Using Frequent Fixed or Variable-Length POS Ngrams or Skip-Grams for Blog Authorship Attribution

Yao Pokou, Philippe Fournier-Viger, Chadia Moghrabi

Using Frequent Fixed or Variable-Length POS Ngrams or Skip-grams for Blog Authorship Attribution

Yao Jean Marc Pokou[1], Philippe Fournier-Viger[1,2], and Chadia Moghrabi[1]

[1] Dept. of Computer Science, Université de Moncton, Moncton, NB, Canada
[2] School of Natural Sciences and Humanities, Harbin Institute of technology Shenzhen Grad. School, Shenzhen, Guangdong, China
{eyp3705, philfv, chadia.moghrabi}@umoncton.ca

Abstract. Authorship attribution is the process of identifying the author of an unknown text from a finite set of known candidates. In recent years, it has become increasingly relevant in social networks, blogs, emails and forums where anonymous posts, bullying, and even threats are sometimes perpetrated. State-of-the-art systems for authorship attribution often combine a wide range of features to achieve high accuracy. Although many features have been proposed, it remains an important challenge to find new features and methods that can characterize each author and that can be used on non formal or short writings like blog content or emails. In this paper, we present a novel method for authorship attribution using frequent fixed or variable-length part-of-speech patterns (ngrams or skip-grams) as features to represent each author's style. This method allows the system to automatically choose its most appropriate features as those sequences being used most frequently. An experimental evaluation on a collection of blog posts shows that the proposed approach is effective at discriminating between blog authors.

Keywords: authorship attribution, part-of-speech patterns, top-k POS sequences, frequent POS patterns, skip-grams, ngrams, blogs

1 Introduction

Authorship analysis is the process of examining the characteristics of a piece of work to identify its authors [1]. The forerunners of authorship attribution (AA) are Medenhall who studied the plays of Shakespeare in 1887 [2], and Mosteller and Wallace who studied the disputed authorship of the Federalist Papers in 1964 [3]. Beside identifying the authors of works published anonymously, popular applications of AA techniques include authenticating documents, detecting plagiarism, and assisting forensic investigations.

In recent years, an emerging application of AA has been to analyze online texts, due to the increasing popularity of Internet-based communications, and the need to find solutions to problems such as online bullying, and anonymous

threats. However, this application raises novel challenges, as online texts are written in an informal style, and are often short, thus providing less information about their authors. The accuracy achieved with well-established features for AA like function words, syntactic and lexical markers suffers when applied to short-form messages [4], as many of the earliest studies have been tested on long formal text such as books. It is thus a challenge to find new features and methods that are applicable to informal or short texts.

Studies on AA focused on various types of features that are either syntactic or semantic [5]. Patterns based on syntactic features of text are considered reliable because they are unconsciously used by authors. Syntactic features include frequencies of ngrams, character ngrams, and function words [1]. Baayen, van Halteren, and Tweedie rewrote the frequencies rules for AA based on two syntactically annotated samples taken from the Nijmegen corpus [6]. Semantic features take advantage of words' meanings and their likeness, as exemplified by Clark and Hannon whose classifier quantifies how often an author uses a specific word instead of its synonyms [7]. Moreover, state-of-the-art systems for AA often combine a wide range of features to achieve higher accuracy. For example, the JStylo system offers more than 50 configurable features [8].

Syntactic markers, however, have been less studied because of their language-dependent aspect. This paper studies complex linguistic information carried by part-of-speech (POS) *patterns* as a novel approach for authorship attribution of informal texts. The hypothesis is that POS patterns more frequently appearing in texts, could accurately characterize authors' styles.

The contributions of this paper are threefold. First, we define a novel feature for identifying authors based on fixed or variable length POS patterns (ngrams or skip-grams). A signature is defined for each author as the intersection of the top k most frequent POS patterns found in his or her texts, that are less frequent in texts by other authors. In this method, the system automatically chooses its most appropriate features rather than have them predefined in advance. Second, a process is proposed to use these signatures to perform AA. Third, an experimental evaluation using blog posts of 10 authors randomly chosen from the *Blog Authorship Corpus* [9] shows that the proposed approach is effective at inferring authors of informal texts. Moreover, the experimental study provides answers to the questions of how many patterns are needed, whether POS skip-grams or ngrams perform better, and whether fixed length or variable-length patterns should be used.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the proposed approach. Section 4 presents the experimental evaluation. Finally, section 5 draws the conclusions.

