Author Age Prediction from Text using Linear Regression

Dong Nguyen **Noah A. Smith** **Carolyn P. Rosé**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{dongn,nasmith,cprose}@cs.cmu.edu

Abstract

While the study of the connection between discourse patterns and personal identification is decades old, the study of these patterns using language technologies is relatively recent. In that more recent tradition we frame author age prediction from text as a regression problem. We explore the same task using three very different genres of data simultaneously: blogs, telephone conversations, and online forum posts. We employ a technique from domain adaptation that allows us to train a joint model involving all three corpora together as well as separately and analyze differences in predictive features across joint and corpus-specific aspects of the model. Effective features include both stylistic ones (such as POS patterns) as well as content oriented ones. Using a linear regression model based on shallow text features, we obtain correlations up to 0.74 and mean absolute errors between 4.1 and 6.8 years.

1 Introduction

A major thrust of research in sociolinguistics is to understand the connection between the way people use language and their community membership, where community membership can be construed along a variety of dimensions, including age, gender, socioeconomic status and political affiliation. A person is a member of a multiplicity of communities, and thus the person's identity and language are influenced by many factors.

In this paper we focus on the relationship between age and language use. Recently, machine learning methods have been applied to determine the age of persons based on the language that they utter. Studies of the stylistic and content-based features that predict age or other personal characteristics yield new insights into the connection between discourse and identity. However, that connection is known to be highly contextual, such as whether the data were collected synchronously or asynchronously, through typed or spoken interaction, or whether participants can see one another or not. Recent work in the area of domain adaptation raises awareness about the effect of contextual factors on the generality of text prediction models.

Our first contribution to this literature is an investigation of age prediction using a multi-corpus approach. We present results and analysis across three very different corpora: a blog corpus (Schler et al., 2006), a transcribed telephone speech corpus (Cieri et al., 2004) and posts from an online forum on breast cancer. By using the domain adaptation approach of Daumé III (2007), we train a model on all these corpora together and separate the global features from corpus-specific features that are associated with age.

A second contribution is the investigation of age prediction with age modeled as a continuous variable rather than as a categorical variable. Most prior research on age prediction has framed this as a two-class or three-class classification problem (e.g., Schler et al., 2006 and Garera and Yarowsky, 2009). In our work, modeling age as a continuous variable is interesting not only as a more realistic representation of age, but also for practical benefits of joint modeling of age across corpora since the bound-

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 115–123,
Portland, OR, USA, 24 June 2011.

aries for discretizing age into a categorical variable in prior work have been chosen heuristically and in a corpus-dependent way, making it hard to compare performance across different kinds of data.

In the remainder of the paper, we first discuss related work and present and compare the different datasets. We then outline our approach and results. We conclude with discussion and future work.

2 Related work

Time is an important factor in sociolinguistic analysis of language variation. While a thorough review of this work is beyond the scope of this paper, Eckert (1996) gives an overview of the literature on age as a sociolinguistic variable. Linguistic variation can occur as an individual moves through life, or as a result of changes in the community itself as it moves through time. As an added complexity, Argamon et al. (2007) found connections between language variation and age and gender. Features that were used with increasing age were also used more by males for any age. Features that were used with decreasing age were used more by females. In other work, the same features that distinguish male and female writing also distinguish non-fiction and fiction (Argamon et al., 2003). Thus, the separate effects of age, time period, gender, topic, and genre may be difficult to tease apart in naturalistic data where many of these variables are unknown.

Recently, machine learning approaches have been explored to estimate the age of an author or speaker using text uttered or written by the person. This has been modeled as a classification problem, in a similar spirit to sociolinguistic work where age has been investigated in terms of differences in distributions of characteristics between cohorts. In the sociolinguistic literature, cohorts such as these are determined either *etically* (arbitrary, but equal age spans such as decades) or *emically* (related to life stage, such as adolescence etc.). In machine learning research, these cohorts have typically been determined for practical reasons relating to distribution of age groups within a corpus, although the boundaries sometimes have also made sense from a life stage perspective. For example, researchers have modeled age as a two-class classification problem with boundaries at age 40 (Garera and Yarowsky, 2009) or 30 (Rao et al., 2010). Another line of work has looked at modeling age estimation as a three-class classification problem (Schler et al., 2006; Goswami et al., 2009), with age groups of 13-17, 23-27 and 33-42. In addition to machine learning experiments, other researchers have published statistical analyses of differences in distribution related to age and language and have found similar patterns.

