Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation

Conrad Sanderson and Simon Guenter
Australian National University, Canberra, ACT 0200, Australia
National ICT Australia, Locked Bag 8001, ACT 2601, Australia

Abstract

We present an investigation of recently proposed character and word sequence kernels for the task of authorship attribution based on relatively short texts. Performance is compared with two corresponding probabilistic approaches based on Markov chains. Several configurations of the sequence kernels are studied on a relatively large dataset (50 authors), where each author covered several topics. Utilising Moffat smoothing, the two probabilistic approaches obtain similar performance, which in turn is comparable to that of character sequence kernels and is better than that of word sequence kernels. The results further suggest that when using a realistic setup that takes into account the case of texts which are not written by any hypothesised authors, the amount of training material has more influence on discrimination performance than the amount of test material. Moreover, we show that the recently proposed author unmasking approach is less useful when dealing with short texts.

1 Introduction

Applications of authorship attribution include plagiarism detection (e.g. college essays), deducing the writer of inappropriate communications that were sent anonymously or under a pseudonym (e.g. threatening or harassing e-mails), as well as resolving historical questions of unclear or disputed authorship. Specific examples are the Federalist papers (Hanus and Hagenauer, 2005; Mosteller, 1984) and the forensic analysis of the Unabomber manifesto (Foster, 2001).

* *Appears in Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia, 2006, pp. 482–491. Published by Association for Computational Linguistics (ACL).*

Within the area of automatic author attribution, recently it has been shown that encouraging performance can be achieved via the use of probabilistic models based on n-grams (Clement and Sharp, 2003) and Markov chains of characters and words (Peng et al., 2004). Diederich et al. (2003) showed that Support Vector Machines (SVMs), using the bag-of-words kernel, can obtain promising performance, while in another study, SVMs with kernels based on character collocations obtained mixed performance (Corney, 2003). Gamon (2004) utilised SVMs with syntactic and semantic features to obtain relatively minor accuracy improvements over the use of function word frequencies and part-of-speech trigrams. Koppel & Schler (2004) proposed a word-level heuristic, resembling recursive feature elimination used for cancer classification (Guyon et al., 2002; Huang and Kecman, 2005), to obtain *author unmasking* curves. The curves were processed to obtain feature vectors that were in turn classified in a traditional SVM setting.

The studies listed above have several limitations. In (Clement and Sharp, 2003), a rudimentary probability smoothing technique was used to handle n-grams which were unseen during the training phase. In the dataset used by (Peng et al., 2004) each author tended to stick to one or two topics, raising the possibility that the discrimination was based on topic rather than by author style.

In (Corney, 2003; Gamon, 2004; Peng et al., 2004; Koppel and Schler, 2004) the datasets were rather small in terms of the number of authors, indicating the results may not be generalisable. Specifically, in (Corney, 2003) the largest dataset contains texts from five authors, in (Gamon, 2004) from three, while in (Peng et al., 2004) and (Koppel and Schler, 2004) from ten.

In (Clement and Sharp, 2003; Gamon, 2004; Peng et al., 2004), the attribution of a given document was forced to one of the authors from a set of possible authors (i.e. a closed set identifi-

cation setup), thus not taking into account the realistic case of having a document which was not written by any of the authors. In (Koppel and Schler, 2004), the unmasking method was evaluated exclusively on books, raising the question as to whether the method is applicable to considerably shorter texts.

Lastly, all of the studies used different datasets and experiment setups, thus making a quantitative performance comparison of the different approaches infeasible.

Recently, various practical character and word sequence kernels have been proposed (Cancedda et al., 2003; Leslie et al., 2004; Vishwanathan and Smola, 2003) for the purposes of text and biological sequence analysis. This allows kernel based techniques (such as SVMs) to be used in lieu of traditional probabilistic approaches based on Markov chains. In comparison to the latter, SVMs have the advantage of directly optimising the discrimination criterion.

This paper has four main aims: **(i)** to evaluate the usefulness of sequence kernel based approaches for the task of authorship attribution; **(ii)** to compare their performance with two probabilistic approaches based on Markov chains of characters and words; **(iii)** to appraise the applicability of the author unmasking approach for dealing with short texts; and **(iv)** to address some of the limitations of the previous studies.

Several configurations of the sequence kernels are studied. The evaluations are done on a relatively large dataset (50 authors) where each author covers several topics. Rather than using long texts (such as books), in almost all of the experiments the amount of training and test material per author is varied from approx. 300 to 5000 words for both cases. Moreover, rather than using a closed set identification setup, the evaluations are done using a verification setup. Here, a given text material is classified as either having been written by a hypothesised author or as not written by that author (i.e. a two class discrimination task).

The paper is structured as follows. Section 2 describes author attribution systems based on Markov chains of characters and words, followed by a description of the corresponding sequence kernel based approaches in Section 3. Section 4 provides an empirical performance comparison of the abovementioned approaches, while in Section 5 the author unmasking method is appraised. Section 6 concludes the paper by presenting the main findings and suggesting future directions.