2 Related Work

The use of authorship attribution techniques goes back to the 19th century with the studies of Shakespeare's work. This section concentrates on those using POS tagging or informal short texts.

Stamatatos et al. exclusively used natural language processing (NLP) tools for their three-level text analysis. Along with the token and analysis levels, the phrase level concentrated on the frequencies of single POS tags. They analyzed a corpus of 300 texts by 10 authors, written in Greek. Their method achieved around 80% accuracy by excluding some less significant markers [10].

Gamon combined syntactic and lexical features to accurately identify authors [11]. He used features such as the frequencies of ngrams, function words, and POS trigrams from a set of eight POS tags obtained using the NLPWin system. Over 95% accuracy was obtained. However, its limitation was its evaluation using only three texts, written by three authors [11].

More recently, Sidorov et al. introduced syntactic ngrams (sngrams) as a feature for AA. Syntactic ngrams are obtained by considering the order of elements in syntactic trees generated from a text, rather than by finding n contiguous elements appearing in a text. They compared the use of sngrams with ngrams of words, POS, and characters. The 39 documents by three authors from Project Gutenberg were classified using SVM, J48 and Naïve Bayes implementations provided by the WEKA library. The best results were obtained by SVM with sngrams [12]. The limitations of this work are an evaluation with only three authors, a predetermined length of ngrams, and that all 11,000 ngrams/sngrams were used.

Variations of ngrams were also considered. For example, skip-grams were used by García-Hernández et al. [13]. Skip-grams are ngrams where some words are ignored in sentences with respect to a threshold named the skip step. Ngrams are the specific case of skip-grams where the skip step is equal to 0. The number of skip-grams is very large and their discovery needs complex algorithms [12]. To reduce their number, a cut-off frequency threshold is used [13].

POS ngrams have also been used for problems related to AA. Litvinova et al. used the frequencies of 227 POS bigrams for predicting personality [14].

Unlike previous work using POS patterns (ngrams or skip-grams), the approach presented here considers not only fixed length POS patterns but also variable length POS patterns. Another distinctive characteristic is that it finds only the k most frequent POS patterns in each text (where k is set by the user), rather than using a large set of patterns or using a predetermined cut-off frequency threshold. This allows the proposed approach to use a very small number of patterns to create a signature for each author, unlike many previous works that compute the frequencies of hundreds or thousands of ngrams or skip-grams.

3 The Proposed Approach

The proposed approach takes as input a training corpus C_m of texts written by m authors. Let $A = \{a_1, a_2, ...a_m\}$ denote the set of authors. Each author a_i $(1 \leq i \leq m)$ has a set of z texts $T_i = \{t_1, t_2, \ldots t_z\}$ in the corpus. The proposed approach is applied in three steps.

Preprocessing. The first step is to prepare blogs from the corpus so that they can be used for generating author signatures. All information that does

not carry an author's style is removed, such as images, tables, videos, and links. In addition, each text is stripped of punctuation and is split into sentences using the Rita NLP library (`http://www.rednoise.org/rita/`). Then, every text is tagged using the Standford NLP Tagger (`http://nlp.stanford.edu/software/`) as it produced 97.24 % accuracy on the Penn Treebank Wall Street Journal corpus [15]. Since the main focus is analyzing how sentences are *constructed* by authors rather than the choice of words, words in texts are discarded and only the *patterns* about parts of speech are maintained. Thus, each text becomes a set of sequences of POS tags. For example, Table 1 shows the transformation of six sentences from an author in the corpus.

Table 1: An example of blog text transformation.

#	Original Sentence	Transformed Sentence into POS Sequences
1	Live From Ohio Its Youth Night.	JJ IN NNP PRP$ NNP NNP.
2	So the Youth Rally was tonight.	IN **DT** NNP NNP VBD **NN**.
3	I think it was a success.	PRP VBP PRP VBD **DT NN**.
4	Maybe not.	RB RB.
5	The activity looked complicated and confusing but luckily Tim and I didn't have to do it instead we got to (...) right.	**DT NN** VBD JJ CC JJ CC RB NNP CC PRP VBP VBP TO VB PRP RB PRP VBD TO (...) JJ.
6	And finally the skit.	CC RB **DT NN**.

