (laughter)	91.38		
+ % of short utt.	91.43		
+ % of auxiliaries	91.48		
+ % of subord-clauses, “mm”	91.58		
+ % of Participation (in utt.)	91.63		
+ % of Passive usage	91.68		9.17%
Switchboard Corpus (51.6% of sides are female)			
Gender Genie	55.94	-350%	
Ngram (Boulis & Ostendorf, 05)	90.22	<i>Ref.</i>	
+ Partner Model	91.58	13.91%	
+ Speaker rate, % of fillers	91.71		
+ Mean utt. len., % of Ques.	91.96		
+ % of Passive usage	92.08		
+ % of (laughter)	92.20		20.25%

Table 4: Results showing improvement in accuracy of gender classifier using partner-model and sociolinguistic features

Model	Acc.	Error Reduc.
Fisher Corpus		
Ngram (Boulis & Ostendorf, 05)	90.50	<i>Ref.</i>
+ Partner Model	91.60	11.58%
+ Socioling. Features	91.70	12.63%
Switchboard Corpus		
Ngram (Boulis & Ostendorf, 05)	92.78	<i>Ref.</i>
+ Partner Model	93.81	14.27%
+ Socioling. Features	96.91	57.20%

Table 5: Aggregate results on a “per-speaker” basis via majority consensus on different conversations for the respective speaker. The results on Switchboard are significantly higher due to more conversations per speaker as compared to the Fisher corpus

with the work reported by Boulis and Ostendorf (2005)), all of the above models can be easily extended to per-speaker evaluation by pooling in the predictions from multiple conversations of the same speaker. Table 5 shows the result of each model on a per-speaker basis using a majority vote of the predictions made on the individual conversations of the respective speaker. The consensus model when applied to Switchboard corpus show larger gains as it has 9.38 conversations per speaker on average as compared to 1.95 conversations per speaker on average in Fisher. The results

on Switchboard corpus show a very large reduction in error rate of more than 57% with respect to the standard algorithm, further indicating the usefulness of the partner-sensitive model and richer sociolinguistic features when more conversational evidence is available.

8 Application to Arabic Language

It would be interesting to see how the Boulis and Ostendorf (2005) model along with the partner-based model and sociolinguistic features would extend to a new language. We used the LDC Gulf Arabic telephone conversation corpus (Linguistic Data Consortium, 2006). The training set consisted of 499 conversations, and the test set consisted of 200 conversations. Each speaker participated in only one conversation, resulting in the same number of training/test speakers as conversations, and thus there was no overlap in speakers/partners between training and test sets. Only non-lexical sociolinguistic features were used for Arabic in addition to the ngram features. The results for Arabic are shown in table 6. Based on prior distribution, always guessing the most likely class for gender (“male”) yielded 52.5% accuracy. We can see that the Boulis and Ostendorf (2005) model gives a reasonably high accuracy in Arabic as well. More importantly, we also see consistent performance gains via partner modeling and sociolinguistic features, indicating the robustness of these models and achieving final accuracy of 96%.

9 Application to Email Genre

A primary motivation for using only the speaker transcripts as compared to also using acoustic properties of the speaker (Bocklet et al., 2008) was to enable the application of the models to other new genres. In order to empirically support this motivation, we also tested the performance of the models explored in this paper on the Enron email corpus (Klimt and Yang, 2004). We manually annotated the sender’s gender on a random collection of emails taken from the corpus. The resulting training and test sets after preprocessing for header information, reply-to’s, forwarded messages consisted of 1579 and 204 emails respectively.

In addition to ngram features, a subset of sociolinguistic features that could be extracted for email were also utilized. Based on the prior distribution, always guessing the most likely class (“male”) resulted in 63.2% accuracy. We can see from Table 7 that the Boulis and Ostendorf (2005)

Model	Acc.	Error Reduc.
Gulf Arabic (52.5% sides are male)		
Ngram (Boulis & Ostendorf, 05)	92.00	<i>Ref.</i>
+ Partner Model	95.00	
+ Mean word len.	95.50	
+ Mean utt. len.	96.00	50.00%

Table 6: Gender classification results for a new language (Gulf Arabic) showing consistent improvement gains via partner-model and sociolinguistic features.

