Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi
Department of Computer Science
Stony Brook University
NY 11794, USA
{rsarawgi, kgajulapalli, ychoi}@cs.stonybrook.edu

Abstract

Sociolinguistic theories (e.g., Lakoff (1973)) postulate that women's language styles differ from that of men. In this paper, we explore statistical techniques that can learn to identify the gender of authors in modern English text, such as web blogs and scientific papers. Although recent work has shown the efficacy of statistical approaches to gender attribution, we conjecture that the reported performance might be overly optimistic due to non-stylistic factors such as topic bias in gender that can make the gender detection task easier. Our work is the first that consciously avoids gender bias in topics, thereby providing stronger evidence to gender-specific styles in language beyond topic. In addition, our comparative study provides new insights into robustness of various stylometric techniques across topic and genre.

1 Introduction

Sociolinguistic theories (e.g., Lakoff (1973)) postulate that women's language styles differ from that of men with respect to various aspects of communication, such as discourse behavior, body language, lexical choices, and linguistic cues (e.g., Crosby and Nyquist (1977), Tannen (1991), Argamon et al. (2003), Eckert and McConnell-Ginet (2003), Argamon et al. (2007)). In this paper, we explore statistical techniques that can learn to identify the gender of authors in modern English text, such as web blogs and scientific papers, motivated by sociolinguistic theories for gender attribution.

There is a broad range of potential applications across computational linguistics and social science where statistical techniques for gender attribution can be useful: e.g., they can help understanding demographic characteristics of user-created web text today, which can provide new insight to social science as well as intelligent marketing and opinion mining. Models for gender attribution can also help tracking changes to gender-specific styles in language over different domain and time. Gender detectors can be useful to guide the style of writing as well, if one needs to assume the style of a specific gender for imaginative writing.

Although some recent work has shown the efficacy of machine learning techniques to gender attribution (e.g., Koppel et al. (2002), Mukherjee and Liu (2010)), we conjecture that the reported performance might be overly optimistic under scrutiny due to non-stylistic factors such as topic bias in gender that can make the gender detection task easier. Indeed, recent research on web blogs reports that there is substantial gender bias in topics (e.g., Janssen and Murachver (2004), Argamon et al. (2007)) as well as in genre (e.g., Herring and Paolillo (2006)).

In order to address this concern, we perform the first comparative study of machine learning techniques for gender attribution after deliberately removing gender bias in topics and genre. Furthermore, making the task even more realistic (and challenging), we experiment with *cross-topic* and *cross-genre* gender attribution, and provide statistical evidence to gender-specific styles in language beyond topic and genre. Five specific questions we aim to investigate are:

Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 78–86,
Portland, Oregon, USA, 23–24 June 2011.

Q1 Are there truly gender-specific characteristics in language? or are they confused with gender preferences in topics and genre?

Q2 Are there deep-syntactic patterns in women's language beyond words and shallow patterns?

Q3 Which stylometric analysis techniques are effective in detecting characteristics in women's language?

Q4 Which stylometric analysis techniques are robust against domain change with respect to topics and genre?

Q5 Are there gender-specific language characteristics even in modern scientific text?

From our comparative study of various techniques for gender attribution, including two publicly available systems - Gender Genie[1] and Gender Guesser[2] we find that (1) despite strong evidence for deep syntactic structure that characterizes gender-specific language styles, such deep patterns are not as robust as shallow morphology-level patterns when faced with topic and genre change, and that (2) there are indeed gender-specific linguistic signals that go beyond topics and genre, even in modern and scientific literature.

2 Related Work

The work of Lakoff (1973) initiated the research on women's language, where ten basic characteristics of women's language were listed. Some exemplary ones are as follows:

1 Hedges: e.g., "kind of", "it seems to be"

2 Empty adjectives: e.g., "lovely", "adorable", "gorgeous"

3 Hyper-polite: e.g., "would you mind ...", "I'd much appreciate if ..."

4 Apologetic: e.g., "I am very sorry, but I think that ..."

5 Tag questions: e.g., "you don't mind, do you?"

Many sociolinguists and psychologists consequently investigated on the validity of each of the above assumptions and extended sociolinguistic theories on women's language based on various controlled experiments and psychological analysis (e.g., Crosby and Nyquist (1977), McHugh and Hambaugh (2010)).

