Comparing Sentence-Level Features for Authorship Analysis in Portuguese

Rui Sousa Silva[1,3], Luís Sarmento[2],
Tim Grant[1], Eugénio Oliveira[2], and Belinda Maia[3]

[1] Centre for Forensic Linguistics at Aston University
[2] Faculdade de Engenharia da Universidade do Porto - DEI - LIACC
[3] CLUP - Centro de Linguística da Universidade do Porto

Abstract. In this paper we compare the robustness of several types of stylistic markers to help discriminate authorship at sentence level. We train a SVM-based classifier using each set of features separately and perform sentence-level authorship analysis over corpus of editorials published in a Portuguese quality newspaper. Results show that features based on POS information, punctuation and word / sentence length contribute to a more robust sentence-level authorship analysis.

1 Introduction

Authorship analysis consists in identifying an author from a limited number of candidates (see [2]), and is increasingly relevant in cases of plagiarism detection and information filtering. Previous research on forensic linguistics has shown that there are several stylistic markers that help determine the authorship of a text *independently* of the topic of those texts. For example, Eagleson [1] claims that context-independent features are related to grammatical and lexical choices, including *syntactic structure*, *morphological inflections*, *vocabulary*, *spelling* and *punctuation*. Grant [2] uses a sophisticated Discriminant Function Analysis (DFA) to determine which variables help discriminate between the texts of three different authors, i.e. which variables are the best predictors to attribute texts to authors. He concluded that DFA is able to tell which of three authors is most likely to have written a queried text, and obtain an indication of the weight of evidence for each attribution, using a further analysis of the probabilities. He further concluded that the system proposed, both strong and conservative, is trade-off between a robust method against mis-attribution and the conservatism in terms of the number of texts not firmly attributed. Hirst and Feiguina [3] perform authorship discrimination based on syntactic analysis, in particular on the frequency of bigrams of syntactic labels, obtained from partial parsing of the text, that they treat as pseudo-words, and consider their relative frequencies. They concluded that bigrams of syntactic labels are more discriminating that other features such as frequencies of rewrite rules, even with fragments of little more than 200 words (in which case the accuracy was boosted by using features such as unigram POS frequencies). All these however require strings of text of considerable length.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 51–54, 2010.

In this exploratory study, we investigate authorship analysis in Portuguese texts at *sentence level*. Performing authorship analysis at this level raises the additional problem that style markers should be able to work with short text strings and intra-sentence information only (i.e. very low frequency counts). We compare the robustness of several types of stylistic markers extracted at sentence-level to help discriminate the authorship of sentences using an SVM (*Support Vector Machine*)-based classifier over a corpus of editorials published in a Portuguese quality newspaper.

2 Stylistic Features for Authorship Analysis

We focus on the following potential and observable markers of authorship, which are to be extracted at *sentence level* with minimum linguistic processing:

- *POS-based features*: Computation of the frequency of each POS label found in the sentence, including function words and tense information in the case of verbs. Words found "POS-ambiguous" and "unknown" (e.g. neologisms) are also included in the POS-based features since their discriminatory power is potentially relevant.
- *Punctuation*: Frequency information about the usage of commas, "strong" punctuation marks, quotes and brackets.
- *Length*: Quantitative features such as the number of characters per word, the number of words per sentence, the number of 1 to 20-letter words and the number of words of 20+ letters.
- *Suffixation - superlatives and diminutives*: Suffixes are found vary greatly among authors (i.e., they are largely idiolectal) in that they act as optional modifiers. We consider two particular forms of affixation: superlatives and diminutive forms.
- *Pronouns*: Information about explicit usage of relative and personal pronouns, whose use is dependent on the individual choices of the authors.
- *Conjunctions*: Information about the use of seven types of conjunctions, as the use of dependent and independent clauses is also highly idiolectal.

It is important to emphasise that *all* feature sets listed are *content agnostic*, which is intended to isolate our experiments from the potentially significant impact that content could have on authorship attribution, especially if columnists tend to focus their posts on certain preferential topics (e.g. economics vs. politics).

3 Experimental Set-Up

We built a corpus of editorials and opinion articles posted by columnists of a Portuguese quality daily newspaper[1]. The corpus comprises 915 posts by 23

[1] Jornal de Notícias – `http://www.jn.pt`

commentators from November 2008 to September 2009. From these we selected the top three most productive columnists - denoted by C_1, C_2, C_3 - who write editorials more than once a week covering a wide variety of issues, and not specialised in any specific topic domain. Commentator C_1 writes many short editorials (176 posts with 5.1 sentences / post), while commentators C_2 and C_3 write less frequently (74 and 51 posts each), but often longer editorials (17.9 and 17.7 sentences / post respectively).