Signature extraction. The second step of the proposed approach is to extract a signature for each author, defined as a set of part-of-speech patterns (POSP) annotated with their respective frequency. Signature extraction has four parameters: the number of POS patterns (ngrams or skip-grams) to be found k, the minimum pattern length n, the maximum length x, and the maximum gap $maxgap$ allowed between POS tags. Frequent patterns of POS tags are extracted from each text. The hypothesis is that each text may contain patterns of POS tags unconsciously left by its author, representing his/her writing style, and could be used to identify that author accurately. For each text t, the k most frequent POS patterns are extracted using a general-purpose sequential pattern mining algorithm [16]. Let POS denote the set of POS tags. Consider a sentence $w_1, w_2, \ldots w_y$ consisting of y part-of-speech tags, and a parameter $maxgap$ (a positive integer). A n-skip-gram is an ordered list of tags $w_{i_1}, w_{i_2}, \ldots w_{i_n}$ where $i_1, i_2, \ldots i_n$ are integers such that $0 < i_j - i_{j-1} \leq maxgap + 1 (1 < j \leq n)$. A n-skip-gram respecting the constraint of $maxgap = 0$ (i.e. no gaps are allowed) is said to be a ngram. For a given text, the *frequency* of a sequence seq is the number of sequences (sentences) from the text containing seq. Similarly, the *relative frequency* of a sequence is its frequency divided by the number of sequences in the text. For example, the frequency of the 5-skip-gram $\langle NNP, VBP, VBP, PRP, RB \rangle$ is 1 in the transformed sentences of Table 1,

while the frequency of the 2-skip-gram $\langle DT, NN \rangle$ is 4 (this pattern appears in the second, third, fifth, and sixth transformed sentences).

In the following, the term *part-of-speech pattern* of a text t, abbreviated as $(POSPt)^k_{n,x}$, or *patterns*, is used to refer to the k most frequent POS patterns extracted from a text, annotated with their relative frequency, respecting the *maxgap* constraint, and having a length no less than n and not greater than x.

Then, the *POS patterns of an author* a_i in all his/her texts are found. They are denoted as $(POSPa_i)^k_{n,x}$ and defined formally as the union of the POS patterns found in all of his/her texts: $(POSPa_i)^k_{n,x} = \bigcup_{t \in T_i} (POSPt)^k_{n,x}$.

Table 2: Part-of-speech patterns and their relative frequencies.

Pattern	Relative Frequency (%)	Part-of-speech Description
NN	66.6	Noun, singular or mass
DT-NN	66.6	Determiner - Noun
NNP	50.0	Proper noun, singular
VBP-VBP-PRP	16.6	Verb, non-3rd person singular-verb, non-3rd person singular - Personal pronoun
NNP-VBP-VBP-PRP-RB	16.6	Proper noun, singular - Verb, non-3rd person singular-verb - Verb, non-3rd person singular-verb -Personal pronoun- Adverb

Then, the signature of each author a_i is extracted. The signature s_{a_i} of author a_i is the intersection[3] of the POS patterns of his/her texts T_i. The signature is formally defined as: $(s_{a_i})^k_{n,x} = \bigcap_{t \in T_i} (POSPt)^k_{n,x}$.

For instance, the part-of-speech patterns $(POSP_{1029959})^4_{1,5}$ of the blogger 1029959 from our corpus for $maxgap = 3$ are shown in Table 2. In this table, POS patterns are ordered by decreasing frequency. It can be seen that the patterns Noun (NN), and Determiner - Noun (DT-NN) appear in four of the six sentences shown in Table 1. Note that the relative frequency of each pattern is calculated as the relative frequency over all texts containing the pattern.