While most theories in socioliguistics and psychology focus on a small set of cognitively identifiable patterns in women's language (e.g., the use of tag questions), some recent studies in computer science focus on investigating the use of machine learning techniques that can learn to identify women's language from a bag of features (e.g., Koppel et al. (2002), Mukherjee and Liu (2010)).

Our work differs from most previous work in that we consciously avoid gender bias in topics and genre in order to provide more accurate analysis of statistically identifiable patterns in women's language. Furthermore, we compare various techniques in stylometric analysis within and beyond topics and genre.

3 Dataset without Unwanted Gender Bias

In this section, we describe how we prepared our dataset to avoid unwanted gender bias in topics and genre. Much of previous work has focused on formal writings, such as English literature, newswire articles and the British Natural Corpus(BNC) (e.g., Argamon et al. (2003)), while recent studies expanded toward more informal writing such as web blogs (e.g., Mukherjee and Liu (2010)). In this work, we chose two very different and prominent genre electronically available today: *web blogs* and *scientific papers*.

Blogs: We downloaded blogs from popular blog sites for 7 distinctive topics:[3] *education*, *travel*, *spirituality*, *entertainment*, *book reviews*, *history* and *politics*. Within each topic, we find 20 articles written by male authors, and additional 20 articles written by female authors. We took the effort to match articles written by different gender even at the subtopic level. For example, if we take a blog written about the TV show "How I met your mother" by a female author, then we also find a blog written by a

[1] http://bookblog.net/gender/genie.php

[2] Available at http://www.hackerfactor.com/GenderGuesser.php

[3] wordpress.com, blogspot.com & nytimes.com/interactive/blogs/directory.html

male author on the same show. Note that previous research on web blogs does not purposefully maintain balanced topics between gender, thereby benefiting from topic bias inadvertently. From each blog, we keep the first 450 (+/- 20) words preserving the sentence boundaries.[4] We plan to make this data publically available.

Scientific Papers: Scientific papers have not been studied in previous research on gender attribution. Scientific papers correspond to very formal writing where gender-specific language styles are not likely to be conspicuous (e.g., Janssen and Murachver (2004)).

For this dataset, we collected papers from the researchers in our own Natural Language Processing community. We randomly selected 5 female and 5 male authors, and collected 20 papers from each author. We tried to select these authors across a variety of subtopics within NLP research, so as to reduce potential topic-bias in gender even in research. It is also worthwhile to mention that authors in our selection are highly established ones who have published over multiple subtopics in NLP.

Similarly as the blog dataset, we keep the first 450 (+/- 20) words preserving the sentence boundaries. Some papers are co-authored by researchers of mixed gender. In those cases, we rely on the gender of the advisory person as she or he is likely to influence on the abstract and intro the most.

4 Statistical Techniques

In this section, we describe three different types of statistical language models that learn patterns at different depth. The first kind is based on probabilistic context-free grammars (PCFG) that learn *deep long-distance syntactic patterns* (Section 4.1). The second kind is based on token-level language models that learn *shallow lexico-syntactic patterns* (Section 4.2). The last kind is based on character-level language models that learn *morphological patterns* on extremely short text spans (Section 4.3). Finally, we describe the bag-of-word approach using the maximum entropy classifier (Section 4.4).

[4]Note that existing gender detection tools require a minimum 300 words for appropriate identification.

4.1 Deep Syntactic Patterns using Probabilistic Context free Grammar

A probabilistic context-free grammar (PCFG) captures syntactic regularities beyond shallow ngram-based lexico-syntactic patterns. Raghavan et al. (2010) recently introduced the use of PCFG for authorship attribution for the first time, and demonstrated that it is highly effective for learning stylistic patterns for authorship attribution. We therefore explore the use of PCFG for gender attribution. We give a very concise description here, referring to Raghavan et al. (2010) for more details.

(1) Train a generic PCFG parser G_o on manually tree-banked corpus such as WSJ or Brown.

(2) Given training corpus D for gender attribution, tree-bank each training document $d_i \in D$ using the PCFG parser G_o.

