collectively, and to follow the hacking tenets set out by Steven Levy: 'All information should be free. Mistrust authority – promote decentralization. You can create art and beauty on a computer.'

Evaluating Unmasking for Cross-Genre Authorship Verification

Kestemont, Mike
mike.kestemont@ua.ac.be
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Luyckx, Kim
kim.luyckx@ua.ac.be
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Daelemans, Walter
walter.daelemans@ua.ac.be
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Crombez, Thomas
thomas.crombez@ua.ac.be
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

In this paper we will stress-test a recently proposed technique for computational authorship verification, 'unmasking' (Koppel et al. 2004, 2007), which has been well received in the literature (Stein et al. 2010). The technique envisages an experimental set-up commonly referred to as 'authorship verification', a task generally deemed more difficult than so-called 'authorship attribution' (Koppel et al. 2007). We will apply the technique to authorship verification across genres, an extremely complex text categorization problem that so far has remained unexplored (Stamatatos 2009). We focus on five representative contemporary English-language authors. For each of them, the corpus under scrutiny contains several texts in two genres (literary prose and theatre plays).

1. Background: cross-genre authorship verification

In authorship verification, the given text may have been written by one of the candidate authors, but could also be written by none of them. Note that this *open case* scenario is typical of forensic applications: the author of e.g. a bomb letter is not necessarily among the suspect candidate authors. In the case of a suicide letter (potentially faked by a murderer), it is highly likely that this is the only suicide letter the victim ever wrote. In absence of similar material, it is difficult to extract reliable style markers from pre-existing writings to determine authorship of the letter.

Authorship *across genres* is an issue that is being paid all too little attention in present-day research. The few remarks that have been made on this issue agree that authorship attribution is difficult within a single textual genre, even more difficult with several topics involved, and likely to be extremely difficult with several genres involved (Luyckx & Daelemans 2011). Although it is generally assumed that an author will display stable style characteristics throughout his oeuvre, irrespective of genre, this remains speculative in the absence of systematic empirical investigation. Consequently, cross-genre authorship verification deserves much more attention than it has attracted so far.

2. Unmasking

Unmasking is a fairly complex meta-learning approach to authorship verification. Koppel et al. (2007) observed in earlier experiments that a small number of features had a lot of discriminatory power. It is indeed common for authors to use 'a small number of features in a consistently different way between works'. Such features often relate to topic-related, narrative, or thematic differences. As a result, a limited number of features can wrongfully maximize the differences in writing style between two works of identical authorship.

The unmasking approach tests the *robustness* of a stylistic model by deliberately impairing it over a number of iterations, each time removing those features that are most discriminative between the two texts. The resulting 'degradation curves' display many sudden drops in accuracy: when the most telling features are removed during each iteration, it becomes increasingly difficult to differentiate between two texts. In the case of two texts of non-identical authorship, however, a far larger number of features is discriminative, causing less dramatic drops in accuracy during degradation. Using training material in the form of a series of same-author and different-author degradation curves, Koppel et al. (2007) try to verify whether previously unseen degradation curves are of (non-)identical authorship.

The unmasking technique is especially attractive for authorship verification across genres, because of the interference between genre markers and authorial style markers. It might help remedy genre-related artifacts in that superficial genre-related differences between same-author texts in different genres will be filtered out easily and removed from the model early in the degradation process. After the removal of these non-essential stylistic features, one could hypothesize that only features more relevant for authorial identity will be preserved.

3. Methodology and evaluation

Our unmasking implementation closely adheres to the original description of the procedure. In the experiments, we have used the same generic parameter settings as tentatively adopted by Koppel et al. (2007): a *chunk size* of 500 tokens, n=250, m=10 and k=3. The main difference is, that a 'leave-one-text- out validation' is carried out on these curves for evaluation purposes, whereas k-fold cross-validation was applied in the original paper. We train an SVM classifier on the training curves and have it classify each of the test curves as a *same-author* or *different-author* curve. When all predictions have been collected, one can report on the overall classification accuracy and macro-averaged F1-score.

4. Corpus and selection of texts

The corpus we collected for the experiments in cross-genre authorship verification consists of published texts by five contemporary authors: Edward Bond, David Mamet, Harold Pinter, Sam Shepard, and Arnold Wesker. The main criterion for selecting an author was the availability of texts in more than one literary genre. Theatre and prose were the genres these five authors were most productive in, so these were chosen for the experiments. In our corpus, applying a text length threshold of 10,000 words (cf. Sanderson & Guenter 2006) resulted in 11 prose works and 23 theatre plays. We experimented with a complete matrix of authors and genres, allowing both intra-genre and cross-genre experiments for all authors. Digitization of the material involved three steps: scanning, OCR'ing, and manual post-correction.