Moreover, the POS patterns of an author a_i may contain patterns having unusual frequencies that truly characterize the author's style, but also patterns representing common sentence structures of the English language. To tell apart these two cases, a set of reference patterns and their frequencies is extracted to be used with each signature for authorship attribution. Extracting this set of reference patterns is done with respect to each author a_i by computing the union of all parts of speech of the other authors[4]. This set is formally defined as follows. The *Common POS patterns of all authors excluding an author* a_i is the union of all the other $POSPa$, that is: $(CPOSa_i)^k_{n,x} = \bigcup_{a \in A \wedge a \neq a_i} (POSPa)^k_{n,x}$.

[3] A less strict intersection could also be used, requiring occurrences in some or the majority of texts rather than all of them.

[4] A subset of all other authors can also be used if the set of other authors is large.

For example, the common POS patterns computed using authors 206953 and 2369365, and excluding author 1029959, are the patterns DT, IN and PRP. The relative frequencies (%) of these patterns for author 206953 are 67.9%, 69.6% and 79.1%, while the relatives frequencies of these patterns for author 2369365 are 64.2%, 63.0% and 58.7%. Note that the relative frequency of each pattern in CPOS is calculated as the relative frequency over all texts containing the pattern.

The revised signature of an author a_i after removing the common POS patterns of all authors excluding a_i is defined as: $(s'_{ai})^k_{n,x} = (s_{ai})^k_{n,x} \setminus (CPOSa_i)^k_{n,x}$. When the revised signature of each author $a_1, a_2, ...a_m$ has been extracted, the collection of revised signatures $s'^{,k}_{n,x} = \{(s'_{a1})^k_{n,x}, (s'_{a2})^k_{n,x}, \ldots, (s'_{am})^k_{n,x}\}$ is saved.

Authorship attribution. The third step of the proposed approach is to use the generated signatures to perform authorship attribution, that is to identify the author a_u of an anonymous text t_u that was not used for training. The algorithm takes as input an anonymous text t_u, the sets of signatures $s'^{,k}_{n,x}$ and the parameters n, x and k. The algorithm first extracts the part-of-speech patterns in the unknown text t_u with their relative frequencies. Then, it compares the patterns found in t_u and their frequencies with the patterns in the signature of each author using a similarity function. Each author and the corresponding similarity are stored as a tuple in a list. Finally, the algorithm returns this list sorted by decreasing order of similarity. This list represents a ranking of the most likely authors of the anonymous text t_u. Various metrics may be used to define similarity functions. In this work, the Pearson correlation was chosen as it provided better results in initial experiments.

4 Experimental Evaluation

Experiments were conducted to assess the effectiveness of the proposed approach for authorship attribution using either fixed or variable-length ngrams or skip-grams of POS patterns. Our Corpus consists of 609 posts from 10 bloggers, obtained from the *Blog Authorship Corpus* [9]. Blog posts are written in English and were originally collected from the *Blogger* website (`https://www.blogger.com/`). The ten authors were chosen randomly. Note that non-verbal expressions (emoticons, smileys or interjections used in web-blogs or chatrooms such as *lol, hihi, and hahaaa* were not removed because consistent part-of-speech tags were returned by the tagger for the different blogs. The resulting corpus has a total of $265,263$ words and $19,938$ sentences. Details are presented in Table 3.

Then, each text was preprocessed (as explained in section 3). Eighty percent (80%) of each text was used to extract each author's signature (training), and the remaining 20% was used to perform the unknown authorship attribution by comparing each text t_u with each author signature (testing). This produced a ranking of the most likely author to the least likely author, for each text.

Table 3: Corpus statistics.

Author Id	Post Count	Word Count	Sentence Count
1029959	83	29,799	2,314
2069530	125	40,254	2,409
2369365	119	24,148	1,692
3182387	87	30,375	1,752
3298664	34	26,052	2,417
3420481	63	19,063	1,900
3454871	37	16,322	1,722
3520038	45	21,312	1,698
3535101	1	24,401	1,865
3701154	15	33,537	2,169
Total	609	265,263	19,938

4.1 Influence of parameters n, x, k, and $maxgap$ on overall results

Recall that our proposed approach takes four parameters as input, i.e. the minimum and maximum length of part-of-speech patterns n and x, the maximum gap $maxgap$, and k the number of patterns to be extracted in each text. The influence of these parameters on authorship attribution success was first evaluated. For our experiment, parameter k was set to 50, 100, and 250. For each value of k, the length of the part-of-speech patterns was varied from $n = 1$ to $x = 4$. Moreover, the $maxgap$ parameter was set to 0 (ngrams), and from 1 to 3 (skip-grams). For each combination of parameters, we measured the *success rate*, defined as the number of correct predictions divided by the number of predictions. Tables 4, 5, 6, and 7 respectively show the results obtained for $maxgap = 0, 1, 2, 3$, for various values of k, n and x. Furthermore, in these tables, results are also presented by ranks. The row R_z indicates the number of texts where the author was predicted as one of the z most likely authors, divided by the total number of texts (success rate). For example, R_3 indicates the percentage of texts where the author is among the three most likely authors as predicted by the proposed approach. Since there are 10 authors in the corpus, results are shown for R_z varied from 1 to 10.

The first observation is that the best overall results are obtained by setting $n = 2, x = 2$, $k = 250$ and $maxgap = 0$. For these parameters, the author of an anonymous text is correctly identified 73.3% of the time, and 86.6% as one of the two most likely authors (R_2).

The second observation is that excellent results can be achieved using few patterns to build signatures ($k = 250$). Note that our approach was also tested with other values of k such as 50, but it did not provide better results than $k = 250$ (results are not shown due to space limitation). This is interesting as it means that signatures can be extracted using a very small number of the k most frequent POS patterns (as low as $k = 100$) and still characterize well the writing style of authors. This is in contrast with previous works that generally

used a large number of patterns to define an author's signature. For example, Argamon et al. have computed the frequencies of 685 trigrams [17] and Sidorov et al. computed the frequencies of 400/11,000 ngrams/sngrams [12]. By Occam's Razor, it can be argued that models with less patterns (simpler) may be less prone to overfitting.

Third, it can be observed that good results can also be obtained using POS skip-grams. This is interesting since skip-grams of words or POS have received considerably less attention than ngrams in previous studies. Moreover, to our knowledge no studies had previously compared the results obtained with ngrams and skip-grams for authorship attribution of informal and short texts. Fourth, it is found that using skip-grams of fixed length (bigrams) is better than using patterns of variable length. This provides an answer to the important question of whether fixed length POS patterns or variable-length POS patterns should be used. This question was not studied in previous works.

Table 4: Overall classification results using ngrams (maxgap = 0)

(a) Top-k, for k=100.

Success rate in %					
n, x	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
R_1	56.7	53.3	56.7	**63.3**	40.0
R_2	73.4	60.0	63.4	**76.6**	70.0
R_3	80.1	73.3	73.4	76.6	73.3
R_4	83.4	83.3	83.4	79.9	80.0
R_5	86.7	90.0	90.1	79.9	86.7
R_6	90.0	90.0	90.1	86.6	90.0
R_7	90.0	90.0	90.1	86.6	90.0
R_8	96.7	96.7	96.8	89.9	96.7
R_9	96.7	96.7	96.8	96.6	100.0
R_{10}	100.0	100.0	100.0	100.0	100.0

(b) Top-k, for k=250.

Success rate in %					
n, x	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
R_1	60.0	60.0	56.7	**73.3**	46.7
R_2	73.3	76.7	73.4	**86.6**	70.0
R_3	83.3	86.7	86.7	86.6	83.3
R_4	86.6	86.7	90.0	86.6	86.6
R_5	86.6	93.4	90.0	89.9	89.9
R_6	93.3	93.4	93.3	89.9	93.2
R_7	93.3	93.4	96.6	89.9	96.5
R_8	96.6	96.7	96.6	89.9	99.8
R_9	96.6	96.7	96.6	96.6	99.8
R_{10}	100.0	100.0	100.0	100.0	100.0

4.2 Influence of parameters n, x and k on authorship attribution for each author

The previous subsection studied the influence of parameters n, x and k on the ranking of authors for all anonymous texts. This subsection analyzes the results for each author separately. Tables 8a and 8b respectively show the success rates attributed to each author (rank R_1) for $maxgap = 0$, $k = 100$ and $k = 250$, when the other parameters are varied.

It can be observed that for most authors, at least 66.7% of texts are correctly attributed. For example, for $n = 2$, $x = 2$ and $k = 250$, four authors have all texts correctly identified, four have 66.7% of their texts correctly classified, and

Table 5: Overall classification results using skip-grams with maxgap = 1

(a) Top-k, for k=100.

Success rate in %					
n, x	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
R_1	56.7	50.0	50.0	**66.7**	50.0
R_2	70.0	60.0	60.0	**80.0**	73.3
R_3	80.0	83.3	80.0	83.3	76.6
R_4	86.7	83.3	80.0	86.6	86.6
R_5	86.7	86.6	83.3	86.6	86.6
R_6	90.0	89.9	90.0	89.9	93.3
R_7	90.0	89.9	90.0	89.9	93.3
R_8	93.3	93.2	93.3	93.2	96.6
R_9	93.3	93.2	93.3	96.5	96.6
R_{10}	100.0	100.0	100.0	100.0	100.0

(b) Top-k, for k=250

Success rate in %					
n, x	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
R_1	63.3	56.7	60.0	**70.0**	56.7
R_2	80.0	76.7	76.7	**83.3**	70.0
R_3	86.7	83.4	83.4	83.3	76.7
R_4	90.0	83.4	90.1	83.3	86.7
R_5	90.0	90.1	93.4	83.3	90.0
R_6	90.0	90.1	93.4	86.6	93.3
R_7	93.3	93.4	93.4	89.9	96.6
R_8	96.6	96.7	96.7	93.2	96.6
R_9	96.6	96.7	96.7	96.5	99.9
R_{10}	100.0	100.0	100.0	100.0	100.0

Table 6: Overall classification results using skip-grams with maxgap = 2

(a) Top-k, for k=100.

Success rate in %					
n, x	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
R_1	53.3	46.7	50.0	**63.3**	56.7
R_2	73.3	70.0	63.3	**76.6**	70.0
R_3	80.0	73.3	73.3	79.9	73.3
R_4	83.3	80.0	80.0	79.9	80.0
R_5	86.6	83.3	80.0	83.2	86.7
R_6	86.6	86.6	90.0	83.2	93.4
R_7	89.9	89.9	90.0	83.2	93.4
R_8	93.2	89.9	90.0	89.9	93.4
R_9	93.2	93.2	93.3	96.6	96.7
R_{10}	100.0	100.0	100.0	100.0	100.0

(b) Top-k, for k=250.

Success rate in %					
n, x	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
R_1	60.0	56.7	56.7	63.3	**66.7**
R_2	70.0	73.4	73.4	83.3	**76.7**
R_3	83.3	80.1	76.7	83.3	80.0
R_4	83.3	83.4	83.4	86.6	80.0
R_5	86.6	86.7	86.7	86.6	90.0
R_6	89.9	90.0	90.0	89.9	93.3
R_7	93.2	93.3	93.3	89.9	93.3
R_8	96.5	96.6	96.6	89.9	100.0
R_9	96.5	96.6	96.6	96.6	100.0
R_{10}	100.0	100.0	100.0	100.0	100.0

two have 33.3% of texts correctly classified. Overall, it can be thus found that the proposed approach performs very well.

It can also be found that some authors were harder to classify (author 3420481 and 3454871). After investigation, we found that the reason for the incorrect classification is that both authors have posted a same very long blog post, which represents about 50% of the length of their respective corpus. This content has a distinct writing style, which suggests that it was not written by any of these authors. Thus, it had a great influence on their signatures, and led to the poor classification of these authors.

In contrast, some authors were very easily identified with high success rate and high accuracy (cf. Table 9 for accuracies). For example, all texts by authors

Table 7: Overall classification results using skip-grams with maxgap = 3

(a) Top-k, for k=100

Success rate in %					
n, x	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
R_1	46.7	46.7	46.7	**66.7**	50.0
R_2	73.4	70.0	66.7	**73.4**	66.7
R_3	73.4	76.7	76.7	80.1	70.0
R_4	83.4	80.0	76.7	80.1	80.0
R_5	83.4	83.3	80.0	86.8	80.0
R_6	83.4	83.3	86.7	86.8	83.3
R_7	90.1	86.6	86.7	86.8	83.3
R_8	90.1	89.9	90.0	90.1	90.0
R_9	93.4	93.2	93.3	96.8	96.7
R_{10}	100.0	100.0	100.0	100.0	100.0

(b) Top-k, for k=250.