(3) For each gender γ, train a new gender-specific PCFG parser G_γ using only those tree-banked documents in D that correspond to gender γ.

(4) For each test document, compare the likelihood of the document determined by each gender-specific PCFG parser G_γ, and the gender corresponding to the higher score.

Note that PCFG models can be considered as a kind of language models, where probabilistic context-free grammars are used to find the patterns in language, rather than n-grams. We use the implementation of Klein and Manning (2003) for PCFG models.

4.2 Shallow Lexico-Syntactic Patterns using Token-level Language Models

Token-based (i.e. word-based) language models have been employed in a wide variety of NLP applications, including those that require stylometric analysis, e.g., authorship attribution (e.g., Uzner and Katz (2005)), and Wikipedia vandalism detection (Wang and McKeown, 2010). We expect that token-based language models will be effective in learning shallow lexico-syntactic patterns of gender specific language styles. We therefore experiment with unigram, bigram, and trigram token-level models, and name them as TLM(n=1), TLM(n=2), TLM(n=3), respectively, where TLM stands for **T**oken-based

	lexicon based		deep syntax	morphology			b.o.w.	shallow lex-syntax		
Data Type	Gender Genie	Gender Guesser	PCFG	CLM n=1	CLM n=2	CLM n=3	ME	TLM n=1	TLM n=2	TLM n=3
Male Only	**72.1**	68.6	53.4	65.8	69.0	63.4	57.6	67.1	67.8	66.2
Female Only	27.1	06.4	**74.8**	57.6	73.6	76.8	73.8	60.1	64.2	64.2
All	50.0	37.5	64.1	61.70	**71.3**	70.3	65.8	63.7	66.1	65.4

Table 1: Overall Accuracy of Topic-Balanced Gender Attribution on Blog Data **(Experiment-I)**

Language **M**odels. We use the LingPipe package[5] for experiments.

4.3 Shallow Morphological Patterns using Character-level Language Models

Next we explore the use of character-level language models to investigate whether there are morphological patterns that characterize gender-specific styles in language. Despite its simplicity, previous research have reported that character-level language models are effective for authorship attribution (e.g., Peng et al. (2003b)) as well as genre classification (e.g., Peng et al. (2003a), Wu et al. (2010)). We experiment with unigram, bigram, and trigram character-level models, and name them as CLM(n=1), CLM(n=2), CLM(n=3), respectively, where CLM stands for **C**haracter-based **L**anguage **M**odels. We again make use of the LingPipe package for experiments.

Note that there has been no previous research that directly compares the performance of character-level language models to that of PCFG based models for author attribution, not to mention for gender attribution.

4.4 Bag of Words using Maximum Entropy (MaxEnt) Classifier

We include Maximum Entropy classifier using simple unigram features (bag-of-words) for comparison purposes, and name it as ME. We use the MALLET package (McCallum, 2002) for experiments.

5 Experimental Results

Note that our two datasets are created to specifically answer the following question: *are there gender-specific characteristics in language beyond gender preferences in topics and genre?* One way to answer this question is to test whether statistical models can detect gender attribution on a dataset that is drastically different from the training data in topic and genre. Of course, it is a known fact that machine learning techniques do not transfer well across different domains (e.g., Blitzer et al. (2006)). However, if they can still perform considerably better than random prediction, then it would prove that there is indeed gender-specific stylometric characteristics beyond topic and genre. In what follows, we present five different experimental settings across two different dataset to compare in-domain and cross-domain performance of various techniques for gender attribution.

5.1 Experiments with Blog Dataset

First we conduct two different experiments using the blog data in the order of increasing difficulty.

[Experiment-I: Balanced Topic] Using the web blog dataset introduced in Section 3, we perform gender attribution (classification) task on balanced topics. For each topic, 80% of the documents are used for training and remaining ones are used for testing, yielding 5-fold cross validation. Both training and test data have balanced class distributions so that random guess would yield 50% of accuracy. The results are given in Table 1. Note that the "overall accuracy" corresponds to the average across the five folds.