5. Intra-genre experiments

Figure 1 shows degradation curves for an experiment on the eleven prose works in the corpus. Solid lines represent *same-author* curves, whereas dotted lines represent *different-author* curves. All curves display downward slopes, with decreasing cross-validation accuracies, as more predictive features get eliminated in each iteration. For *same-author* curves, however, it is clearly visible that the effect of degradation generally sets off sooner and more dramatically. *Different-author* curves are more robust and yield higher cross-validation accuracies, even when a large number of strongly discriminative features is deleted. Intersections between both curve types are minimal. A leave-one-text-out validation test on this set of curves confirms the success of the approach: the overall accuracy amounts to 96%, which is only just over the F1 score of 95%. This result confirms the

potential of unmasking for authorship verification in prose work collections.

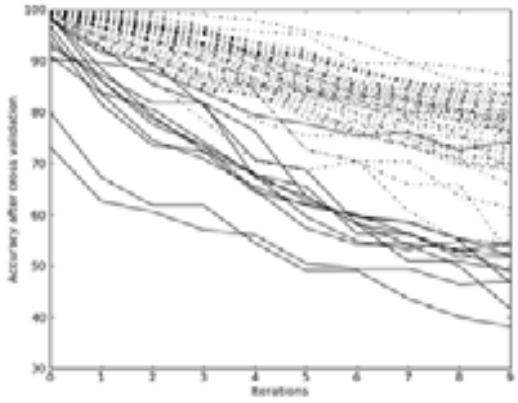

Figure 1: Unmasking on prose texts by five authors

A second experiment has been carried out on the 23 theatrical works in the corpus. Figure 2 displays a much less clear-cut differentiation of the *same-author* curves and their *different-author* counterparts, suggesting that the unmasking approach (with its default settings) is less effective for the theatrical section of the corpus. The leave-one-text out validation confirms this, yielding an overall accuracy of 84% and an F1 score of 62%.

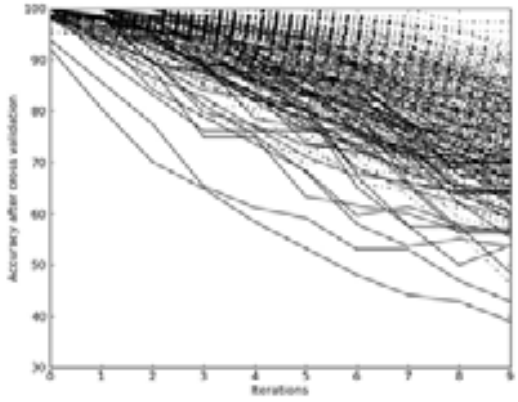

Figure 2: Unmasking on theatre plays by five authors

6. Cross-genre experiments

So far, the unmasking procedure has been mainly investigated for text pairs within the same text variety, although Koppel *et al.* (2007) report on a successful application of the technique to Hebrew-Aramaic texts across different topics. It is an interesting question whether the degradation differences between *same- author* and *different-author* curves would also hold for pairs of texts that do not belong to the same genre. A leave-one-text-out validation, however, shows poor performance of unmasking in this experiment, with an overall accuracy of 77% and a macro-averaged F1 of 56%.

7. Interpretation

After unmasking has been applied, the individual degradation curves allow for interpretation of results. Figure 3 visualizes the elimination process for Pinter's play *The Caretaker* and Mamet's prose text *The Old Religion*, who were personal friends. Mamet even acknowledged Pinter as a key influence on his work. The limited degradation in accuracy demonstrates that these Mamet and Pinter texts appear to adopt well-distinguishable styles. Figure 3 shows early elimination of names of principal characters (*davies*, *mick* and *aston* vs. *mark* and *pete*), personal pronouns that relate to a text's narrative perspective (*i, you*), and colloquial language (*aint*). Moreover, typical genre-features (e.g. the director's indication *pause*) are deleted as anticipated.