Model	Acc.	Error Reduc.
Enron Email Corpus (63.2% sides are male)		
Ngram (Boulis & Ostendorf, 05)	76.78	<i>Ref.</i>
+ % of subor-claus., Mean word len., Type-token ratio	80.19	
+ % of pronouns.	80.50	16.02%

Table 7: Application of Ngram model and sociolinguistic features for gender classification in a new genre (Email)

model based on lexical features yields a reasonable performance with further improvements due to the addition of sociolinguistic features, resulting in 80.5% accuracy.

10 Application to New Attributes

While gender has been studied heavily in the literature, other speaker attributes such as age and native/non-native status also correlate highly with lexical choice and other non-lexical features. We applied the ngram-based model of Boulis and Ostendorf (2005) and our improvements using our partner-sensitive model and richer sociolinguistic features for a binary classification of the age of the speaker, and classifying into native speaker of English vs non-native.

Corpus details for Age and Native Language:

For age, we used the same training and test speakers from Fisher corpus as explained for gender in section 3 and binarized into greater-than or less-than-or-equal-to 40 for more parallel binary evaluation. For predicting native/non-native status, we used the 1156 non-native speakers in the Fisher corpus and pooled them with a randomly selected equal number of native speakers. The training and test partitions consisted of 2000 and 312 speakers respectively, resulting in 3267 conversation sides for training and 508 conversation sides for testing.

Age \geq 40		Age $<$ 40	
well	0.0330	im thirty	-0.0266
im forty	0.0189	actually	-0.0262
thats right	0.0160	definitely	-0.0226
forty	0.0158	like	-0.0223
yeah well	0.0153	wow	-0.0189
uhhuh	0.0148	as well	-0.0183
yeah right	0.0144	exactly	-0.0170
and um	0.0130	oh wow	-0.0143
im fifty	0.0126	everyone	-0.0137
years	0.0126	i mean	-0.0132
anyway	0.0123	oh really	-0.0128
isnt	0.0118	mom	-0.0112
daughter	0.0117	im twenty	-0.0110
well i	0.0116	cool	-0.0108
in fact	0.0116	think that	-0.0107
whether	0.0111	so	-0.0107
my daughter	0.0111	mean	-0.0106
pardon	0.0110	pretty	-0.0106
gee	0.0109	thirty	-0.0105
know laughter	0.0105	hey	-0.0103
this	0.0102	right now	-0.0100
oh	0.0102	cause	-0.0096
young	0.0100	im actually	-0.0096
in	0.0100	my mom	-0.0096
when they	0.0100	kinda	-0.0095

Table 8: Top 25 ngram features for Age ranked by weights assigned by the linear SVM model

Results for Age and Native/Non-Native:

Based on the prior distribution, always guessing the most likely class for age (age less-than-or-equal-to 40) results in 62.59% accuracy and always guessing the most likely class for native language (non-native) yields 50.59% accuracy.

Table 9 shows the results for age and native/non-native speaker status. We can see that the ngram-based approach for gender also gives reasonable performance on other speaker attributes, and more importantly, both the partner-model and sociolinguistic features help in reducing the error rate on age and native language substantially, indicating their usefulness not just on gender but also on other diverse latent attributes.

Table 8 shows the most discriminative ngrams for binary classification of age, it is interesting to see the use of “well” right on top of the list for older speakers, also found in the sociolinguistic studies for age (Macaulay, 2005). We also see that older speakers talk about their children (“my daughter”) and younger speakers talk about their parents (“my mom”), the use of words such as “wow”, “kinda” and “cool” is also common in younger speakers. To give maximal consistency/benefit to the Boulis and Ostendorf (2005) n-gram-based model, we did not filter the self-reporting n-grams such as “im forty” and “im thirty”, putting our sociolinguistic-literature-based and discourse-style-based features at a relative disadvantage.

