Personality profiling from text: language features tied to personality across corpora

William R. Wright
University of Hawaiʻi at Mānoa
Dept. of Information and Computer Sciences
1680 East-West Road, POST 317
Honolulu, HI 96822 USA
wrightwr@hawaii.edu

David N. Chin
University of Hawaiʻi at Mānoa
Dept. of Information and Computer Sciences
1680 East-West Road, POST 317
Honolulu, HI 96822 USA
chin@hawaii.edu

1. INTRODUCTION

Studies that correlate personality with language features use homogeneous data sets from a single domain, such as Facebook posts, emails from one company, and student essays from one university. Such single-domain correlations may not generalize well to other domains. Therefore it is important to study which language features are associated with personality across multiple domains. This paper reports preliminary results from the first such cross-domain study of correlations between personality and language usage.

2. PERSONALITY AND TEXT

Personality traits are consistent patterns in a person's behavior over time—particularly behavior that observers consider when forming an opinion about how an individual's behavior differs significantly from others. A prevailing model of human personality, the Five Factor Model, places such behavior in five dimensions: *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, and *openness*.

Language usage tends to reveal a lot about someone's personality. The advent of computer technology, particularly digital storage and retrieval of text allows us to examine language usage. When relevant, such as in e-mail exchanges, speech acts may predict personality (e.g. the disagreeable person is apt to repeat demands without offering a variety of other speech acts), and punctuation and word sentiment certainly do.

Word frequency (bag-of-words counts) along with overall stem and word counts comprise some of the most intuitive and common features extracted from text. Since word usage is quite context dependent, we are interested in examining aspects of language usage that are less so. Part of speech n-grams preserve information about how a speaker is using language while decoupling from specific words, which are very context dependent. Also n-grams combining both words and part of speech present a compromise between pure word usage and grammatical usage.

3. EXPERIMENTAL DETAILS

We worked with two sets of participants. The first set is a new corpus that we collected ourselves: group of 49 web forum users to whom we administered an personality test consisting of 50 items from the IPIP [2], and gathered their forum postings. The second set that we used was a group of 2,588 university students in North America who each wrote freely for 20 minutes in English [1]. If a writer stopped writing, the computer would stop the clock until typing resumed. The essays span the time period 2005 through 2008, and the average of the essay word counts is 787. Each student also took the Five Factor Inventory, a personality questionnaire. To preserve anonymity, the essays and personality scores are assigned ID numbers in place of participants' names. The two groups differ significantly in the average size of participant texts: 787 for the essays, 57,983 for the forums.

3.1 Features extracted

We extracted various POS n-grams, and sometimes hybrid POS and word n-grams. Table 2 shows a few examples of these features and the text that underlies them. The hybrid features provide context about word usage that simple word counts lack. To extract the features, we first tokenized each participant essay and then extracted the features of interest; when possible used pre-existing tools. Although the statistics computed were straightforward, we chose to use standard, well-tested statistics libraries to avoid errors. The POS tagger we used was an implementation of that presented in [3]; the tagger is trained on manually tagged Wall Street Journal articles.

3.2 Feature Selection

The most populous sample (the Essays corpus) has 2588 participants. We chose a number 1/5th our sample size: 517. We took the 517 most frequent features F in the Essays corpus (the corpus with the most participants) and ignored the rest. Then we extracted F from the Forum texts (49 participants) as well. Of the 517 features, we show here only the ones with $p < 0.1$ for both corpora.

4. RESULTS

To determine the independent relatedness of these features to the personality dimensions, we computed the Pearson correlations, Table 1(c, e), between feature frequencies, normalized to document length, and scores in the given personality dimension, Table 1(a). The p-values tell us the probably of the null hypothesis. An encouraging aspect of this early result is that far more features of interest are related to Conscientiousness than any other personality dimension. This suggests a close relationship between usage of these language features and the speaker's personality in the Conscientiousness dimension. It is difficult to imagine an unrelated process that would cause such a significant difference.