Success rate in %					
n, x	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
R_1	63.3	60.0	56.7	**66.7**	66.7
R_2	76.6	73.3	70.0	**83.4**	73.4
R_3	83.3	80.0	73.3	86.7	73.4
R_4	86.6	83.3	76.6	86.7	76.7
R_5	86.6	83.3	83.3	86.7	80.0
R_6	86.6	86.6	86.6	90.0	83.3
R_7	89.9	93.3	89.9	90.0	93.3
R_8	96.6	93.3	93.2	96.7	100.0
R_9	96.6	93.3	93.2	96.7	100.0
R_{10}	100.0	100.0	100.0	100.0	100.0

Table 8: Success rate per author using skip-grams with maxgap = 0 (ngrams)

(a) Top-k, for k=100.

Success rate per author in %					
Authors	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
1029959	33.3	33.3	33.3	66.7	33.3
2069530	100.0	100.0	100.0	66.7	33.3
2369365	33.3	33.3	33.3	66.7	33.3
3182387	100.0	100.0	100.0	100.0	33.3
3298664	100.0	100.0	100.0	100.0	66.7
3420481	0.0	0.0	0.0	0.0	0.0
3454871	0.0	0.0	0.0	33.3	0.0
3520038	66.7	66.7	100.0	66.7	66.7
3535101	66.7	33.3	33.3	66.7	66.7
3701154	66.7	66.7	66.7	66.7	66.7

(b) Top-k, for k=250.

Success rate per author in %					
Authors	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
1029959	33.3	33.3	33.3	**66.7**	0.0
2069530	100.0	100.0	100.0	**100.0**	100.0
2369365	33.3	0.0	33.3	**66.7**	0.0
3182387	100.0	100.0	100.0	**100.0**	100.0
3298664	100.0	100.0	100.0	**100.0**	100.0
3420481	0.0	33.3	0.0	**33.3**	0.0
3454871	0.0	0.0	33.3	**33.3**	33.3
3520038	100.0	66.7	66.7	**100.0**	33.3
3535101	66.7	66.7	33.3	**66.7**	66.7
3701154	66.7	100.0	66.7	**66.7**	33.3

3182387 and 3298664 were almost always correctly classified for the tested parameter values in these tables. The reason is that these authors have distinctive writing styles in terms of part-of-speech patterns.

5 Conclusions

A novel approach using fixed or variable-length patterns of part-of-speech skip-grams or n-grams as features was presented for blog authorship attribution. An experimental evaluation using blog posts from 10 blog authors has shown that authors are accurately classified with more than 73.3% success rate, and an average accuracy of 94.7%, using a small number of POS patterns (e.g. $k = 250$),

Table 9: Accuracy per author using ngrams ($maxgap = 0$)

(a) Top-k, for k=100.

Authors	n,x				
	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
1029959	0.833	0.867	0.867	0.933	0.933
2069530	0.933	0.967	0.967	0.967	0.900
2369365	0.867	0.867	0.867	0.800	0.933
3182387	1.000	0.967	0.967	1.000	0.833
3298664	0.933	0.867	0.900	1.000	0.967
3420481	0.833	0.867	0.867	0.867	0.700
3454871	0.867	0.900	0.900	0.900	0.900
3520038	0.967	0.967	1.000	0.900	0.867
3535101	0.933	0.900	0.900	0.967	0.900
3701154	0.967	0.900	0.900	0.933	0.867

(b) Top-k, for k=250.

Authors	n,x				
	$1,2$	$1,3$	$1,4$	$2,2$	$3,3$
1029959	0.933	0.867	0.867	0.933	0.833
2069530	0.933	0.900	0.967	1.000	1.000
2369365	0.900	0.900	0.867	0.833	0.800
3182387	1.000	1.000	1.000	1.000	1.000
3298664	0.833	0.867	0.833	0.967	1.000
3420481	0.833	0.867	0.867	0.900	0.767
3454871	0.867	0.900	0.900	0.933	0.900
3520038	1.000	0.967	0.967	0.967	0.767
3535101	0.933	0.933	0.900	0.967	0.933
3701154	0.967	1.000	0.967	0.967	0.933

and that it is unnecessary to create signatures with a large number of patterns. Moreover, it was found that using fixed length POS bigrams provided better results than using POS ngrams or using a larger gap between POS tags, and that using fixed length patterns is preferable to using variable-length patterns. To give the reader a feel of the relative efficacy of this approach on traditional long texts, we present a summary table of results on a corpus of 30 books by ten 19th century authors (c.f. Table 11 for comparative results) [18, 19]. In future work, we plan to develop other features and also evaluate the proposed features on other types of texts.