2 Markov Chain Based Approaches

The opinion on how likely a given text X was written by author A, rather than any other author, can be found by a log likelihood ratio:

$$\mathcal{O}_{A,G}(X) = |e_z(X)|^{-1} \log [\, p_A(e_z(X)) \,/\, p_G(e_z(X)) \,]$$

where $z \in \{words, chars\}$, $e_z(X)$ extracts an ordered set of items from X (where the items are either words or characters, indicated by z), $|e_z(X)|^{-1}$ is used as a normalisation for varying number of items, while $p_A(e_z(X))$ and $p_G(e_z(X))$ estimate the likelihood of the text having been written by author A and a *generic* author[1], G, respectively.

Given a threshold t, text X is classified as having been written by author A when $\mathcal{O}_{A,G}(X) > t$, or as written by someone else when $\mathcal{O}_{A,G}(X) \leq t$. The $|e_z(X)|^{-1}$ normalisation term allows for the use of a common threshold (i.e. shared by all authors), which facilitates the interpretation of performance (e.g. via the use of the Equal Error Rate (EER) point on a Receiver Operating Characteristic (ROC) curve (Ortega-Garcia et al., 2004)).

Appropriating a technique originally used in language modelling (Chen and Goodman, 1999), the likelihood of author A having written a particular sequence of items, $X = (i_1, i_2, \cdots, i_{|X|})$, can be approximated using the joint probability of all present m-th order Markov chains:

$$p_A(X) \approx \prod_{j=(m+1)}^{|X|} p_A\left(i_j | i_{j-m}^{j-1}\right) \qquad (1)$$

where i_{j-m}^{j-1} is a shorthand for $i_{j-m} \cdots i_{j-1}$ and m indicates the length of the history. Given training material for author A, denoted as X_A, the maximum likelihood (ML) probability estimate for a particular m-th order Markov chain is:

$$p_A^{ml}\left(i_j | i_{j-m}^{j-1}\right) = \mathcal{C}\left(i_{j-m}^{j} | X_A\right) / \mathcal{C}\left(i_{j-m}^{j-1} | X_A\right) \qquad (2)$$

where $\mathcal{C}\left(i_{j-m}^{j} | X_A\right)$ is the number of times the sequence i_{j-m}^{j} occurs in X_A. For chains that have not been seen during training, elaborate smoothing techniques (Chen and Goodman, 1999) are utilised to avoid zero probabilities in Eqn. (1).

The probabilities for the generic author are estimated from a dataset comprised of texts from many authors.

In this work we utilise interpolated Moffat smoothing[2], where the probability of an m-th or-

[1] A *generic* author is a composite of a number of authors.

[2] Moffat smoothing is often mistakenly referred to as Witten-Bell smoothing. Witten & Bell (1991) referred to this technique as *Method C* and cited Moffat (1988).

der chain is a linear interpolation of its ML estimate and the smoothed probability estimate of the corresponding $(m\text{-}1)$-th order chain:

$$p_A^{mof}\left(i_j|i_{j\text{-}m}^{j\text{-}1}\right) = \alpha_{i_{j\text{-}m}^{j\text{-}1}}\, p_A^{ml}\left(i_j|i_{j\text{-}m}^{j\text{-}1}\right) + \beta_{i_{j\text{-}m}^{j\text{-}1}}\, p_A^{mof}\left(i_j|i_{j\text{-}(m\text{-}1)}^{j\text{-}1}\right)$$

where $\alpha_{i_{j\text{-}m}^{j\text{-}1}} = 1 - \beta_{i_{j\text{-}m}^{j\text{-}1}}$, and

$$\beta_{i_{j-m}^{j-1}} = \frac{\left|i_j : \mathcal{C}\left(i_{j-m}^{j-1} i_j | X_A\right) > 0\right|}{\left|i_j : \mathcal{C}\left(i_{j-m}^{j-1} i_j | X_A\right) > 0\right| + \sum_{i_j} \mathcal{C}\left(i_{j-m}^{j} | X_A\right)}$$

Here, $\left|i_j : \mathcal{C}\left(i_{j-m}^{j-1} i_j | X_A\right) > 0\right|$ is the number of unique $(m{+}1)$-grams that have the same i_{j-m}^{j-1} history items. Further elucidation of this method is given in (Chen and Goodman, 1999; Witten and Bell, 1991).

The $(m\text{-}1)$-th order probability will typically correlate with the m-th order probability and has the advantage of being estimated from a larger number of examples (Chen and Goodman, 1999). The 0-th order probability is interpolated with the uniform distribution, given by: $p_A^{unif} = 1/|V_A|$, where $|V_A|$ is the vocabulary size (Chen and Goodman, 1999).

When an m-th order chain has a history (i.e. the items i_{j-m}^{j-1}) which hasn't been observed during training, a back-off to the corresponding reduced order chain is done[3]:

$$\text{if } \mathcal{C}\left(i_{j-m}^{j-1}|X_A\right) = 0,\ p_A^{mof}\left(i_j|i_{j-m}^{j-1}\right) = p_A^{mof}\left(i_j|i_{j-(m-1)}^{j-1}\right)$$

Note that if the 0-th order chain also hasn't been observed during training, we are effectively backing off to the uniform distribution.

A caveat: the training dataset for an author can be much smaller (and hence have a smaller vocabulary) than the combined training dataset for the generic author, resulting in $p_A^{unif} > p_G^{unif}$. Thus when a previously unseen chain is encountered there is a dangerous bias towards author A, i.e., $p_A^{mof}\left(i_j|i_{j-m}^{j-1}\right) > p_G^{mof}\left(i_j|i_{j-m}^{j-1}\right)$. To avoid this, p_A^{unif} must be set equal to p_G^{unif}.

3 Sequence Kernel Based Approaches

Kernel based techniques, such as SVMs, allow the comparison of, and discrimination between, vectorial as well as non-vectorial objects. In a binary SVM, the opinion on whether object X belongs to class -1 or +1 is given by:

$$\mathcal{O}_{+1,-1}(X) = \sum_{j=1}^{|S|} \lambda_j\, y_j\, k(s_j, X) + b \qquad (3)$$

where $k(X_A, X_B)$ is a symmetric kernel function which reflects the degree of similarity between objects X_A and X_B, while $S = (s_j)_{j=1}^{|S|}$ is a set of support objects with corresponding class labels $(y_j \in \{-1, +1\})_{j=1}^{|S|}$ and weights $\Lambda = (\lambda_j)_{j=1}^{|S|}$. The kernel function, b as well as sets S and Λ define a hyperplane which separates the +1 and -1 classes. Given a training dataset, quadratic programming based optimisation is used to maximise the separation margin[4] (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004).

Recently, kernels for measuring the similarity of texts based on sequences of characters and words have been proposed (Cancedda et al., 2003; Leslie et al., 2004; Vishwanathan and Smola, 2003). One kernel belonging to this family is:

$$k(X_A, X_B) = \sum\nolimits_{q \in Q^\star} w_q\, \mathcal{C}(q|X_A)\, \mathcal{C}(q|X_B) \qquad (4)$$

where $Q^\star$ represents all possible sequences, in X_A and X_B, of the symbols in Q. In turn, Q is a set of possible symbols, which can be characters, e.g. $Q = \{$ 'a', 'b', 'c', $\cdots\}$, or words, e.g. $Q = \{$'kangaroo', 'koala', 'platypus', $\cdots\}$. Furthermore, $\mathcal{C}(q|X)$ is the number of occurrences of sequence q in X, and w_q is the weight for sequence q. If the sequences are restricted to have only one item, Eqn. (4) for the case of words is in effect a bag-of-words kernel (Cancedda et al., 2003; Shawe-Taylor and Cristianini, 2004).

In this work we have utilised weights that were dependent only on the length of each sequence, i.e. $w_q = w_{|q|}$. By default $w_{|q|} = 0$, modified by one of the following functions:

$$\begin{aligned}
\text{specific length: } & w_{|q|} = 1, \text{ if } |q| = \tau \\
\text{bounded range: } & w_{|q|} = 1, \text{ if } |q| \in [1, \tau] \\
\text{bounded linear decay: } & w_{|q|} = 1 + \tfrac{1-|q|}{\tau}, \text{ if } |q| \in [1, \tau] \\
\text{bounded linear growth: } & w_{|q|} = |q| \,/\, \tau, \text{ if } |q| \in [1, \tau]
\end{aligned}$$

where τ indicates a user defined maximum sequence length.

To allow comparison of texts with different lengths, a normalised version (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004) of the kernel can be used:

$$\widehat{k}(X_A, X_B) = k(X_A, X_B) \,/\, \sqrt{k(X_A, X_A)\, k(X_B, X_B)}$$

with constraints $|X_A| \geq 1$ and $|X_B| \geq 1$.

It has been suggested that SVM discrimination based on character sequence kernels in effect utilises a noisy version of stemming (Cancedda et al., 2003). As such, word sequence kernels could be more effective than character sequence

[3] Personal correspondence with the authors of (Chen and Goodman, 1999).

[4] Based on preliminary experiments, the regularisation constant C, used in SVM training, was set to 100.

kernels, since proper word stems, instead of full words, can be explicitly used. However, it must be noted that Eqn. (4) implicitly maps texts to a feature space which has one dimension for each of the possible sequences comprised of the symbols from Q (Cancedda et al., 2003). When using words, the number of unique symbols (i.e. $|Q|$) can be much greater than when using characters (e.g. 10,000 vs 100); furthermore, for a given text the number of words is always smaller than the number of characters. For a given sequence length, these observations indicate that for word sequence kernels the implicit feature space representation can have considerably higher dimensionality and be sparser than for character sequence kernels, which could lead to poorer generalisation of the resulting classifier.

4 Evaluation

4.1 "Columnists" Dataset

We have compiled a dataset that is comprised of texts from 50 newspaper journalists, with a minimum of 10,000 words per journalist. Journalists were selected based on their coverage of several topics; any journalist who covered only one specific area (e.g. sports or economics) was not included in the dataset. Apart from removing all advertising material and standardising the representation by converting any unicode characters to their closest ASCII counterparts, no further editing was performed. The dataset is available for use by other researchers by contacting the authors.

4.2 Setup

The experiments followed a verification setup, where a given text material was classified as either having been written by a hypothesised author or as not written by that author (i.e. a two class discrimination task). This is distinct from a closed set identification setup, where a text is assigned as belonging to one author out of a pool of authors. The presentation of an *impostor text* (a text known not to be written by the hypothesised author) will be referred to as an *impostor claim*, while the presentation of a *true text* (a text known to be written by the hypothesised author) will be referred to as a *true claim*.

Table 1: Approximate correspondence between the number of characters and number of words. For comparison purposes, this paper has about 5900 words.

No. characters	1750	3500	7000	14000	28000
No. words	312	625	1250	2500	5000

For a given text, one of the following two classification errors can occur: **(i)** a false positive, where an impostor text is incorrectly classified as a true text; **(ii)** a false negative, where a true text is incorrectly classified as an impostor text. The errors are measured in terms of the false positive rate (FPR) and the false negative rate (FNR). Following the approach often used within the biometrics field, the decision threshold was then adjusted so that the FPR is equal to the FNR, giving Equal Error Rate (EER) performance (Ortega-Garcia et al., 2004; Sanderson et al., 2006).

The authors in the database were randomly assigned into two disjoint sections: **(i)** 10 background authors; **(ii)** 40 evaluation authors. For the case of Markov chain approaches, texts from the background authors were used to construct the generic author model, while for kernel based approaches they were used to represent the negative class. In both cases, text materials each comprised of approx. 28,000 characters were used, via randomly choosing a sufficient number of sentences from the pooled texts. Table 1 shows a correspondence between the number of characters and words, using the average word length of 5.6 characters including a trailing whitespace (found on the whole dataset).

For each author in the evaluation section, their material was randomly split[5] into two continuous parts: training and testing. The split occurred without breaking sentences. The training material was used to construct the author model, while the test material was used to simulate a true claim as well as impostor claims against all other authors' models. Note that if material from the evaluation section was used for constructing the generic author model, the system would have prior knowledge about the writing style of the authors used for the impostor claims.

For each configuration of an approach (where, for example, the configuration is the order of the Markov chains), the above procedure was repeated ten times, with the randomised assignments and splitting being done each time. The final results

[5] By 'randomly split' we mean that the location of the training and testing parts within the text material is random.

were then obtained in terms of the mean and the corresponding standard deviation of the ten EERs (the standard deviations are shown as error bars in the result figures). Based on preliminary experiments, stemming was used for word based approaches (Manning and Schütze, 1999).

4.3 Experiments and Discussion

In the first experiment we studied the effects of varying the order for character and word Markov chain approaches, while the amount of training material was fixed at approx. 28,000 characters and the test material (for evaluation authors) was decreased from approx. 28,000 to 1,750 characters. Results are presented in Fig. 1.

The results show that 2nd order chains of characters generally obtain the best performance. However, the difference in performance between 1st order and 2nd order chains could be considered as statistically insignificant due to the large overlap of the error bars. The best performing word chain approach had an order of zero, with higher orders (not shown) having virtually the same performance as the 0th order. Its performance is largely similar to the 2nd order character chain approach, with the latter obtaining a somewhat lower error rate at 28,000 characters.

The second experiment was similar to the first, with the difference being that the amount of training material *and* test material was decreased from approx. 28,000 to 1,750 characters. The main change between the results of this experiment (shown in Fig. 2) and the previous experiment's results is the faster degradation in performance as the number of characters is decreased. We comment on this effect later.

In the third experiment we utilised SVMs with character sequence kernels and studied the effects of chunk size. As SVMs employ support objects in the definition of the discriminant function (see Section 3), the training material was split into varying size chunks, ranging from approximately 62 to 4000 characters. Each of the chunks can become a support chunk. Naturally, the smaller the chunk size, the larger the number of chunks. As the split was done without breaking sentences, the effective chunk size tended to be somewhat larger. If there is less words available than a given chunk size, then all of the remaining words are used for forming a chunk. Based on preliminary experiments, the bounded range weight function with

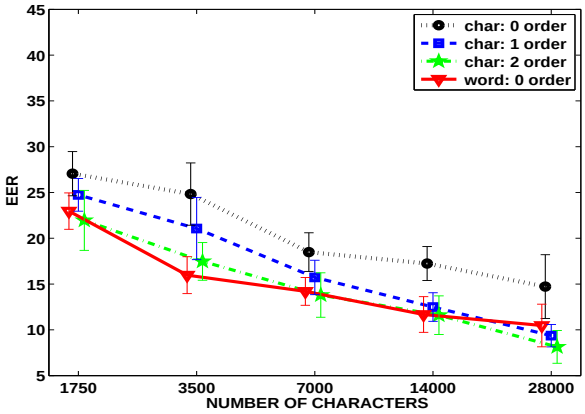

Figure 1: Performance of character and word Markov chain approaches using fixed size training material (approx. 28,000 characters) and varying size test material.

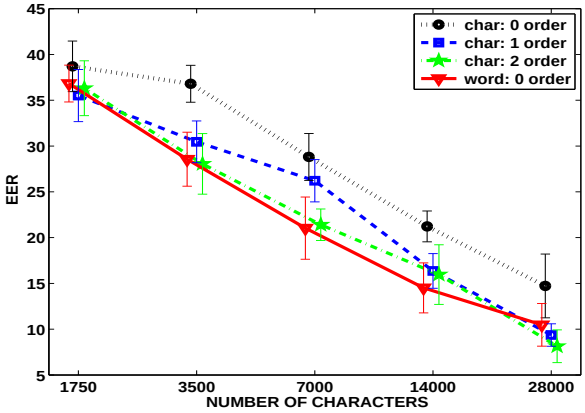

Figure 2: Performance of character and word Markov chain approaches for varying size of training and test material. At each point the size of the training and test materials is equal.

τ=3 was used. The amount of training and test material was equal and three cases were evaluated: 28,000, 14,000 and 7,000 characters. Results, presented in Fig. 3, indicate that the optimum chunk size is approximately 500 characters for the three cases. Furthermore, the optimum chunk size appears to be independent of the number of available chunks for training.

In the fourth experiment we studied the effects of various weight functions and sequence lengths for the character sequence kernel. The amount of training and test material was fixed at approx. 28,000 characters. Based on the results from the previous experiment, chunk size was set at 500. Results for specific length (Fig. 4) suggest that most of the reliable discriminatory information is contained in sequences of length 2. The error rates for the bounded range and bounded lin-

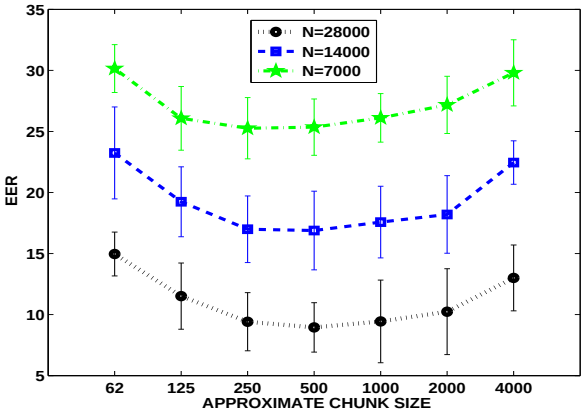

Figure 3: Performance of the character sequence kernel approach for varying chunk sizes. Bounded range weight function with τ=3 was used.

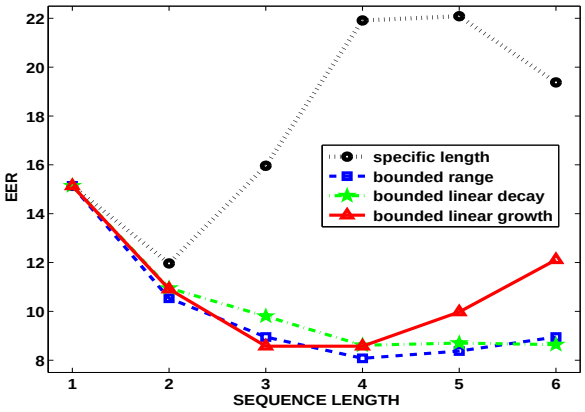

Figure 4: Performance of the character sequence kernel approach for various weight functions. The size of training and test materials was fixed at approx. 28,000 characters. Chunk size of 500 characters was used. Error bars were omitted for clarity.

ear decay functions are quite similar, with both reaching minima for sequences of length 4; most of the improvement occurs when the sequences reach a length of 3. This indicates that while sequences with a specific length of 3 and 4 are less reliable than sequences with a specific length of 2, they contain (partly) complementary information which is useful when combined with information from shorter lengths. Emphasising longer lengths of 5 and 6 (via the bounded linear growth function) achieves a minor, but noticeable, performance degradation. We conjecture that the degradation is caused by the sparsity of relatively long sequences, which affects the generalisation of the classifier.

The fifth experiment was devoted to an evaluation of the effects of chunk size for the word sequence approach. To keep the results comparable with the character sequence approach (third experiment), the training material was split into varying size chunks, ranging from approximately 62 to 8000 characters. Based on the results from the first experiment, the specific length weight function with τ=1 was used[6] (resulting in a bag-of-words kernel).

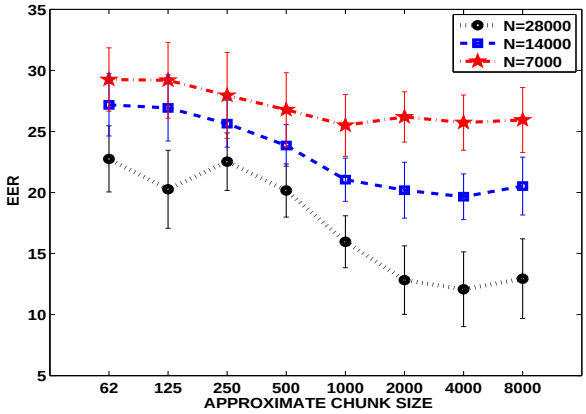

Figure 5: Performance of the word sequence kernel approach for varying chunk sizes. Specific length weight function with τ=1 was used.

The amount of training and test material was equal and three cases were evaluated: 28,000, 14,000 and 7,000 characters. Results, shown in Fig. 5, suggest that the optimum chunk size is approximately 4000 characters for the three cases.

As mentioned in Section 3, for the word based approach the implicit feature space representation can have considerably higher dimensionality and be sparser than for the character based approach. Consequently, longer texts would be required to adequately populate the feature space. This is reflected by the optimum chunk size for the word based approach, which is roughly an order of magnitude larger than the optimum chunk size for the character based approach.

In the sixth experiment we compared the performance of character sequence kernels (using the bounded range function with τ=4) and several configurations of the word sequence kernels. The amount of training material was fixed at approx. 28,000 characters and the test material was decreased from approx. 28,000 to 1,750 characters. Based on the results of previous experiments, chunk size was set to 500 for the character based approach and to 4000 for the word based

[6]Note that for τ=1, all of the weight functions presented in Section 3 are equivalent.

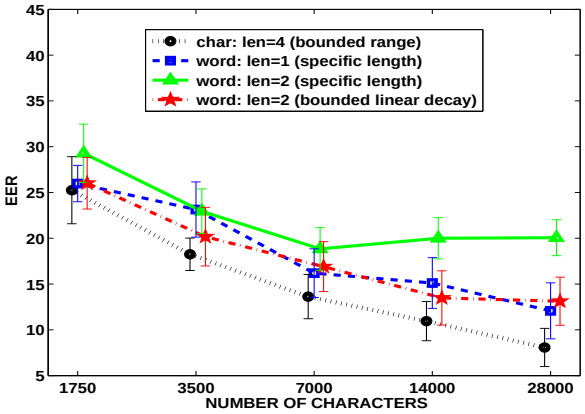

Figure 6: Performance of character and word sequence kernel approaches using fixed size training material (approx. 28,000 characters) and varying size test material.

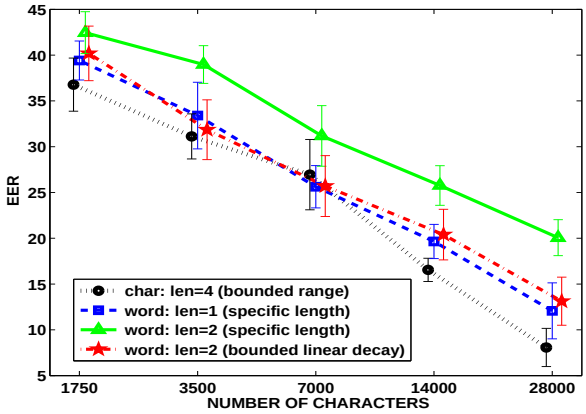

Figure 7: Performance of character and word sequence kernel approaches for varying size of training and test material. At each point the size of the training and test materials is equal.

approach. Fig. 6 shows that word sequences with a specific length of 2 lead to considerably worse performance than sequences of length 1 (i.e. individual words). Furthermore, the best performing combination of lengths (i.e. via the bounded linear decay function[7]) does not provide better performance than using individual words. The character sequence kernels consistently achieve a lower error rate than the best performing word sequence kernel. This suggests that the sparse feature space representation, described in Section 3, is becoming an issue.

The final experiment was similar to the sixth, with the difference being that the amount of training material and test material was decreased from approx. 28,000 to 1,750 characters. As observed for the Markov chain approaches, the main change between the results of this experiment (shown in Fig. 7) and the previous experiment's results is the faster degradation in performance as the number of characters is decreased. Along with the results from experiments 1 and 2, this indicates that the amount of training material has considerably more influence on discrimination performance than the amount of test material.

In Fig. 8 it can be observed that the best performing Markov chain based approach (characters, 2nd order) obtains comparable performance to the character sequence kernel based approach (using the bounded range function with τ=4).

[7]Other combinations of lengths were also evaluated, though the results are not shown here.

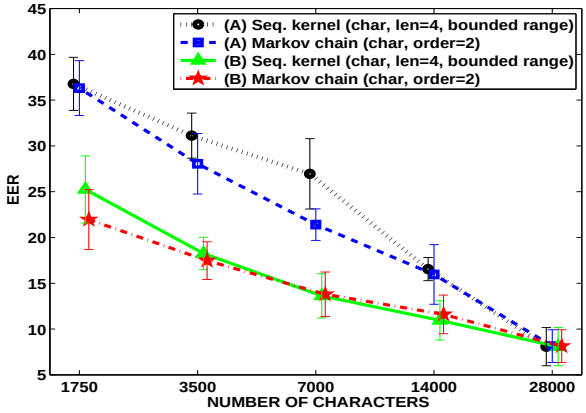

Figure 8: Comparison between the best sequence kernel approach with the best Markov chain approach for two cases: **(A)** varying size of training and test material, **(B)** fixed size training material (approx. 28,000 characters) and varying size test material.

5 Author Unmasking On Short Texts

Koppel & Schler (2004) proposed an alternative method for author verification. Rather than treating the verification problem directly as a two-class discrimination task (as done in Section 4), an "author unmasking" curve is first built. A vector representing the "essential features" of the curve is then classified in a traditional SVM setting. The unmasking procedure is reminiscent of the recursive feature elimination procedure first proposed in the context of gene selection for cancer classification (Guyon et al., 2002).

