Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Authorship Attribution Using Small Sets of Frequent Part-of-Speech Skip-Grams

Yao Jean Marc Pokou[1], **Philippe Fournier-Viger**[1,2], **Chadia Moghrabi**[1]
[1]Dept. of Computer Science, Université de Moncton, Moncton, NB, Canada
[2]School of Natural Sciences and Humanities, Harbin Inst. of Techn., Shenzhen Grad. School, Guangdong, China
{eyp3705, philippe.fournier-viger, chadia.moghrabi}@umoncton.ca

Abstract

Computer-supported authorship attribution provides tools for extracting stylistic features that can help verify or identify the author of text documents. In many situations finding the author of a document is very important, such as the detection of plagiarism for protecting copyrights and forensic support during criminal investigations.

This paper, thus explores a novel stylistic feature with the aim of accurately characterizing an author's work. In particular, the use of part-of-speech skip-grams and an in-house top-k sequential pattern mining algorithm are considered for the task of authorship attribution. A study using a collection of of 30 texts, written by 10 authors, consisting of $2,615,856$ words and $99,903$ sentences, confirms that mining part-of-speech skip-grams in texts facilitates authorship inference.

Introduction

Authorship attribution (AA) is a problem of classification, where an anonymous text (an unlabeled instance) needs to be attributed to an author from a possible set of authors (classes). Authorship attribution has played an important role in many forensic investigations by narrowing the list of suspects (Morton and Michaelson 1990; Crain 1998). Several methods have been developed for AA. But a key problem for AA is to choose appropriate features to accurately classify a set of texts. Features can be of different types (Stamatatos, Fakotakis, and Kokkinakis 2000) such as lexical, semantic (Clark and Hannon 2007) and syntactic (Uzuner, Katz, and Nahnsen 2005). Despite a large number of proposed features, there are no specific markers or sets of features that are well known to be fully accurate in all situations. For this reason, it is crucial to find new markers and methods to improve state-of-the-art systems for authorship attribution.

This paper explores a novel stylistic feature with the aim of accurately characterizing an author's work. It considers the use of part-of-speech skip-grams for the task of authorship attribution. Part-of-speech skip-grams are constructed like part-of-speech n-grams but they allow a skip distance or gap between adjacent part-of-speech (POS) tags that reflects an author's habit of using a particular sentence structure. 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. David Guthrie et al. explored skip-gram modeling and their results demonstrated their usefulness in case of data sparsity compared to traditional n-grams (Guthrie et al. 2006). The novel approach introduced in this paper is accomplished through three main steps. First, a set of training texts with known authors is preprocessed using a part-of-speech tagger. Then, a modified top-k sequential pattern mining algorithm is employed to mine the k most frequent part-of-speech skip-grams in each training text (Fournier-Viger et al. 2013). These frequent patterns are then used to generate a unique signature representing the writing style of each author. Finally, the extracted signatures are used to classify anonymous texts.

The proposed approach is different from previous works using n-grams of parts of speech of fixed size for authorship attribution (Argamon-Engelson, Koppel, and Avneri 1998; Gamon 2004). In this work, skip-grams are used instead of n-grams, that is the proposed approach allows gaps between parts of speech in sentences to better accommodate the personal style of an individual. Furthermore, unlike several previous studies, the proposed approach only extracts the k most frequent skip-grams rather than calculating the frequencies of all POS skip-grams thus reducing the total processing time. Lastly, the proposed approach is also more flexible by discovering part-of-speech skip-grams of various lengths and by considering various gap sizes.

An experimental study is described using a collection of 30 texts, written by 10 authors, consisting of a total of $2,615,856$ words. The study confirms that discovering part-of-speech skip-grams in texts facilitates authorship inference. Moreover, results show that using gaps provides generally better results than using part-of-speech bigrams and trigrams. Lastly, an interesting observation is that using a small set of only 50 to 100 most frequent skip-grams of varied size can lead to high classification accuracy.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset used in this study. Section 4 presents the proposed methodology. Section 5 describes the experimental results. Lastly, section 6 presents the conclusions.

Related Work

Since the 19th century, numerous researches have focused on finding the right set of features to reveal the style of an author. Among the earliest approaches, Medenhall found that every author's word-length curve is preserved through his texts (Mendenhall 1887). Similarly, Mosteller and Wallace proposed a more thorough study using a Bayesian statistical analysis method, which highlighted notable discrimination between the authors of the 12 Federalist Papers anonymously published in 1787-1788 (Mosteller and Wallace 1964).

More recently, lexical and pseudo-syntactic features were used such as part-of-speech trigrams for classifying newspapers and magazine articles (Argamon-Engelson, Koppel, and Avneri 1998). In their experimentation, 720 documents were used for training, each of them containing between 300 to 1,300 words. An error rate of 15% was then achieved for correctly classifying 80 test documents.

Chaski et al. developed reliable methods for authorship attribution. After splitting their corpus data into smaller chunks, they analysed them using a discriminant function with linguistic variables (punctuations, syntactic and lexical terms) that maximize the difference between groups (Chaski 2005).