As an example of one of these studies, Pennebaker and Stone (2003) analyzed the relationship between language use and aging by collecting data from a large number of previous studies. They used LIWC (Pennebaker et al., 2001) for analysis. They found that with increasing age, people tend to use more positive and fewer negative affect words, more future-tense and less past-tense, and fewer self-references. Furthermore, a general pattern of increasing cognitive complexity was seen. Barbieri (2008) uses key word analysis to analyze language and age. Two groups (15–25 and 35–60) were compared. Analysis showed that younger speakers' talk is characterized by slang and swear words, indicators of speaker stance and emotional involvement, while older people tend to use more modals.

Age classification experiments have been conducted on a wide range of types of data, including blogs (Schler et al., 2006; Goswami et al., 2009), telephone conversations (Garera and Yarowsky, 2009), and recently Twitter (Rao et al., 2010). Effective features were both content features (such as unigrams, bigrams and word classes) as well as stylistic features (such as part-of-speech, slang words and average sentence length). These separate published studies present some commonalities of findings. However, based on these results from experiments conducted on very different datasets, it is not possible to determine how generalizable the models are. Thus, there is a need for an investigation of generalizability specifically in the modeling of linguistic variation related to age, which we present in this paper.

Age classification from speech data has been of interest for many years. Recently, age regression using speech features has been explored (Spiegl et al., 2009). Spiegel's system obtained a mean absolute error of approximately 10 years using support vector regression. Van Heerden et al. (2010) explore combining regression estimates to improve age classification. As far as we are aware, we are the first to publish results from a regression model that directly predicts age using textual features.

3 Data description

We explore three datasets with different characteristics. The data was divided into a training, development and test set. Statistics are listed in Table 1.

3.1 Blog corpus

In August 2004 Schler et al. (2006) crawled blogs from `blogger.com`. Information such as gender and age were provided by the users in their respective profiles. Users were divided into three age groups, and each group had an equal number of female and male bloggers. In our experiments, every document consists of all posts from a particular blogger.

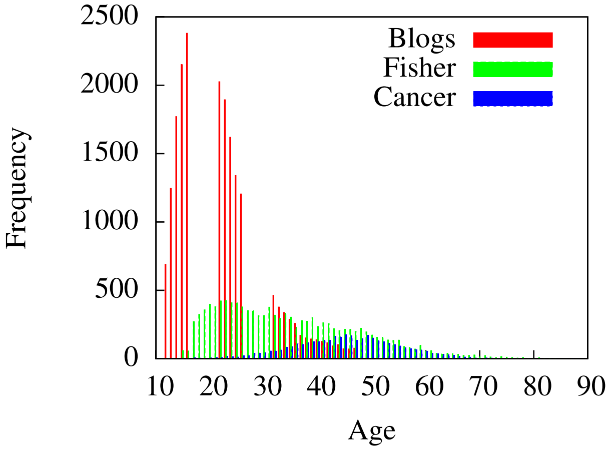
3.2 Fisher telephone corpus

The Fisher corpus (Cieri et al., 2004) contains transcripts of telephone conversations. People were randomly assigned to pairs, and for (almost) every person, characteristics such as gender and age were recorded. Furthermore, for each conversation a topic was assigned. The data was collected beginning December 2002 and continued for nearly one year. In our experiments, we aggregate the data for each person.

3.3 Breast cancer forum

We drew data from one of the most active online forums for persons with breast cancer.[1] All posts and user profiles of the forum were crawled in January 2011. Only a small proportion of users had indicated their age in their profile. We manually annotated the age of approximately 200 additional users with less common ages by looking manually at their posts. An author's age can often be annotated because users tend to make references to their age when they introduce themselves or when telling their treatment history (e.g., *I was diagnosed 2 years ago when I was just 38*). Combining this with the date of the specific post, a birth year can be estimated. Because a person's data can span multiple years, we aggregate all the data per year for each person. Each person was assigned randomly to one of the data splits, to make sure all documents representing the same person appeared in only one split. The dataset contains posts from October 2002 until January 2011.

Figure 1: Comparison of age frequency in datasets.



3.4 Dataset comparison and statistics

The datasets differ in several respects: specificity (general topics versus breast cancer), modality of interaction (telephone conversations versus online forum versus blog post), age distribution, and amount of data per person. The blog and Fisher dataset contain approximately equal amounts of males and females, while the breast cancer dataset is heavily biased towards women.

A comparison of the age distributions of the three corpora is given in Figure 1. The Fisher dataset has the most uniform distribution across the ages, while the blog data has a lot of young persons and the breast cancer forum has a lot of older people. The youngest person in our dataset is 13 years old and the oldest is 88. Note that our blog corpus contains gaps between different age categories, which is an artifact of the experimental approach used by the people who released this dataset (Schler et al., 2006).

Because all datasets were created between 2002 and 2011, we are less likely to observe results due to cohort effects (changes that occur because of collective changes in culture, such as use of the Internet).

[1] `http://community.breastcancer.org`

Table 1: Datasets statistics.

	Blogs		Fisher		Cancer		
Data	**#docs**	**avg #tokens**	**#docs**	**avg #tokens**	**#docs**	**avg #tokens**	**#persons**
Training	9,660	13,042	5,957	3,409	2,330	22,719	1,269
Development	4,830	13,672	2,977	3,385	747	32,239	360
Test	4,830	13,206	2,980	3,376	797	26,952	368

4 Experimental setup

4.1 Linear regression

Given an input vector $\mathbf{x} \in \mathbb{R}^m$, where $x_1, \ldots, x_m$ represent features (also called independent variables or predictors), we find a prediction $\hat{y} \in \mathbb{R}$ for the age of a person $y \in \mathbb{R}$ using a linear regression model: $\hat{y} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$ where β_0 and $\boldsymbol{\beta}$ are the parameters to estimate. Usually, the parameters are learned by minimizing the sum of squared errors. In order to strive for a model with high explanatory value, we use a linear regression model with Lasso (also called L_1) regularization (Tibshirani, 1996). This minimizes the sum of squared errors, but in addition adds a penalty term $\lambda \sum_{j=1}^{m} |\beta_j|$. λ is a constant and can be found by optimizing over the development data. As a result, this method delivers sparse models. We use OWLQN to optimize the regularized empirical risk (Andrew and Gao, 2007; Gao et al., 2007). We evaluate the models by reporting the correlation and mean absolute error (MAE).

4.2 Joint model

To discover which features are important across datasets and which are corpus-specific, we train a model on the data of all corpora using the feature representation proposed by Daumé III (2007). Using this model, the original feature space is augmented by representing each individual feature as 4 new features: a global feature and three corpus-specific features, specifically one for each dataset. Thus for every feature f, we now have f_{global}, f_{blogs}, f_{fisher} and f_{cancer}. For every instance, only the global and the one specific corpus feature are set. For example for a particular feature value x_j for the blog dataset we would have $\langle x_j, x_j, 0, 0 \rangle$. If it would appear in the cancer dataset we would have $\langle x_j, 0, 0, x_j \rangle$. Because the resulting model using L_1 regression only selects a small subset of the features, some features may only appear either as global features or as corpus-specific features in the final model.

4.3 Overview different models

Besides experimenting with the joint model, we are also interested in the performance using only the discovered global features. This can be achieved by applying the weights for the global features directly as learned by the joint model, or retraining the model on the individual datasets using only the global features. In summary, we have the following models:

- INDIV: Models trained on the three corpora individually.
- JOINT: Model trained on all three corpora with features represented as in Daumé III (2007).
- JOINT-Global: Using the learned JOINT model but only keeping the global features.
- JOINT-Global-Retrained: Using the discovered global features by the JOINT model, but *retrained* on each specific dataset.

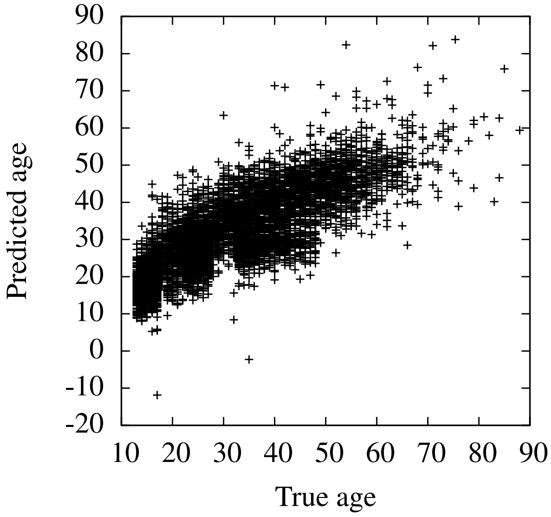
4.4 Features

4.4.1 Textual features

We explore the following textual features; all features are frequency counts normalized by the length (number of tokens) of the document.

- *Unigrams*.
- *POS unigrams* and *bigrams*. Text is tagged using the Stanford POS tagger (Toutanova et al., 2003).
- *LIWC* (Pennebaker et al., 2001). This is a word counting program that captures word classes such as inclusion words (*LIWC-incl:* "with," "and," "include," etc.), causation words (*LIWC-cause:* "because," "hence," etc.), and stylistic characteristics such as percentage of words longer than 6 letters (*LIWC-Sixltr*).

Figure 2: Scatterplot of true and predicted age.



4.4.2 Gender

Because the gender of a person also influences how age is reflected in a person's text or speech (e.g. Argamon et al. (2007)), we add a binary feature for the gender of the person (Male = 1, Female = 0). This feature is only known for the blog and Fisher dataset. For the breast cancer dataset the gender is not known, but we assume they are all women.

5 Results and discussion

As discussed, we experiment with four different models. We explore three different feature sets: only unigrams, only POS, and the full feature set. The results are presented in Table 2. The most important features using the JOINT model with the full feature set (condition 10) are presented in Table 3.

5.1 Quantitative analysis

Overall, similar performance is obtained on the Fisher and blog datasets. The highest correlations were achieved on the Fisher dataset, with a best correlation of $r = 0.742$. This gives an r^2 value of 0.551, indicating that 55% of the variance can be explained by the model. However, a higher mean absolute error (MAE) was observed compared to the blog dataset. This may be caused by the larger spread in distribution of ages in the Fisher dataset. The lowest correlations were observed on the cancer dataset. This is probably caused by the small amount of training instances, the noisy text, and the fact that the ages lie very close to each other.

Overall, the joint model using all features performed best (condition 10). In Figure 2 a plot is presented that relates the true and predicted ages for this condition. We find that for the high ages there are more instances with high errors, probably caused by the small amount of training data for the extreme ages.

We find the correlation metric to be very sensitive to the amount of data. For example, when computing the correlation over the aggregated results of all corpora, we get a much higher correlation (0.830), but the MAE (5.345) is closer to that computed over the individual datasets. However, the MAE is dependent on the age distributions in the corpus, which can be observed by contrasting the MAE on the runs of the Fisher and cancer dataset. This thus suggests that these two measures are complementary and both are useful as evaluation metrics for this task.

