Differentiating Document Type and Author Personality from Linguistic Features

Scott Nowson
Centre for Language Technology
Macquarie University
NSW 2109 Australia
snowson@ics.mq.edu.au

Jon Oberlander
Division of Informatics
University of Edinburgh
Edinburgh EH8 9LW UK
j.oberlander@ed.ac.uk

Abstract

There are many ways to profile a collection of documents. This paper presents highlight from a body of work that has looked at individual differences in the language of personal weblogs. Firstly, we present a measure of linguistic contextuality that can be used to profile and rank genres. When applied to weblogs, we will show they are similar to school essays, yet significantly less contextual than e-mail. We then look at individual variation of language, as due to the personality of the author. We show that with just a few linguistic features, it is possible to explain significant proportions of variance within personality traits.

Keywords Personalised Documents; Multimedia Resource Discovery

1 Introduction

With the increasing amounts of data available to us via the web, and with new types of documents emerging all the time [7] organising large collections is becoming even less-trivial than it has always been. One obvious target for research is to develop the ability to automatically categorise new documents; to tell *between* one type and another. However, with so much data, it is desirable to have further ways to subdivide categories; to make distinctions *within* text types.

This paper is interested in one specific CMC-based document class, the online journal weblog, or 'blog'.This paper introduces two aspects of a larger study [12] which has looked at linguistic features of blogs.

With so many host services, authors with multiple blogs, and the lack of statistics on non-English language blogs, quoting the number of blogs in existence is difficult. However, as an example of their increasing popularity, the host LiveJournal has seen a 10000% increase in registrations annually from the year 2000 to 2005.

With the emergence of so many different genres on the web [7] there is certainly interest in automatically distinguishing document types [16]. However, the fluidity of genres such as blogs and the freedom for individual expression available to authors means there is a great deal of variation within just this one type. This freedom provides the perfect opportunity for the exploration of variation due to individual differences: in the case of this work, personality traits. Just as automatic identification of text types is a desirable target, so is the automatic differentiation of author types.

This paper presents highlights from a larger body of work investigating the linguistic properties of, and variation within, blogs. It first introduces a unitary measure of contextuality that can be applied to texts. Secondly, it briefly introduces trait theories of personality, the linguistic features used in the study, and the collection of a personality labeled corpus of blogs. It will then discuss the results, beginning by placing blogs in a ranking of text genres based on our unitary measure. It will finish by reporting work which shows that there are linguistic features that can be used to distinguish personality traits.

Proceedings of the 11th Australasian Document Computing Symposium, Brisbane, Australia, December 11, 2006.

2 Background

2.1 Contextuality of language

Heylighen and Dewaele [9] explore the notion of implicitness in text, by developing a unitary measure of contextuality. They considered parts-of-speech as they related to *deixis*: that is to say POSs that require anchoring with spatio-temporal context of utterance in order to be properly interpreted; for example pronouns can generally be considered diectic, or highly contextual, while nouns are (generally) non-diectic, or less contextual. Their F-measure is defined as follows:

$$F = 0.5 * [(nounfrq + adjfrq + prepfrq + artfrq) \\ -(pronfrq + verbfrq + advfrq + intfrq) + 100)]$$

The F-measure was use to explore data derived from multiple language and the results were consistent: spoken language scored lower than written language, meaning that the latter is less contextual; fiction is more contextual than newspapers. Of course, there are other factors which can be used to distinguish *between* genres [2, 10]. However, the F-measure has also been used specifically to investigate individual differences between writers *within* a genre, hence the adoption of this measure.

2.2 Personality traits

This work explores personality from the perspective of Costa and McCrae's five-factor model [6], Each factor gives a continuous dimension for personality scoring. The factors, defined here by their facets [11] are: *Neuroticism* (anxiety, angry hostility, depression, self-consciousness, impulsiveness, and vulnerability); *Extraversion* (warmth, gregariousness, assertiveness, activity, excitement-seeking, and positive emotion); *Openness to experience* (fantasy, aesthetics, feelings, actions, ideas, and values); *Agreeableness* (trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness); and *Conscientiousness* (competence, order, dutifulness, achievement striving, self-discipline, and deliberation)

2.3 Linguistic features

The first approaches employed were content analyses, using categorised dictionaries of words. The Linguistic Inquiry and Word Count (LIWC; [13]) is a collection of psychologically-derived, human-constructed sets of words. It has been used previously to study both language and personality [14] and the language of blogs [4]. The MRC psycholinguistic database [5, 17] was originally developed as a resource for researchers, but was applied in this context following Gill [8]. In addition to these top-down features, bottom-up features are included in the form of POS counts from calculating the F-measure (as described in section 2.1) and distinctive word collocations — bigrams and trigrams that proved to be significantly used by one personality sub-group over another.

3 The weblog corpus

3.1 Construction

A corpus of blog text has been gathered [12]. Participants were recruited directly via e-mail to suitable candidates, and indirectly by word-of-mouth: many participants wrote about the study in their blogs. Participants were first required to answer sociobiographic and personality questionnaires. The personality instrument was specifically designed for online completion [3]. Participants rate themselves on 41-items using a 5-point Likert scale, providing scores for the traits described in section 2.2.

After completing this stage, participants were requested to submit one month's worth of prior weblog postings. This month was pre-specified so as to reduce the effects of an individual choosing their 'best' or 'preferred' month. Raw submissions were marked-up using XML, distinguishing post types such as purely personal, commentary reporting of external matters, or direct posting of internet memes such as quizzes. The corpus consisted of 71 participants (47 females, 24 males; average ages 27.8 and 29.4, respectively) and only the text marked as 'personal' from each weblog, approximately 410,000 words. To eliminate undue influence of particularly verbose individuals, the size of each weblog file was truncated at the mean word count plus 2 standard deviations.

3.2 Personality distribution

A common misconception regarding the personality of bloggers is that they are narcissistic exhibitionists; i.e. Extraverted. This assumption appears to be incorrect, since plotting the distribution of Extraversion scores (figure 1) reveals a relatively normal distribution. However, when Openness scores are plotted (figure 2) there is a significant bias in the sample. It is conceivable that blogger are more Open than average; or perhaps there is response bias. However, without a comparison sample of matched non-bloggers, one cannot say for certain. Due to the statistical complications this creates, Openness is not discussed further in this paper.

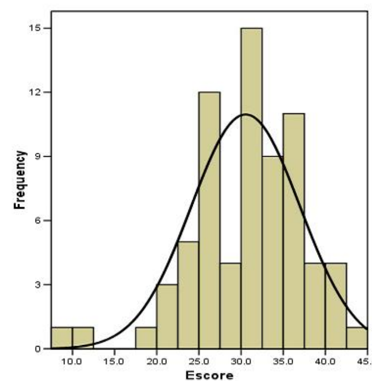

Figure 1: Distribution of Extraversion scores

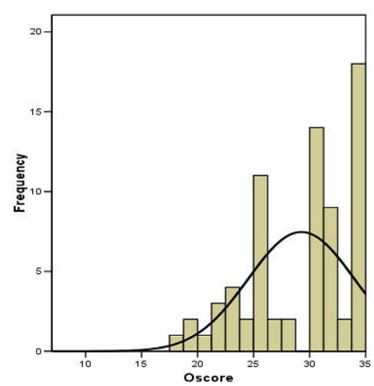

Figure 2: Distribution of Openness scores

4 Between Genres

Looking at blogs as a whole we compare them to a range of genres selected from the British National Corpus (BNC). The BNC consists of over 4000 files, containing over 100 million words of both spoken and written English. Calculating the F-score of a selection of genres from the BNC allows us to place blogs on a scale.

4.1 Method

Using Lee's BNC World Edition Index[1] (2001), 17 genres were selected from the BNC. These included both spoken ($n = 4$) and written ($n = 13$) material. Only files dating from 1985 to 1994 and (for speech) only files with a single speaker were included. Altogether there were 837 files comprised of 23 million words. The original release of the BNC comes pre-tagged, and these tags are algorithmically reduced to the set needed for calculating the F-score of each file. These scores are then averaged to give the F-score of each genre. The F-score for the blog corpus was also computed, and in addition, that of the e-mail corpus of Gill [8].

4.2 Results

When the F-score calculations were completed, the genres ranked as in Table 1. As predicted by Heylighen and Dewaele [9], spoken genres are on the whole more contextual than written, with sermons, lectures, and unscripted speeches scoring the lowest. As expected, unscripted Speeches are more contextual than scripted, while fiction is more contextual than academic writing. Genres appear to be ordered in a plausible manner.