Model	Accuracy
Age (62.6% of sides have age \leq 40)	
Ngram Model	82.27
+ Partner Model	82.77
+ % of passive, mean inter-utt. time , % of pronouns + % of “yeah”	83.02
+ type/token ratio, + % of lipsmacks	83.43
+ % of auxiliaries, + % of short utt.	83.83
+ % of “mm”	83.98
(Reduction in Error)	84.03 (9.93%)
Native vs Non-native (50.6% of sides are non-native)	
Ngram	76.97
+ Partner	80.31
+ Mean word length	80.51 (15.37%)

Table 9: Results showing improvement in the accuracy of age and native language classification using partner-model and sociolinguistic features

11 Conclusion

This paper has presented and evaluated several original techniques for the latent classification of speaker gender, age and native language in diverse genres and languages. A novel partner-sensitive model shows performance gains from the joint modeling of speaker attributes along with partner speaker attributes, given the differences in lexical usage and discourse style such as observed between same-gender and mixed-gender conversations. The robustness of the partner-model is substantially supported based on the consistent performance gains achieved in diverse languages and attributes. This paper has also explored a rich variety of novel sociolinguistic and discourse-based features, including mean utterance length, passive/active usage, percentage domination of the conversation, speaking rate and filler word usage. In addition to these novel models, the paper also shows how these models and the previous work extend to new languages and genres. Cumulatively up to 20% error reduction is achieved relative to the standard Boulis and Ostendorf (2005) algorithm for classifying individual conversations on Switchboard, and accuracy for gender detection on the Switchboard corpus (aggregate) and Gulf Arabic exceeds 95%.

Acknowledgements

We would like to thank Omar F. Zaidan for valuable discussions and feedback during the initial stages of this work.

References

- S. Argamon, M. Koppel, J. Fine, and A.R. Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text-Interdisciplinary Journal for the Study of Discourse*, 23(3):321–346.
- T. Bocklet, A. Maier, and E. Nöth. 2008. Age Determination of Children in Preschool and Primary School Age with GMM-Based Supervectors and Support Vector Machines/Regression. In *Proceedings of Text, Speech and Dialogue; 11th International Conference*, volume 1, pages 253–260.
- C. Boulis and M. Ostendorf. 2005. A quantitative analysis of lexical differences between genders in telephone conversations. *Proceedings of ACL*, pages 435–442.
- J.D. Burger and J.C. Henderson. 2006. An exploration of observable features related to blogger age. In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*, pages 15–20.
- C. Cieri, D. Miller, and K. Walker. 2004. The Fisher Corpus: a resource for the next generations of speech-to-text. In *Proceedings of LREC*.
- J. Coates. 1998. *Language and Gender: A Reader*. Blackwell Publishers.
- Linguistic Data Consortium. 2006. *Gulf Arabic Conversational Telephone Speech Transcripts*.
- P. Eckert and S. McConnell-Ginet. 2003. *Language and Gender*. Cambridge University Press.
- J.L. Fischer. 1958. Social influences on the choice of a linguistic variant. *Word*, 14:47–56.
- JJ Godfrey, EC Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. *Proceedings of ICASSP*, 1.
- S.C. Herring and J.C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.
- J. Holmes and M. Meyerhoff. 2003. *The Handbook of Language and Gender*. Blackwell Publishers.
- H. Jing, N. Kambhatla, and S. Roukos. 2007. Extracting social networks and biographical facts from conversational speech transcripts. *Proceedings of ACL*, pages 1040–1047.
- B. Klimt and Y. Yang. 2004. Introducing the Enron corpus. In *First Conference on Email and Anti-Spam (CEAS)*.
- M. Koppel, S. Argamon, and A.R. Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.
- W. Labov. 1966. The Social Stratification of English in New York City. *Center for Applied Linguistics, Washington DC*.
- H. Liu and R. Mihalcea. 2007. Of Men, Women, and Computers: Data-Driven Gender Modeling for Improved User Interfaces. In *International Conference on Weblogs and Social Media*.
- R.K.S. Macaulay. 2005. *Talk that Counts: Age, Gender, and Social Class Differences in Discourse*. Oxford University Press, USA.
- S. Nowson and J. Oberlander. 2006. The identity of bloggers: Openness and gender in personal weblogs. *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*.
- J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. 2006. Effects of age and gender on blogging. *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*.
- I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. *Proceedings of ASRU*, pages 31–36.
- S. Singh. 2001. A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing*, 16(3):251–264.