For most experiments the joint models show improvement over the individual models. Returning to our question of generality, we can make several observations. First, performance decreases significantly when only using the global features (comparing JOINT and JOINT-Global-retrained), confirming that corpus-specific features are important. Second, learned weights of global features are reasonably generalizable. When using the full feature set, retraining the global features on the corpora directly only gives a slight improvement (e.g. compare conditions 11 and 12). Third, the bias term (β_0) is very corpus-specific and has a big influence on the MAE. For example, when comparing conditions 11 and 12, the correlations are very similar but the MAEs are much lower when the model is retrained. This is a result of adjusting the bias term to the specific dataset. For example the bias term of the model trained on only the blog dataset is 22.45, compared to the bias of 46.11 when trained on the cancer dataset.

In addition, we observe better performance in the cancer dataset when retraining the model using only the global features compared to the initial feature set. This suggests that using the global features might have been an effective method for feature selection to prevent overfitting on this small dataset.

Table 2: Results on the test set, reported with Pearson's correlation (r) and mean absolute error (MAE).

			Blogs		**Fisher**		**Cancer**	
ID	**Model**	**#Features**	***r***	**MAE**	***r***	**MAE**	***r***	**MAE**
Unigrams								
1	INDIV	56,440	0.644	4.236	0.715	7.145	0.426	7.085
2	JOINT	56,440	0.694	4.232	0.723	7.066	0.530	**6.537**
3	JOINT-Global	656	0.605	5.800	0.628	10.370	0.461	16.632
4	JOINT-Global-retrained	656	0.658	4.409	0.675	7.529	0.498	6.797
POS								
5	INDIV	4,656	0.519	5.095	0.553	8.635	0.150	7.699
6	JOINT	4,656	0.563	4.899	0.549	8.657	0.035	8.449
7	JOINT-Global	110	0.495	6.332	0.390	12.232	0.151	19.454
8	JOINT-Global-retrained	110	0.519	5.095	0.475	9.187	0.150	7.699
All features								
9	INDIV	61,416	**0.699**	**4.144**	0.731	6.926	0.462	6.943
10	JOINT	61,416	0.696	4.227	**0.742**	**6.835**	**0.535**	6.545
11	JOINT-Global	510	0.625	5.295	0.650	11.982	0.459	17.472
12	JOINT-Global-retrained	510	0.629	4.633	0.651	7.862	0.490	6.876

5.2 Feature analysis

The most important features using the JOINT model with the full feature set (condition 10) are presented in Table 3. Features associated with a young age have a negative weight, while features associated with old age have a positive weight. For almost all runs and evaluation metrics the full feature set gives the best performance. However, looking at the performance increase, we observe that the unigram only baseline gives strong results. Overall, both stylistic as well as content features are important. For content features, we see that references to family (e.g., "granddaughter" versus "son") as well as to daily life (e.g., "school" versus "job") are very predictive.

Although the performance using only POS tags is lower, reasonable correlations are obtained using only POS tags. In Table 3 we see many POS features associated with old age. This is confirmed when analyzing the whole feature set selected by the JOINT model (condition 10). In this model 510 features are nonzero, 161 of which are POS patterns. Of these, 43 have a negative weight, and 118 have a positive weight. This thus again suggests that old age is characterized more by syntactic effects than young age.

Most important features are consistent with observations from previous research. For example, in the Fisher dataset, similar to findings from classification experiments by Garera and Yarowsky (2009), the word "well" is most predictive of older age. "Like" has the highest association with younger age. This agrees with observations by Barbieri (2008). As was also observed by others, "just" is highly associated with young persons. Consistent with literature that males generally "sound older" than they truly are (Argamon et al., 2007, and others), our male speaker feature has a high negative weight. And, in agreement with previous observations, younger people use more swear words and negative emotions.

The differences between the corpora are reflected in the features that have the most weight. The effective features in the Fisher dataset are more typical of conversational settings and effective features in the cancer dataset are about being pregnant and having kids. Features associated with the blog dataset are typical of the story telling nature of many blog posts.

Comparing the extracted corpus-specific features with the features selected when training on the individual corpora, we do see evidence that the JOINT model separates general versus specific features. For example, the most important features associated with young people in the cancer dataset when only training on the cancer dataset (condition 9) are: *LIWC - Emoticons*, *LIWC - Pronoun*, definitely,

Table 3: Most important features in the JOINT model with all features (condition 10).

(a) Features for younger people.

Global		Blogs		Fisher		Cancer	
like	-1.295	you	-0.387	actually	-0.457	LIWC-Emotic.	-0.188
gender-male	-0.539	went	-0.310	mean	-0.343	young	-0.116
LIWC-School	-0.442	fun	-0.216	everyone	-0.273	history	-0.092
just	-0.354	school	-0.192	definitely	-0.273	mom	-0.087
LIWC-Anger	-0.303	but	-0.189	mom	-0.230	ultrasound	-0.083
LIWC-Cause	-0.290	LIWC-Comma	-0.152	student	-0.182	kids	-0.071
mom	-0.290	go	-0.142	pretty	-0.137	age	-0.069
so	-0.271	POS-vbp nn	-0.116	POS-lrb cd	-0.135	mum	-0.069
definitely	-0.263	thats	-0.115	LIWC-Swear	-0.134	POS-sym rrb	-0.069
LIWC-Negemo	-0.256	well	-0.112	huge	-0.126	discharge	-0.063

(b) Features for older people.

Global		Blogs		Fisher		Cancer	
years	0.601	LIWC - Job	0.514	well	1.644	POS - dt	0.713
POS - dt	0.485	son	0.267	LIWC - WC	0.855	POS - md vb	0.450
LIWC - Incl	0.483	kids	0.228	POS - uh prp	0.504	POS - nn	0.369
POS - prp vbp	0.337	years	0.178	retired	0.492	LIWC - Negate	0.327
granddaughter	0.332	work	0.147	POS - prp vbp	0.430	POS - nn vbd	0.321
grandchildren	0.293	wife	0.142	said	0.404	POS - nnp	0.304
had	0.277	husband	0.137	POS - cc fw	0.358	us	0.287
daughter	0.272	meds	0.112	son	0.353	all	0.266
grandson	0.245	dealing	0.096	subject	0.319	good	0.248
ah	0.243	weekend	0.094	POS - cc cc	0.316	POS - cc nn	0.222

mom, mum, really, *LIWC - Family*, *LIWC - Humans*, thank, and she. The difference in age distribution is reflected in the feature weights. In the JOINT model, the bias term is 24.866. Because most of the persons in the cancer dataset are older, the features associated with young age in the cancer dataset have much lower weights compared to the other datasets.

Because our goal is to compare features across the corpora, we have not exploited corpus-specific features. For example, thread or subforum features could be used for the breast cancer corpus, and for the Fisher dataset, one could add features that exploit the conversational setting of the data.

5.3 Examples

We present examples of text of younger and older persons and connect them to the learned model. The examples are manually selected to illustrate strengths and weaknesses of the model.

5.3.1 Younger people

We first present some examples of text by young persons. The following is an example of a 17-year old in the blog dataset, the system predicted this to be from a 16.48-year-old:

> *I can't sleep, but this time I have school tommorow, so I have to try I guess. My parents got all pissed at me today because I forgot how to do the homework [...]. Really mad, I ended it pissing off my mom and [...] NOTHING! Damn, when I'm at my cousin's I have no urge to use the computer like I do here, [...].*

This example matches with important features determined by the system, containing references to school and parents, and usage of swearing and anger words.

The following are selected turns (T) by a 19-year old (system prediction: 17.37 years) in a conversation in the Fisher dataset.

> *T: yeah it's too i just just freaked out [...]*
> *T: that kinda sucks for them*
> *T: they were they were like going crazy [...]*
> *T: it's like against some law to like*

The text has many informal words such as "kinda" and well as many occurrences of the word "like."

This example is from a 19-year old from the cancer dataset. The system's prediction was far off, estimating an age of 35.48.

> *Im very young and an athlete and I really do not want to look disfigured, especially when I work so hard to be fit. I know it sounds shallow, but Im young and hope to [...] my husband one day :) [...] My grandmother died of breast cancer at 51, and my mother is currently dealing with a cancerous tumor on her ovaries.*

Besides explicit references to being "very young," the text is much more formal than typical texts, making it a hard example.

5.3.2 Older people

The following is a snippet from a 47-year-old (system prediction: 34.42 years) in the blog dataset.

> *[...]In the weeks leading up to this meeting certain of the managers repeatedly asserted strong positions. [...] their previous (irresponsible yet non-negotiable) opinions[...] Well, today's my first Father's day [...]. Bringing a child into this world is quite a responsibility especially with all the fears and challenges we face. [...]*

This matches some important features such as references to jobs, as well as having kids. The many references to the word "father" in the whole text might have confused the model. The following are selected turns (T) by a 73-year old (system prediction: 73.26 years) in a conversation in the Fisher dataset.

> *T: ah thoughts i'm retired right now*
> *T: i i really can't ah think of anyth- think of i would ah ah change considerably ah i'm i'm very i've been very happily married and i have ah three children and six grandchildren*
> *T: yeah that's right well i i think i would do things more differently fair- fairly recently than a long time ago*

This example contains references to being retired and having grandchildren, as well as many usages of "ah". The following is an example of a 70-year old (system prediction: 71.53 years) in the cancer dataset.

> *[...] I was a little bit fearful of having surgery on both sides at once (reduction and lift on the right, tissue expander on the left) [...] On the good side, my son and family live near the plastic surgeon's office and the hospital, [...], at least from my son and my granddaughter [...]*

6 Conclusion

We presented linear regression experiments to predict the age of a text's author. As evaluation metrics, we found correlation as well as mean absolute error to be complementary and useful measures. We obtained correlations up to 0.74 and mean absolute errors between 4.1 and 6.8 years. In three different corpora, we found both content features and stylistic features to be strong indicators of a person's age. Even a unigram only baseline already gives strong performance and many POS patterns are strong indicators of old age. By learning jointly from all of the corpora, we were able to separate generally effective features from corpus-dependent ones.

Acknowledgments

The authors would like to thank the anonymous reviewers for feedback, Michael Heilman for the regression code, and other members of the ARK group for help running the experiments. This work was funded by NSF grants CAREER IIS-1054319 to N.A.S. and IIS-0968485 to C.P.R.

References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of l_1-regularized log-linear models. In *Proc. of ICML*.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat R. Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346.

Shlomo Argamon, Moshe Koppel, James Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: age, gender, and the varieties of self-expression.

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

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *Proc. of LREC*, pages 69–71.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proc. of ACL*.

Penelope Eckert. 1996. Age as a sociolinguistic variable. In *The Handbook of Sociolinguistics*. Oxford: Blackwell.

Jianfeng Gao, Galen Andrew, Mark Johnson, and Kristina Toutanova. 2007. A comparative study of parameter estimation methods for statistical natural language processing. In *Proc. of ACL*.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proc. of ACL-IJCNLP*.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Proc. of ICWSM*.

James W. Pennebaker and Lori D. Stone. 2003. Words of wisdom: Language use over the lifespan. *Journal of Personality and Social Psychology*, 85:291–301.

James W. Pennebaker, Roger J. Booth, and Martha E. Francis, 2001. *Linguistic Inquiry and Word Count (LIWC): A Computerized Text Analysis Program.*

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proc. of SMUC*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*.

Werner Spiegl, Georg Stemmer, Eva Lasarcyk, Varada Kolhatkar, Andrew Cassidy, Blaise Potard, Stephen Shum, Young Chol Song, Puyang Xu, Peter Beyerlein, James Harnsberger, and Elmar Nöth. 2009. Analyzing features for automatic age estimation on cross-sectional data. In *Proc. of INTERSPEECH*.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, 58(1):267–288.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL-HLT*.

Charl van Heerden, Etienne Barnard, Marelie Davel, Christiaan van der Walt, Ewald van Dyk, Michael Feld, and Christian Muller. 2010. Combining regression and classification methods for improving automatic speaker age recognition. In *Proc. of ICASSP*.