Instead of having an author specific model (as in the Markov chain approach) or an author specific SVM, a reference text is used. The text to be clas-

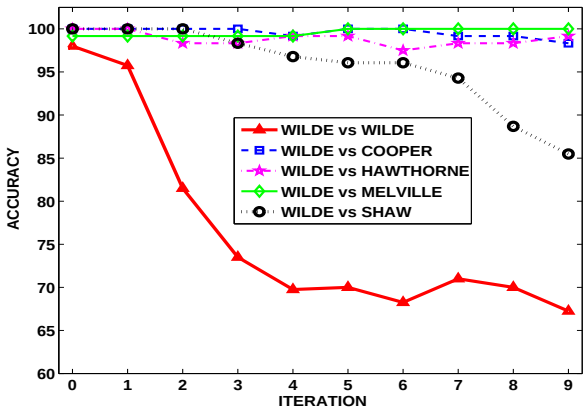

Figure 9: Unmasking of Wilde's *An Ideal Husband* using Wilde's *Woman of No Importance* as well as the works of other authors as reference texts.

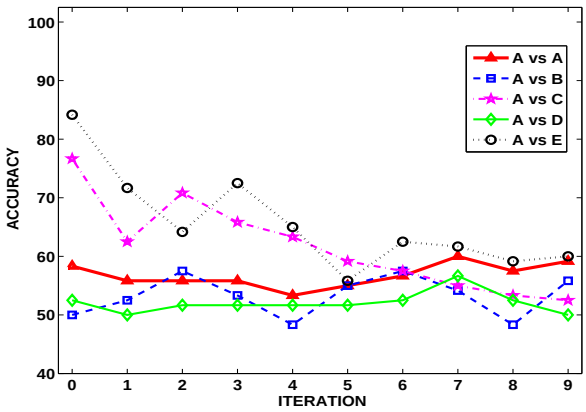

Figure 10: Unmasking of a text from author A from the Columnists dataset, using A's as well as other authors' reference texts.

sified as well as the reference text are divided into chunks; the features representing each chunk are the counts of pre-selected words. The chunks are partitioned into training and test sets. Each point in the author unmasking curve is the accuracy of discriminating (using a linear SVM) between the test chunks from the two texts. At the end of each iteration several of the most discriminative words are removed from further consideration.

The underlying hypothesis is that if the two given texts have been written by the same author, the differences between them will be reflected in a relatively small number of features. Koppel & Schler (2004) observed that for texts authored by the same person, the extent of the accuracy degradation is much larger than for texts written by different authors. Encouraging classification results were obtained for long texts (books available from Project Gutenberg[8]).

In this section we first confirm the unmasking effect for long texts and then show that for shorter texts (i.e. approx. 5000 words), the effect is considerably less distinctive.

For the first experiment we followed the setup in (Koppel and Schler, 2004), i.e. the same books, chunks with a size of approximately 500 words, 10 iterations of removing 6 features, and using 250 words with the highest average frequency in both texts as the set of pre-selected words. For each pair of texts, 10 fold cross-validation was used to obtain 10 curves, which were then represented by a mean curve prior to further processing. Fig. 9 shows curves for unmasking Oscar Wilde's

[8]http://www.gutenberg.org

Table 2: Performance of author unmasking, character sequence kernel approach ($\tau = 4$, bounded range) and character Markov chain approach (2nd order).

Approach	**mean EER**	**std. dev.**
Author unmasking	30.88	4.32
Character sequence kernel	8.08	2.08
Character Markov chain	8.14	1.79

An Ideal Husband using Wilde's *Woman of No Importance* (same-author curve) as well as the works of other authors as reference texts (different-author curves). As can be observed, the unmasking effect is most pronounced for Wilde's text. Furthermore, this figure has a close resemblance to Fig. 2 in (Koppel and Schler, 2004).

In the second experiment we used text materials from the Columnists dataset. Each author's text material was divided into two sections of approximately 5000 words, with the one of the sections randomly selected to be the reference material, leaving the other as the test material. Based on preliminary experiments, the number of pre-selected words was set to 100 (with the highest average frequency in both texts) and the size of the chunks was set to 200 words. The remainder of the unmasking procedure setup was the same as for the first experiment. The setup for verification trials was similar to the setup in Section 4.2, with the difference being that the background authors were used to generate same-author and different-author curves for training the secondary SVM. In all cases features from each curve were extracted, as done in (Koppel and Schler, 2004), prior to further processing.

Table 2 provides a comparison between the performance of the unmasking approach with that of the character sequence kernel and character Markov chain based approaches, as evaluated in Section 4. Fig. 10 shows representative curves resulting from unmasking of the test material from author A, using A's as well as other authors' reference materials. Generally, the unmasking effect for the same-author curves is considerably less pronounced and in some cases it is non-existent. More dangerously, different-author curves often have close similarities to same-author curves. The results and the above observations hence suggest that the unmasking method is less useful when dealing with relatively short texts.