The PCFG model achieves prediction accuracy 64.1%, demonstrating statistical evidence to gender-specific characteristics in syntactic structure. The PCFG model outperforms two publicly available systems - Gender Genie and Gender Guesser, which are based on a fixed list of indicator words. The difference is statistically significant ($p = 0.01 < 0.05$)

[5]Available at `http://alias-i.com/lingpipe/`

Topic	lexicon based		deep syntax	morphology			b.o.w.	shallow lex-syntax		
	Gender Genie	Gender Guesser	PCFG	CLM n=1	CLM n=2	CLM n=3	ME	TLM n=1	TLM n=2	TLM n=3
Per Topic Accuracy (%) for All Authors										
Entertain	50.0	42.5	50.0	52.5	**67.5**	**67.5**	60.0	57.5	57.5	57.5
Book	50.0	42.5	65.0	57.5	67.5	**72.5**	55.0	60.0	67.5	67.5
Politics	35.0	30.0	50.0	47.5	**52.5**	50.0	45.0	**52.5**	**52.5**	**52.5**
History	40.0	35.0	77.5	65.0	**80.0**	**80.0**	55.0	65.0	65.0	65.0
Education	62.5	42.5	55.0	63.0	65.0	**70.0**	63.0	55.0	57.5	52.5
Travel	62.5	37.5	63.0	**65.0**	63.0	63.0	63.0	62.5	**65.0**	**65.0**
Spirituality	50.0	32.5	53.0	**78.0**	**78.0**	**78.0**	50.0	65.0	70.0	72.5
Avg	50.0	37.5	59.0	61.2	**68.3**	**68.3**	55.87	60.0	61.3	61.5
Per Topic Accuracy (%) for Female Authors										
Entertain	25.0	10.0	**85.0**	70.0	50.0	**85.0**	70.0	75.0	75.0	75.0
Book	15.0	15.0	**95.0**	80.0	**95.0**	90.0	85.0	75.0	90.0	90.0
Politics	10.0	05.0	**65.0**	00.0	05.0	00.0	35.0	30.0	30.0	25.0
History	10.0	05.0	**90.0**	70.0	80.0	75.0	70.0	50.0	50.0	50.0
Education	45.0	10.0	80.0	95.0	85.0	90.0	**100.0**	50.0	55.0	50.0
Travel	65.0	00.0	85.0	90.0	**100.0**	**100.0**	**100.0**	85.0	95.0	90.0
Spirituality	20.0	00.0	60.0	65.0	65.0	**70.0**	45.0	50.0	50.0	50.0
Avg	27.1	06.4	**80.0**	67.1	68.6	72.9	72.1	59.3	63.6	61.4
Per Topic Accuracy (%) for Male Authors										
Entertain	75.0	75.0	15.0	35.0	**85.0**	50.0	50.0	40.0	40.0	40.0
Book	**80.0**	70.0	35.0	35.0	40.0	55.0	25.0	45.0	45.0	45.0
Politics	60.0	55.0	35.0	95.0	**100.0**	**100.0**	55.0	75.0	75.0	80.0
History	70.0	65.0	65.0	60.0	80.0	**85.0**	40.0	80.0	80.0	80.0
Education	**80.0**	75.0	30.0	30.0	45.0	50.0	25.0	60.0	60.0	55.0
Travel	60.0	**75.0**	40.0	40.0	25.0	25.0	25.0	40.0	35.0	40.0
Spirituality	80.0	65.0	45.0	**90.0**	**90.0**	85.0	55.0	80.0	90.0	95.0
Avg	**72.1**	68.6	37.9	55.0	66.4	64.2	39.3	60.0	60.8	62.1

Table 2: Per-Topic & Per-Gender Accuracy of Cross-Topic Gender Attribution on Blog Data **(Experiment-II)**

using paired student's t-test.[6]

Interestingly, the best performing approaches are character-level language models, performing substantially better (71.30% for n=2) than both the token-level language models (66.1% for n=2) and the PCFG model (64.10%). The difference between CLM(n=2) and PCFG is statistically significant ($p = 0.015 < 0.05$) using paired student's t-test, while the difference between TLM(n=2) and PCFG is not.

[6]We also experimented with the interpolated PCFG model following Raghavan et al. (2010) using various interpolation dataset, but we were not able to achieve a better result in our experiments. We omit the results of interpolated PCFG models for brevity.