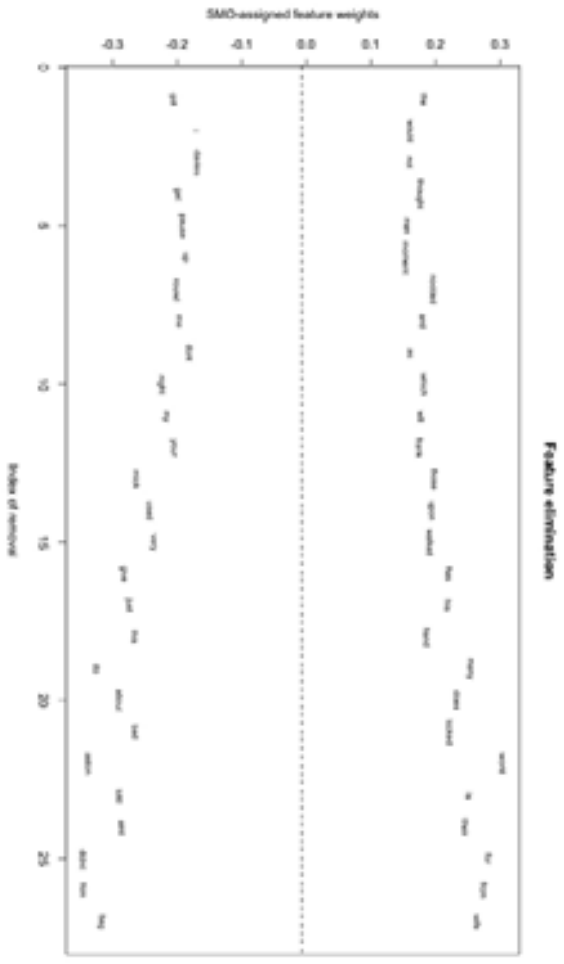

Figure 3: Visualization of the feature elimination process for Pinter's play The Caretaker and Mamet's prose text The Old Religion

8. Conclusion

The experiments reported on in this paper confirm that unmasking is an interesting technique for computational authorship verification, especially yielding reliable results within the genre of (larger) prose works in our corpus. Authorship verification, however, proves much more difficult in the theatrical part of our corpus. The original settings for the various parameters often appear to be genre-specific or even author-specific, so that further research on optimization is desirable. Finally, we have shown that interpretability is an important asset of the unmasking technique.

Funding and Acknowledgements

Kestemont is a Ph.D. fellow of the Research Foundation – Flanders (FWO). The research of Luyckx and Daelemans is partially funded by the FWO project 'Computational Techniques for Stylometry for Dutch'. The research by Crombez is partially funded by the FWO project 'Mass Spectacle in Flanders'. The authors would like to acknowledge Sarah Bekaert's work on the digitization of the corpus.

References

Koppel, M., J. Schler, and E. Bonchek-Dokow (2007). Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research* 8: 1261-1276.

Luyckx, K., and W. Daelemans (2011). The Effect of Author Set Size and Data Size in Authorship Attribution. *Literary and Linguistic Computing* 26(1): 35-55.

Sanderson, C., and S. Guenter (2006). Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia, pp. 482-491.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60(3): 538-556.

Stein, B., N. Lipka, and P. Prettenhofer (2011). Intrinsic Plagiarism Analysis. *Language Resources and Evaluation* 45(1): 63-82.

Koppel, M., and J. Schler (2004). Authorship Verification as a One-Class Classification Problem. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada,

Literary Wikis: Crowd-sourcing the Analysis and Annotation of Pynchon, Eco and Others

Ketzan, Erik
eketzan@lavabit.com
Institut für Deutsche Sprache, Germany

1. Introduction

Annotation of complicated texts – the *Bible*, the works of Shakespeare, experimental fiction – is a familiar concept. In 2005 I had an idea: what if I used a wiki to create such a guide? Would anyone contribute to it? Would anyone read it? Would the results be any good? Since then, hundreds of users have annotations thousands of pages of experimental fiction by the authors Thomas Pynchon and Umberto Eco, and the projects *PynchonWiki* and *Umberto Eco Wiki*.

What have we learned from these projects? What can other digital humanities and crowd-sourcing projects learn from their successes and failures? The Eco wiki project is still ongoing, but I plan to present my findings and insights at DH 2012.

2. The Queen Loana Wiki

When Umberto Eco's novel, *The Mysterious Flame of Queen Loana,* was published a few months later, I launched what I called the *Queen Loana Annotation Project*, a wiki organized by chapter and page.[1] Eco's novel was a perfect test case for the experiment I envisioned, a literary annotation wiki. First, Eco frequently quotes texts without attribution, which led to wiki entries like:

P. 15, 'you always said you could resist anything but temptation'

quotation from *Lady Windermere's Fan* by Oscar Wilde.

Second, many references in the novel were confusing to readers, making my wiki a *useful* resource. According to a *Village Voice* review at the time, 'Early reviews have dismissed *Mysterious Flame* as nostalgic and at times so personal as to be impenetrable. Eco concedes he wrote it with his own generation in mind. "It's a book for Italian people of my age".' Thanks to the wiki, though, readers could easily read up on all those references, with entries like this: