Readability for author profiling?
Notebook for PAN at CLEF 2013

Lee Gillam

Department of Computing, University of Surrey, UK
l.gillam@surrey.ac.uk

Abstract. This paper briefly describes the approach taken to the Author Profiling task at PAN 13. It describes the simple features used, and the origins in thinking around text readability as a mechanism for identification, and the predictive model used which may have beneficially omitted classes, as well as offering commentary on the results obtained.

1 Introduction

The Author Profiling task was new to PAN2013, albeit with some relationship to the PAN2012 cyberpredator task (e.g. Vartapetiance and Gillam 2012). The two output features of interest for this task are age and gender, with only 10s, 20s and 30s as suitable values for age, and male or female for gender, for a collection of English and Spanish blogs and product reviews.

Our approach is based on simple text readability characteristics, eschewing linguistic processing, and some fairly straightforward assumptions. We achieved total accuracy of 0.32 (8^{th} position at time of writing – of 21, though with 64 registered) for English, against the best accuracy overall of 0.389, and accuracy of 0.25 (13^{th}) versus best of 0.42 for Spanish. Notably, we achieved the fastest runtime overall of 615,347.

In this short paper, we outline the approach taken at the University of Surrey to Author Profiling at PAN 13. First, in section 2, we offer a brief overview of text readability as provides the inspiration for our approach. In section 3, we briefly state our simple approach. Section 4 shows results obtained, and Section 5 concludes with considerations for future work.

2 Text Readability

Elsewhere, we have published on text readability for authoring assistance (Newbold and Gillam 2010a), in relation to video annotation (Newbold and Gillam 2010b) and for ranking in information retrieval with one of the forefathers of text readability measures (Newbold, McLaughlin and Gillam 2010). We have presented (Newbold and Gillam 2010b, Table II), core characteristics of common measures of text readability – namely, of Flesch, Kincaid, Fog, SMOG and ARI. In their

formulation, all use sentence length and either the number of characters per word, number of syllables per word, or a count of the number of complex words. To determine whether readability could be more broadly applied to author profiling, and therefore whether more of our prior work might be of benefit to such a task in future, we settled on a simple measurable combination for the distributions of sentence lengths and word lengths in the texts – which would relate most closely with ARI. A simply hypothesis, here, would be that longer words would be used with greater frequency at higher ages, and that females might write in longer sentences than males.

Results obtained do indicate greater scope for exploration of text readability as a proxy for author age and gender, albeit with a large degree of uncertainty over how far one could push such an approach or in which direction. The approach used is described briefly in the next section.

3 Author Profiles

Our approach, coded in Python – and so potentially with scope for better speed – first removes all markup, and then uses ".", "!" and "?" characters as sentence delimiters. For sentences of more than 35 characters, we get a simple approximation of word length as character length divided by 6, and collect distributions of both sentence lengths and word lengths for the text.

From the two collected distributions, we produce one number each for sentence length and word length by summing the numbers of sentences/words seen at 50% and 90% (i.e. approximating average sentence/word length plus a value of tendency towards using long sentences and words). Clearly such an approach is not likely to be reliable when just a few sentences are involved, but no further analysis has yet been done to ascertain accuracy as sentence number increases.

To determine rules by which to label, we made use of the popular Weka software from the University of Waikato, and in particular the J48 implementation of the C4.5 decision tree. We were looking for a simple tree to offer good generalization (complex trees, suggesting overfitting, would not be suitable).

It was noticeable from this that training data were not evenly distributed, with data for 30s age group tending to dominate and hence to bias rules produced, The somewhat weak quantity of 10s data tends to explain why the label does not get produced in the resulting rule – and if such data imbalance is carried into testing data would suggest something else about a suitable baseline.

Primarily through trial and error, it was clear that treating age and gender separately produced the simplest tree(s). The resulting rules are:

AGE:

```
if( word <= 10):  return "20s"
else: if( sentence <= 108):  return "30s"
      else: if( word <= 11): return "20s"
            else: return "30s"
```

```
GENDER:
    if( sentence <= 28):
        if( word <= 18): return "male"
        else:
            if( sentence < 17):
                if( word <= 21): return "female"
                else: return "male"
            else:
                return "male"
    else:
        if(word <= 11): return "male"
        else: return "female"
```

We did not differentiate English and Spanish, so variation in results across these collections may suggest something about the need to do so in future – and consequently, for our own interests, for Spanish readability analysis.

4 Results

The key numbers of interest relate to accuracy. Our approach, which eschews linguistic processing, achieves total accuracy of 0.32 (8th position at time of writing – 21 participants) for English, against the best accuracy overall of 0.389, and accuracy of 0.25 (13th) versus best of 0.42 for Spanish. Notably, we achieved the fastest runtime overall of 615,347 milliseconds for the entire collection (next fastest was 1,729,618ms).

5 Conclusions

Although only 8th, the simplicity of the approach suggests that there may be greater scope for assessing the use of more complex readability approaches in future, of the kinds we have covered in several previous papers. However, it is not certain at the time of writing how much of a part is played by any imbalances in data presented since data characteristics are not available. Baseline figures are considered as 1/3 for age and 1/2 for gender. However, if 30s data were dominant, and severely so in contrast to 10s, then simply selecting 30s each time could produce quite good figures for accuracy. For the approach considered here, we ended up never labeling 10s, and so would consider our own baseline at 0.25 overall (in contrast to what should be 0.167 if all were equal). Without the data, it is difficult to tell whether 0.32 is a good score, or just an artifact of these rules applied to an unbalanced set of data.

Acknowledgements

The author gratefully recognizes prior contributions of Neil Newbold in relation to text readability, and also of Harry McLaughlin, originator of the SMOG measure in 1969 ("SMOG Grading — a New Readability Formula" (PDF). Journal of Reading 12 (8): 639–646).

This work has been supported in part by the EPSRC and JISC (EP/I034408/1) the UK's Technology Strategy Board (TSB, 169201), and also the efforts of the PAN13 organizers in system provision and managing the submissions.

References

Newbold, N. and Gillam, L., 2010a, The Linguistics of Readability: The Next Step for Word Processing. Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids (CL&W 2010). At NAACL-HTL 2010.

Newbold, N. and Gillam, L., 2010b, Text Readability within Video Retrieval Applications: A Study On CCTV Analysis. Journal of Multimedia 5(2):123-141 (Special Issue on Visual Information Engineering), Academy Publisher.

Newbold, N., McLaughlin, H. and Gillam, L., 2010, Rank by Readability: Document Weighting for Information Retrieval. Information Retrieval Facility Conference 2010. Available from Springer LNCS 6107

Vartapetiance, A. and Gillam, L., 2012, Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification. Proceedings of the 4th PAN workshop