As will be seen in the following experiment (**Experiment-II**) using the Blog dataset as well, the performance of PCFG models is very close to that of unigram language models. As a result, one might wonder whether PCFG models are learning any useful syntactic pattern beyond terminal productions that can help discriminating gender-specific styles in language. This question will be partially answered in the fourth experiment (**Experiment-IV**) using the Scientific Paper dataset, where PCFG models demonstrate considerably better performance over the unigram language models.

Following Raghavan et al. (2010), we also exper-

	lexicon based		deep syntax	morphology			b.o.w.	shallow lex-syntax		
Data Type	Gender Genie	Gender Guesser	PCFG	CLM n=1	CLM n=2	CLM n=3	ME	TLM n=1	TLM n=2	TLM n=3
Male Only	85.0	63.0	59.0	96.0	94.0	**99.0**	62.0	68.0	68.0	68.0
Female Only	9.0	0.0	36.0	10.0	8.0	18.0	**61.0**	34.0	32.0	32.0
All	47.0	31.5	47.5	53.0	51.0	58.5	**61.5**	51.0	50.0	50.0

Table 3: Overall Accuracy of Cross-Topic /Cross-Genre Gender Attribution on Scientific Papers **(Experiment-III)**

imented with ensemble methods that linearly combine the output of different classifiers, but we omit the results in Table 1, as we were not able to obtain consistently higher performance than the simple character-level language models in our dataset.

[Experiment-II: Cross-Topic] Next we perform cross-topic experiments using the same blog dataset, in order to quantify the robustness of different techniques against topic change. We train on 6 topics, and test on the remaining 1 topic, making 7-fold cross validation. The results are shown in Table 2, where the top one third shows the performance for all authors, the next one third shows the performance with respect to only female authors, the bottom one third shows the performance with respect to only male authors.

Again, the best performing approaches are based on character-level language models, achieving upto 68.3% in accuracy. PCFG models and token-level language models achieve substantially lower accuracy of 59.0% and 61.5% respectively. Per-gender analysis in Table 1 reveals interesting insights into different approaches. In particular, we find that Gender Genie and Gender Guesser are biased toward male authors, attributing the majority authors as male. PCFG and ME on the other hand are biased toward female authors. Both character-level and token-level language models show balanced distribution between gender. We also experimented with ensemble methods, but omit the results as we were not able to obtain higher scores than simple character-level language models.

From these two experiments so far, we find that PCFG models and word-level language models are neither as effective, nor as robust as character-level language models for gender attribution. Despite overall low performance of PCFG models, this result suggests that PCFG models are able to learn gender-specific syntactic patterns, albeit the signals from deep syntax seem much weaker than those of very shallow morphological patterns.

5.2 Experiments with Scientific Papers

Next we present three different experiments using the scientific data, in the order of decreasing difficulty.

[Experiment-III: Cross-Topic & Cross-Genre] In this experiment, we challenge statistical techniques for gender attribution by changing both topics and genre across training and testing. To do so, we train models on the blog dataset and test on the scientific paper dataset. Notice that this is a dramatically harder task than the previous two experiments.

Note also that previous research thus far has not reported experiments such as this, or even like the previous one. It is worthwhile to mention that our goal in this paper is not domain adaptation for gender attribution, but merely to quantify to what degree the gender-specific language styles can be traced across different topics and genre, and which techniques are robust against domain change.

The results are shown in Table 5. Precisely as expected, the performance of all models drop significantly in this scenario. The two baseline systems – Gender Genie and Gender Guesser, which are not designed for formal scientific writings also perform worse in this dataset. Table 4 discussed in the next experiment will provide more insight into this by providing per-gender accuracy of these baseline systems.

From this experiment, we find a rather surprising message: although the performance of most statistical approaches decreases significantly, notice that most approaches perform still better than random (50%) prediction, achieving upto 61.5% accuracy.