We then randomly selected a set of 750 sentences for each of these commentators and trained a binary classifier to discriminate sentences written by each commentator. The authorship of a given test sentence is thus determined by the binary classifier that produces the highest classification score. Since the scores produced by all the three binary classifiers may be quite low – which can reflect the fact that the classifier has a very low confidence level in its result or the sentence is somehow difficult to differentiate – we introduced a threshold on the minimum value of classification score to be considered valid, c_{min}. Only classification scores higher than c_{min} were considered, which means that if none of the three classifiers reaches that threshold the test sentence at stake remains *unclassified*, i.e. no authorship is attributed to it.

We opted for using SVM as the classification algorithm for their well-known robustness in several text classification settings. We used the *SVM-light* [4] implementation. In all our experiments we used the default *SVM-light* parameters (including the choice for a linear kernel). In order to obtain Precision vs. Recall curves, we attempted authorship attribution with different values on the threshold c_{min}. We performed 5-fold cross-validation in all our experiments, and we micro-averaged partial Precision and Recall results.

4 Results and Analysis

We ran the training and classification procedure using *only one* of the six subgroup of markers described in Section 2 at a time. Figure 1 presents the precision vs. recall curves for all the six runs plus an additional curve for the run made using *all* stylistic markers. As expected, the performance obtained using any of the subgroup of markers alone is lower than the performance obtained using all the stylistic features. Among all groups of markers, *POS-based* ones seem to carry more information, performing almost as well as all subgroups of markers together. Two other groups perform reasonably well alone: *Punctuation* and *Length*. Interestingly, these subgroups use practically *no lexical information*, but instead a rather simple statistics related to punctuation and word / sentence length. The other groups of stylistic markers are not so *robust* (i.e., their performance drops sharply) since they tend to occur in only a limited number of sentences. However, it is important to emphasize that *all* groups of features seem to carry some relevant information for authorship analysis, as the results obtained using all groups of markers shows.

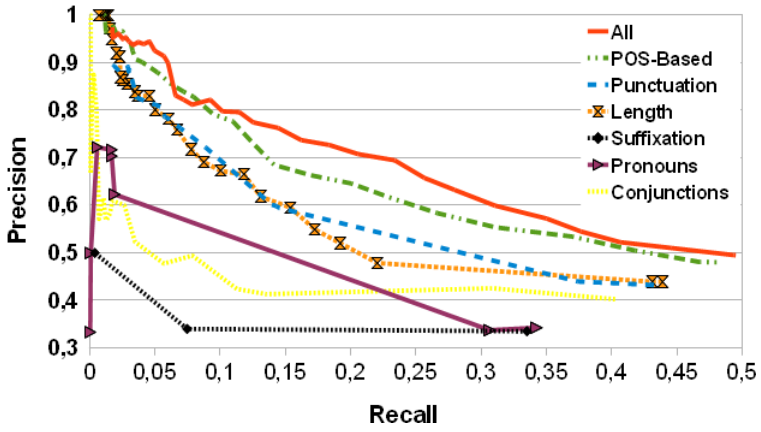

Fig. 1. Precision vs. Recall curves for each subgroup of stylistic markers

5 Conclusions and Future Work

This experiment confirms our initial assumptions that *content-agnostic* features can effectively be used for authorship analysis at sentence level. Among all the stylistic features, the excellent performance rate of punctuation stands out, which demonstrates that punctuation is one of the most robust stylistic features analysed. Affixes (superlatives and diminutives) and pronouns do not demonstrate enough robustness to perform *sentence-level* authorship attribution. Unsurprisingly, simple quantitative data (i.e. word and sentence length) perform well overall, with results that are similar to those obtained by punctuation.

Acknowledgments

This work was partially supported by grant SFRH/BD/23590/2005 FCT-Portugal, co-financed by POSI, and by grant SFRH/BD/47890/2008 FCT-Portugal, co-financed by POPH/FSE.

References

1. Eagleson, R.: Forensic analysis of personal written texts: a case study. In: Gibbons, J. (ed.) Forensic Linguistics: An Introduction to Language in the Justice System, pp. 362–373. Longman, Harlow (1994)
2. Grant, T.: Quantifying evidence in forensic authorship analysis. The International Journal of Speech, Language and the Law 14(1), 1–25 (2007)
3. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Lit Linguist Computing 22(4), 405–417 (2007)
4. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)