(b)	(c)	(d)	(e)	(f)
Feature	Essays corr	p-val	Forum corr	p-val
(Openness)				
NNS|IN	**-0.041**	**0.0379**	**-0.289**	**0.0440**
DT|lot	**-0.040**	**0.0439**	**-0.291**	**0.0424**
(Conscientious-ness				
JJ|NN	**-0.076**	**0.0001**	**-0.409**	**0.0035**
so|JJ	**0.073**	**0.0002**	**0.358**	**0.0116**
PRP|RB	**0.063**	**0.0014**	**0.282**	**0.0494**
have|RB	**0.062**	**0.0016**	**0.315**	**0.0274**
NN|of	**-0.055**	**0.0049**	**-0.336**	**0.0183**
to|get	**0.054**	**0.0062**	**0.361**	**0.0109**
i|RB	**0.050**	**0.0109**	**0.293**	**0.0407**
VBG|to	**0.047**	**0.0171**	**0.324**	**0.0231**
IN|i	**0.039**	**0.0464**	**0.340**	**0.0167**
(Extraversion)				
BOS|DT	-0.067	0.0006	-0.271	0.0596
it|.	**-0.047**	**0.0177**	**-0.295**	**0.0399**
(Agreeableness)				
for|PRP	**0.077**	**0.0001**	**0.331**	**0.0203**
(Neuroticism)				
JJ|and	**0.043**	**0.0281**	**0.317**	**0.0266**
NNS|IN|DT	**-0.040**	**0.0414**	**-0.286**	**0.0464**

Table 1: Features related to personality across each corpus. Boldface when $p < 0.05$ for both corpora. An index defining each POS tag is available online: www2.hawaii.edu/˜wrightwr/supporting/pos_tags.html

Corpus	Personality Dimension	Personality Score	Feature
forum	open	-1.692	NN|to|VB
see if there's the **possibility to get** more calories			
forum	open	-2.734	NN|to|VB
I have the **tendency to tip** a minimum			
essay	open	-1.281	NN|to|VB
not in the **mood to do** it also having to move out			
essay	cons	2.214	VBG|to|VB
I am **trying to sleep** because I came down with			
forum	cons	2.009	VBG|to|VB
I don't snack often if i'm **trying to lose** weight			
essays	extra	-2.994	DT|NN
should have started writing at **a time** that was easier			
forum	extra	-2.415	DT|NN
high end CPU is a little bit of **a waste** for gaming			
forum	agree	1.307	for|PRP
I have a lot if respect **for you** brah.			
essay	neur	1.624	PRP|feel
I feel like I'm in summer camp.			

Table 2: Instances of language features. Boldface indicates the words associated with the feature label.

For the Essays corpus, the effect sizes are small, whereas for the Forum corpus they are consistently larger, Table 1(c, e). The sparsity and variance of the Essays features may be constraining their predictive impact; Essays features have an average frequency of 4 whereas Forum features have an average frequency of 248. The Forum texts are generally much longer than the Essays, so that constraint is removed.

5. FUTURE WORK

It may be possible to extend this work by including coarse-grained parts of speech (e.g. noun phrases) extracted by chunking tools. Further examination of additional corpora may establish the generalizability of our language features to a variety of populations. Also compelling explanations of why particular POS n-grams are indicative of personality would be of great interest in directing the exploration of new text features useful for personality prediction. Finally, the sparseness of some corpora encourages analysis of features measured by assigning $\{1, 0\}$ when $\{present, absent\}$, a practice that is sometimes useful when working with sparse features.

6. ACKNOWLEDGMENT

We extend grateful thanks to James W. Pennebaker for making this research possible by sharing the essays and personality scores.

7. REFERENCES

[1] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker. Lexical predictors of personality type. In *in 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[2] Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96, 2006.

[3] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.