6 Main Findings and Future Directions

In this paper we investigated the use of character and word sequence kernels for the task of authorship attribution and compared their performance with two probabilistic approaches based on Markov chains of characters and words. The evaluations were done on a relatively large dataset (50 authors), where each author covered several topics. Rather than using the restrictive closed set identification setup, a verification setup was used which takes into account the realistic case of texts which are not written by any hypothesised authors. We also appraised the applicability of the recently proposed author unmasking approach for dealing with relatively short texts.

In the framework of Support Vector Machines, several configurations of the sequence kernels were studied, showing that word sequence kernels do not achieve better performance than a bag-of-words kernel. Character sequence kernels (using sequences with a length of 4) generally have better performance than the bag-of-words kernel and also have comparable performance to the two probabilistic approaches.

A possible advantage of character sequence kernels over word-based kernels is their inherent ability to do partial matching of words. Let us consider two examples. **(i)** Given the words "negotiation" and "negotiate", the character sequence kernel can match "negotiat", while a standard word-based kernel requires explicit word stemming beforehand in order to match the two related words (as done in our experiments). **(ii)** Given the words "negotiation" and "desalination", a character sequence kernel can match the common ending "ation". Particular word endings may be indicative of a particular author's style; such information would not be picked up by a standard word-based kernel.

Interestingly, the bag-of-words kernel based approach obtains worse performance than the corresponding word based Markov chain approach. Apart from the issue of sparse feature space representation, factors such as the chunk size and the setting of the C parameter in SVM training can also affect the generalisation performance.

The results also show that the amount of training material has more influence on discrimination performance than the amount of test material; about 5000 training words are required to obtain relatively good performance when using between 1250 and 5000 test words.

Further experiments suggest that the author unmasking approach is less useful when dealing with relatively short texts, due to the unmasking effect being considerably less pronounced than for long texts and also due to different-author unmasking curves having close similarities to the same-author curves.

In future work it would be useful to appraise composite kernels (Joachims et al., 2001) in order to combine character and word sequence kernels. If the two kernel types use (partly) complementary information, better performance could be achieved. Furthermore, more sophisticated character sequence kernels can be evaluated, such as mismatch string kernels used in bioinformatics, where mutations in the sequences are allowed (Leslie et al., 2004).

Acknowledgements

The authors thank the anonymous reviewers as well as Simon Burton, Ari Chanen, Arvin Dehghani, Etienne Grossmann, Adam Kowalczyk and Silvia Richter for useful suggestions and discussions.

National ICT Australia (NICTA) is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

References

N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders. 2003. Word-sequence kernels. *J. Machine Learning Research*, 3:1059–1082.

S.F. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13:359–394.

R. Clement and D. Sharp. 2003. Ngram and Bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing*, 18(4):423–447.

M. W. Corney. 2003. Analysing e-mail text authorship for forensic purposes. Master's thesis, Queensland University of Technology, Australia.

J. Diederich, J. Kindermann, E. Leopold, and G. Paass. 2003. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2):109–123.

D. W. Foster. 2001. *Author Unknown: On the Trail of Anonymous*. Henry Holt & Company, 2nd ed.

M. Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proc. 20th Int. Conf. Computational Linguistics (COLING)*, pages 611–617, Geneva.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422.

P. Hanus and J. Hagenauer. 2005. Information theory helps historians. *IEEE Information Theory Society Newsletter*, 55(September):8.

T.-M. Huang and V. Kecman. 2005. Gene extraction for cancer diagnosis by support vector machines – an improvement and comparison with nearest shrunken centroid method. In *Proc. 15th Int. Conf. Artificial Neural Networks (ICANN)*, pages 617–624, Warsaw.

T. Joachims, N. Cristianini, and J. Shawe-Taylor. 2001. Composite kernels for hypertext categorisation. In *Proc. 18th Int. Conf. Machine Learning (ICML)*, pages 250–257, Massachusetts.

M. Koppel and J. Schler. 2004. Authorship verification as a one-class classification problem. In *Proc. 21st Int. Conf. Machine Learning (ICML)*, Banff, Canada.

C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. Noble. 2004. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476.

C.D Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

A. Moffat. 1988. A note on the PPM data compression algorithm, Res. Report. 88/7, Dept. Comput. Sci., University of Melbourne, Australia.

F. Mosteller. 1984. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer, 2nd edition.

J. Ortega-Garcia, J. Bigun, D. Reynolds, and J. Gonzalez-Rodriguez. 2004. Authentication gets personal with biometrics. *IEEE Signal Processing Magazine*, 21(2):50–62.

F. Peng, D. Schuurmans, and S. Wang. 2004. Augmenting naive Bayes classifiers with statistical language models. *Information Retrieval*, 7:317–345.

C. Sanderson, S. Bengio, and Y. Gao. 2006. On transforming statistical models for non-frontal face verification. *Pattern Recognition*, 39(2):288–302.

B. Schölkopf and A. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, USA.

J. Shawe-Taylor and N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, UK.

S.V.N. Vishwanathan and A. Smola. 2003. Fast kernels for string and tree matching. In *Advances in Neural Information Processing Systems (NIPS) 15*, pages 569–576, Cambridge. MIT Press.

I.H. Witten and T.C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Information Theory*, 37(4):1085–1094.