Data Type	lexicon based		deep syntax	morphology			b.o.w.	shallow lex-syntax		
	Gender Genie	Gender Guesser	PCFG	CLM n=1	CLM n=2	CLM n=3	ME	TLM n=1	TLM n=2	TLM n=3
Per Author Accuracy (%) for All Authors										
All	47.0	31.5	**76.0**	73.0	72.0	**76.0**	70.50	63.5	62.5	62.5
Per Author Accuracy (%) for Male Authors										
A	80.0	55.0	75.0	**100.0**	**100.0**	**100.0**	45.0	45.0	40.0	40.0
B	90.0	75.0	75.0	80.0	70.0	**85.0**	55.0	45.0	40.0	40.0
C	95.0	55.0	85.0	85.0	90.0	**95.0**	90.0	90.0	90.0	90.0
D	85.0	65.0	**100.0**	95.0	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
E	75.0	65.0	**90.0**	70.0	85.0	80.0	70.0	70.0	70.0	60.0
Avg	85.0	63.0	85.0	86.0	89.0	**92.0**	72.0	70.0	68.0	66.0
Per Author Accuracy (%) for Female Authors										
F	15.0	0.0	95.0	05.0	30.0	75.0	**100.0**	85.0	85.0	85.0
G	5.0	0.0	25.0	55.0	70.0	**85.0**	75.0	80.0	**85.0**	**85.0**
H	10.0	0.0	65.0	**70.0**	45.0	35.0	40.0	35.0	30.0	30.0
I	15.0	0.0	80.0	**85.0**	45.0	50.0	65.0	35.0	35.0	35.0
J	0.0	0.0	70.0	**85.0**	**85.0**	**85.0**	65.0	50.0	50.0	60.0
Avg	9.0	0.0	67.0	60.0	55.0	66.0	**69.0**	57.0	57.0	59.0

Table 4: Per-Author Accuracy of Cross-Topic Gender Attribution for Scientific Papers **(Experiment-IV)**

Considering that the models are trained on drastically different topics and genre, this result suggests that there are indeed gender-specific linguistic signals beyond different topics and genre. This is particularly interesting given that scientific papers correspond to very formal writing where gender-specific language styles are not likely to be conspicuous (e.g., Janssen and Murachver (2004)).

[Experiment-IV: Cross-Topic] Next we perform cross-topic experiment, only using the scientific paper dataset. Because the stylistic difference in genre is significantly more prominent than the stylistic difference in topics, this should be a substantially easier task than the previous experiment. Nevertheless, previous research to date has not attempted to evaluate gender attribution techniques across different topics. Here we train on 4 authors per gender (8 authors in total), and test on the remaining 2 authors, making 5-fold cross validation. As before, the class distributions are balanced in both training and test data.

The experimental results are shown in Table 4, where we report per-author, per-gender, and overall average accuracy. As expected, the overall performance increase dramatically, as models are trained on articles in the same genre. It is interesting to see how Gender Genie and Gender Guesser are extremely biased toward male authors, achieving almost zero accuracy with respect to articles written by female authors. Here the best performing models are PCFG and CLM(n=3), both achieving 76.0% in accuracy. Token-level language models on the other hand achieve significantly lower performance.

Remind that in the first two experiments based on the blog data, PCFG models and token-level language models performed similarly. Given that, it is very interesting that PCFG models now perform just as good as character-level language models, while outperforming token-level language models significantly. We conjecture following two reasons to explain this:

- First, scientific papers use very formal language, thereby suppressing gender-specific lexical cues that are easier to detect (e.g., empty words such as "lovely", "gorgeous" (Lakoff, 1973)). In such data, deep syntactic patterns play a much stronger role in detecting gender specific language styles. This also indirectly

	lexicon based		deep syntax	morphology			b.o.w.	shallow lex-syntax		
Data Type	Gender Genie	Gender Guesser	PCFG	CLM n=1	CLM n=2	CLM n=3	ME	TLM n=1	TLM n=2	TLM n=3
Male Only	85.0	63.0	86.0	**92.0**	**92.0**	91.0	86.0	86.0	87.0	88.0
Female Only	9.0	0.0	84.0	88.0	87.0	**92.0**	91.0	83.0	84.0	86.0
All	47.0	31.5	85.0	90.0	88.50	**91.50**	88.50	85.0	85.5	87.0

Table 5: Overall Accuracy of Topic-Balanced Gender Attribution on Scientific Papers **(Experiment-V)**

addresses the concern raised in **Experiment-I & II** as to whether the PCFG models are learning any syntactic pattern beyond terminal productions that are similar to unigram language models.