Table 1: Average F-score of selected genres from BNC

Genre	Ave F
Sermons	42.4
Lectures on Social Science	44.3
Unscripted Speeches	44.4
Fiction Prose	46.3
Personal Letters	49.7
Sports Mailing List E-Mails	50.0
E-Mail Corpus	*50.8*
Scripted speeches	53.0
School Essay	53.2
Blog Corpus	*53.3*
Biography	56.3
Non Academic Social Science	56.9
Nat Broadsheet Social	57.5
Professional Letters	57.5
Nat Broadsheet Editorial	58.1
Nat Broadsheet Science	60.0
University Essays	60.3
Academic Social Science	60.6
Nat Broadsheet Reportage	62.2

As one might expect, the e-mail corpus is very similar to the E-Mails taken from the BNC; proximity to Personal Letters follows from this. It can be seen that the blogs are scored as being significantly less contextual than the e-mails (t=3.54, DF=174, p<.001), scoring similarly to School-level essays.

4.3 Discussion

That blogs are more contextual than e-mail can be explained by considering some of the situational factors involved in deixis. Heylighen and Dewaele describe four categories: the *persons* involved, the *space* of the communication, the *time*, and the prior *discourse*. The e-mail corpus consists of two emails per subject, written to a good friend. Blogs however, as a property of being published online, can be read by anyone; hence, to at least some degree, they are written with such readers in mind. Bloggers therefore cannot assume as large a shared context, if any, with their readers as writers of e-mails composed for friends.

Not knowing the reader means the writer can assume less about their knowledge of any places, or *spaces* that are discussed. Similarly, since one cannot know when a reader will encounter their blog, or if they have read it previously, the writer can assume less about the *time* and *discourse* contexts.

In sum, it appears that the F-measure, a measure of contextuality of language, is a reasonable method for distinguishing *between* genres. In fact, the ordering on genres is very similar to that found by Biber [2] when ranking via his involved/informational factor.

5 Within Genre

We have so far explored a method for distinguishing between genres. We now report an exploration into the blog genre considering the personality of the author.

5.1 Method

In section 2.3 we introduced a number of linguistic features, namely the categories of the LIWC and MRC along with word n-grams. Firstly we describe the creation of the n-gram set.

Only 2/3-grams with a corpus frequency $\geq$5 were included to allow accurate log-likelihood G^2 statistics to be computed [15]. Distinct collocations are identified via a three way comparison between the high and low groups (defined as one standard deviation above and below the mean score) of each trait and a third, neutral group. This neutral group contains all those individuals who fell in the medium group for *all four traits in the study*. Hence, this approach selects features using only a *subset* of the corpus. N-gram software was used to identify and count collocations within a sub-corpus [1]. For each feature found, its frequency and relative frequency are calculated. This permits relative frequency ratios and log-likelihood comparisons to be made between High-Low, High-Neutral and Low-Neutral. Only features that prove distinctive for the H or L groups with a significance of $p < .01$ are included in the feature set.

Once all the features were identified the relative frequencies of each were computed for each individual

[1]Available at http://clix.to/davidlee00

author. These were then correlated (Pearsons r) with the personality trait scores. Any features which correlated with at least marginal significance ($p < .1$) were considered to show a relationship with the personality trait in question. This produces a set of related features (drawn from the LIWC, MRC, F-measure and n-grams) for each trait.

In order to explore just how much of a relationship these features had with personality when combined, multiple linear regression was used. For this analysis, the traits are considered the dependent variables, while the correlating features are considered independent. The results of these analyses will provide a further sub-set of features which, when combined, explain the greatest percentage of the variation within the personality scores.

5.2 Result

In mind of space considerations, the full equations resulting from the regression analyses are not included here. Table 2 shows how much of the variance is explained, by how many independent variables along with how significant the result is.

Trait	# of features	R^2	p
N score	10	**.67**	**.000**
E score	8	**.55**	**.000**
A score	8	**.65**	**.000**
C score	8	**.66**	**.000**

Table 2: Multiple regression analysis with personality scores

The third column, the R^2 value, can be seen as the percentage of variance explained by the independent variables. So it is clear that a combination of 10 linguistic features accounts for 67% of the variation in Neuroticism. Similarly, 55% of Extraversion, 65% of Agreeableness and 66% of Conscientiousness can each explained by combinations of 8 features.

5.3 Discussion

These results show that just a small number of linguistic features can account for a great deal of variance. What this shows is that there are linguistic features that can be used to differentiate between personality types. In the case of Conscientiousness for example, calculating the relative frequency of just 8 features in a text offers a reasonably reliable tool to identify high scorers from low. While these results do not translate directly into automatic classification, they are a promising start.

It is interesting to note which features proved most useful. Though exact details are not given here, it must be brought to the readers attention, that the majority of the features retained in the analyses were from the n-gram sets. In fact only 6 of the 34 features were not n-grams. N-gram frequency is trivial to compute for individual documents. This suggests that n-grams would be a reasonable base from which to begin experimentation in automated classification.

It is worth noting that the methodology here is perhaps slightly naïve. The use of the neutral group in identifying the distinct collocations was intended to minimise over-fitting in the correlation and regression analyses. However, it remains the case that there were only 71 subjects, and data-sparseness is likely.

6 Final words

There are many ways to separate documents. This paper has considered doing so by genre, as well as by author type. The unitary measure employed here, the F-measure, whilst perhaps not lending itself to automatic classification, is a useful way to consider the differences between genres. It has proved particularly useful in highlighting the differences between the CMC genres of blogs and e-mails. In the second study reported we have shown that there are features which can be used to detect personality traits. In combination, these explains considerable levels of variation within the language used by different personality types. This suggests that it might not be such a wild idea to consider the automatic classification of text by author personality.

Acknowledgements We would like to thank Robert Dale for his invaluable comments and encouragement. The first author also acknowledges funding from the UK Economic and Social Research Council.

References

[1] Satanjeev Banerjee and Ted Pedersen. The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, 2003.

[2] Douglas Biber. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, 1988.

[3] Tom Buchanan. Online implementation of an ipip five factor personality inventory. Available at `http://users.wmin.ac.uk/~buchant/wwwffi/introduction.html`, *accessed 07/10/06*, 2001.

[4] Michael A. Cohn, Matthias R. Mehl and James W. Pennebaker. Linguistic markers of psychological change surrounding september 11. *Psychological Science*, Volume 15, pages 687–693, 2004.

[5] M. Coltheart. The mrc psycholinguistic database. *Quarterly Journal of Experimental Psychology*, Volume 33, Number A, pages 407–505, 1981.

[6] Paul T. Costa and Robert R. McCrae. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Odessa, FL: Psychological Assessment Resources, 1992.

[7] Kevin Crowston and Marie Williams. Reproduced and emergent genres of communication on the world wide web. *The Information Society*, Volume 16, Number 3, pages 201–216, 2000.

[8] Alastair J. Gill. *Personality and Language: The projection and perception of personality in computer-mediated communication*. Ph.D. thesis, University of Edinburgh, 2004.

[9] Francis Heylighen and Jean-Marc Dewaele. Variation in the contextuality of language: an empirical measure. *Foundations of Science*, Volume 7, pages 293–340, 2002.

[10] Max Louwerse, Philip M. McCarthy, Danielle S. McNamara and Arthur C. Graesser. Variation in language and cohesion across written and spoken registers. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 1035–1040, Hillsdale, NJ, 2004. LEA.

[11] Gerald Matthews, Ian J. Deary and Martha C. Whiteman. *Personality Traits*. Cambridge University Press, Cambridge, 2nd edition, 2003.

[12] Scott Nowson. *The Language of Weblogs: A study of genre and individual differences*. Ph.D. thesis, University of Edinburgh, 2006.

[13] James W. Pennebaker and Martha E. Francis. *Linguistic Inquiry and Word Count: LIWC*. Erlbaum Publishers, 1999.

[14] James W. Pennebaker and Laura King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, Volume 77, pages 1296–1312, 1999.

[15] Paul Rayson. *Wmatrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University, 2003.

[16] Maria Santini. Clustering web pages to identify emerging textual patterns. RECITAL 2005, Dourdan, 2005.

[17] Michael Wilson. MRC psycholinguistic database: Machine usable dictionary. Technical report, Oxford Text Archive, Oxford, 1987.