Acknowledgements. This work is financed by a National Science and Engineering Research Council (NSERC) of Canada research grant, and the Faculty of Research and Graduate Studies of the Université de Moncton.

Table 10: Summary of Best Success Rate Results (Best R_1 Rank)

Markers	Ngrams		Skip-grams		Ngrams vs Skip-grams	
Parameters	k	50	k	250	k	250
	n,x	$1,2$	n,x	$3,3$	n,x	$2,2$
			$maxgap$	1	$maxgap$	0
Results in %	R_1	73.3	R_1	70.0	R_1	73.3
	R_2	83.3	R_2	76.6	R_2	86.6
	R_3	90.0	R_3	86.6	R_3	86.6
Nb. of words	2,615,856 (30 books, 10 authors)		idem (30 books, 10 authors)		265,263 (609 posts, 10 authors)	

References

1. Koppel, M., Schler, J., Argamon, S.: Authorship attribution: What's easy and what's hard? Available at SSRN 2274891 (2013)

2. Mendenhall, T.C.: The characteristic curves of composition. Science **9**(214) (1887) 237–246
3. Mosteller, F., Wallace, D.: Inference and disputed authorship: The federalist. (1964)
4. Grant, T.: Text messaging forensics: Txt 4n6: idiolect free authorship analysis? (2010)
5. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Computational linguistics **26**(4) (2000) 471–495
6. Baayen, H., van Halteren, H., Tweedie, F.: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing **11**(3) (1996) 121–132
7. Clark, J.H., Hannon, C.J.: A classifier system for author recognition using synonym-based features. In: Proc. 6th Mexican Intern. Conf. on Artificial Intelligence. Springer (2007) 839–849
8. McDonald, A.W., Afroz, S., Caliskan, A., Stolerman, A., Greenstadt, R.: Use fewer instances of the letter i: Toward writing style anonymization. In: Privacy Enhancing Technologies, Springer (2012) 299–318
9. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. Volume 6. (2006) 199–205
10. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Computer-based authorship attribution without lexical measures. Computers and the Humanities **35**(2) (2001) 193–214
11. Gamon, M.: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: Proc. 20th Intern. Conf. on Computational Linguistics, Association for Computational Linguistics (2004) 611
12. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications **41**(3) (2014) 853–860
13. García-Hernández, R.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: Finding maximal sequential patterns in text document collections and single documents. Informatica **34**(1) (2010)
14. Litvinova, T., Seredin, P., Litvinova, O.: Using part-of-speech sequences frequencies in a text to predict author personality: a corpus study. Indian Journal of Science and Technology **8**(S9) (2015) 93–97
15. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proc. 2003 Conf. North American Chapter of the ACL - Human Language Technologies. (2003) 173–180
16. Fournier-Viger, P., Gomariz, A., Gueniche, T., Mwamikazi, E., Thomas, R.: TKS: Efficient mining of top-k sequential patterns. In: Proc. 9th Intern. Conf. on Advanced Data Mining and Applications. Springer (2013) 109–120
17. Argamon-Engelson, S., Koppel, M., Avneri, G.: Style-based text categorization: What newspaper am i reading. In: Proc. AAAI Workshop on Text Categorization. (1998) 1–4
18. Pokou, J.M., Fournier-Viger, P., Moghrabi, C.: Authorship attribution using small sets of frequent part-of-speech skip-grams. In: Proc. 29th Intern. Florida Artificial Intelligence Research Society Conference. (2016) 86–91
19. Pokou, J.M., Fournier-Viger, P., Moghrabi, C.: Authorship attribution using variable-length part-of-speech patterns. In: Proc. 7th Intern. Conf. on Agents and Artificial Intelligence. (2016) 354–361