- Second, our dataset is constructed in such a way that the training and test data do not share articles written by the same authors. Furthermore, the authors are chosen so that the main research topics are substantially different from each other. Therefore, token-based language models are likely to learn topical words and phrases, and suffer when the topics change dramatically between training and testing.

[Experiment-V: Balanced Topic] Finally, we present the conventional experimental set up, where topic distribution is balanced between training and test dataset. This is not as interesting as the previous two scenarios, however, we include this experiment in order to provide a loose upper bound. Because we choose each different author from each different sub-topic of research, we need to split articles by the same author into training and testing to ensure balanced topic distribution. We select 80% of articles from each author as training data, and use the remaining 20% as test data, resulting in 5-fold cross validation.

This is the easiest task among the three experiments using the scientific paper data, hence the performance increases substantially. As before, character-level language models perform the best, with CLM n=3 reaching extremely high accuracy of 91.50%. All other statistical approaches perform very well achieving at least 85% or higher accuracy.

Note that token-level language models perform very poorly in the previous experimental setting, while performing close to the top performer in this experiment. We make the following two conclusions based on the last two experiments:

- Token-level language models have the tendency of learning topics words, rather than just stylometric cues.
- When performing cross-topic gender attribution (as in **Experiment-IV**), PCFG models are more robust than token-level language models.

6 Conclusions

We postulate that previous study in gender attribution might have been overly optimistic due to gender specific preference on topics and genre. We perform the first comparative study of machine learning techniques for gender attribution consciously removing gender bias in topics. Rather unexpectedly, we find that the most robust approach is based on character-level language models that learn morphological patterns, rather than token-level language models that learn shallow lexico-syntactic patterns, or PCFG models that learn deep syntactic patterns. Another surprising finding is that we can trace statistical evidence of gender-specific language styles beyond topics and genre, and even in modern scientific papers.

Acknowledgments

We thank reviewers for giving us highly insightful and valuable comments.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere:

Age, gender and the varieties of selfexpression. In *First Monday, Vol. 12, No. 9*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Faye Crosby and Linda Nyquist. 1977. The female register: an empirical study of lakoff's hypotheses. In *Language in Society, 6*, pages 313 – 322.

Penelope Eckert and Sally McConnell-Ginet. 2003. Language and gender. Cambridge University Press.

Susan C. Herring and John C. Paolillo. 2006. Gender and genre variations in weblogs. In *Journal of Sociolinguistics, Vol. 10, No. 4.*, pages 439 –459.

Anna Janssen and Tamar Murachver. 2004. The relationship between gender and topic in gender-preferential language use. In *Written Communication, 21*, pages 344– 367.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics.

Moshe Koppel, Shlomo Argamon, and Anat Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, June.

Robin T. Lakoff. 1973. Language and woman's place. In *Language in Society, Vol. 2, No. 1*, pages 45 – 80.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Maureen C. McHugh and Jennifer Hambaugh. 2010. She said, he said: Gender, language, and power. In *Handbook of Gender Research in Psychology. Volume 1: Gender Research in General and Experimental Psychology*, pages 379 – 410.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 207–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fuchun Peng, Dale Schuurmans, Vlado Keselj, and Shaojun Wang. 2003a. Language independent authorship attribution with character level n-grams. In *EACL*.

Funchun Peng, Dale Schuurmans, and Shaojun Wang. 2003b. Language and task independent text categorization with simple language models. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL*, pages 38–42, Uppsala, Sweden, July. Association for Computational Linguistics.

Deborah Tannen. 1991. You just don't understand: Women and men in conversation. Ballantine Books.

Ozlem Uzner and Boris Katz. 2005. A Comparative Study of Language Models for Book And Author Recognition. In *Second International Joint Conference on Natural Language Processing:Full Papers*, pages 1969–980. Association for Computational Linguistics.

William Yang Wang and Kathleen R. McKeown. 2010. "got you!": Automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In *23rd International Conference on Computational Linguistics (Coling 2010)*, page 1146?1154.

Zhili Wu, Katja Markert, and Serge Sharoff. 2010. Fine-grained genre classification using structural learning algorithms. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 749–759, Uppsala, Sweden, July. Association for Computational Linguistics.