More deep linguistic features like syntactic information was also used in the task of Authorship attribution by Baayen et al. They used around 20,000 words from two English books and wrote authorship attribution frequency rules based on the syntactic annotations of their corpus (Baayen, Van Halteren, and Tweedie 1996). The combination of shallow linguistic features (lexical, semantic) with syntactic markers such as parts of speech can help achieve highly accurate identification of authors of short texts as demonstrated by the study of Gamon et al. (Gamon 2004). They used multiple features such as the frequencies of function words extracted using the NLPWin system, and frequencies of part-of-speech trigrams obtained by the system. Gamon et al. achieved 85% classification accuracy when using all the above listed features together and concluded that removing deep linguistic analysis features (part-of-speech trigrams) decreased accuracy.

Sidorov et al. (Sidorov et al. 2014) used syntactic n-grams (sn-grams) where elements are selected not by their order of appearance in the text but rather their position in the syntatic tree. Their experiment showed that sn-grams put the emphasis on syntactic relations between words. Sidorov et al. used sn-grams of words among several types of sn-grams: part-of-speech, character, mixed sn-grams. They used three classifiers from Weka (NormalizedPolyKernel of the SMO, Naive Bayes, and J48). Syntactic n-grams topped traditional n-grams in the experiments conducted on a corpus of 39 documents by three authors (Sidorov et al. 2014).

Data

The data used for the experiments is extracted from Project Gutenberg [1] and consists of 30 books from 10 different English novelists from the XIX century, each of them having exactly 3 books (for a total of $2,615,856$ words and $99,903$ sentences). The total number of words/sentences in the corpus of each author is as follows: Catharine Traill (276,829/ 6,588), Emerson Hough (295,166/ 15,643), Henry Addams (447,337/ 14,356), Herman Melville (208,662/ 8,203), Jacob Abbott (179,874/ 5,804), Louisa May Alcott (220,775/ 7,769), Lydia Maria Child (369,222/ 15,159), Margaret Fuller (347,303/ 11,254), Stephen Crane (214,368/ 12,177), and Thornton W. Burgess (55,916/ 2,950).

[1] https://www.gutenberg.org/

Methodology

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 composed of three phases, which are described in the following subsections.

Phase 1: Data Preparation and Transformation

In the first phase, texts from the corpus are prepared by removing content which does not reflect the author's style, such as illustrations. Then, because the proposed method uses part-of-speech skip-grams, all punctuations from the texts are also removed, using the Rita Natural Language processing library (Howe 2009). Then, each word is replaced by its corresponding one of 36 part-of-speech (POS) tags in each sentence using the Standford NLP tagger as it offered a 97.24 % accuracy on the Penn Treebank Wall Street Journal corpus (Toutanova et al. 2003). For example, consider the following three sentences from the book Eight Cousins by Louisa May Alcott, which will be used to illustrate the proposed method : *Running down the long hall, she peeped out at both doors, but saw nothing feathered except a draggle-tailed chicken under a burdock leaf. She listened again, and the sound seemed to be in the house. Away she went, much excited by the chase, and following the changeful song, it led her to the china-closet door.* After the first phase, the following sentence sequences are obtained: *VBG RP DT JJ NN PRP VBD RP IN DT NNS CC VBD NN JJ IN DT VBN NN IN DT NN NN. PRP VBD RB CC DT JJ VBD TO VB IN DT NN. RB PRP VBD RB VBN IN DT NN CC VBG DT JJ NN PRP VBD PRP TO DT NN NN.*

Phase 2: Signature Extraction

The second phase of the proposed approach consists of extracting a signature for each author. The signature of an author is a set of part-of-speech skip-grams (PSG) annotated with their respective frequency. Our approach has four parameters: the number of part-of-speech skip-grams to be found k, the minimum sequence length n, the maximum length x, and the maximum gap $maxgap$ allowed between part-of-speech tags. Note that the part-of-speech skip-grams are of various lengths with various gap sizes, as mentioned earlier. The second phase is performed in two steps.

Extracting skip-grams from each text. The first step consists of extracting part-of-speech skip-grams from each corpus text t. Part-of-speech skip-grams are similar to part-of-speech n-grams but allow a gap between adjacent elements.

Definition 1 (part-of-speech skip-gram) *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 $i_j - i_{j-1} \leq maxgap + 1 (1 < j \leq n)$. Note that part-of-speech n-grams are a special case of part-of-speech skip-grams where $maxgap = 0$ (i.e. no gaps).*

For each text t, the k most frequent POS skip-grams are extracted. The frequency of a skip-gram is defined as the number of sentences containing the skip-gram divided by the total number of sentences. In the following, the term *part-of-speech skip-grams* of t, abbreviated as $(PSGt)^k_{n,x}$, or *patterns*, is used to refer to those POS skip-grams, annotated with their relative frequency.

Creating the signature of each author. The second step is to create a signature for each author. For a given author a_i, this is performed as follows. First, the POS skip-grams appearing in any of the texts written by the author are found.

Definition 2 *The* part-of-speech skip-grams of an author a_i *is a set denoted as $(PSGa_i)^k_{n,x}$ and defined as the union of the POS skip-grams found in all of his/her texts, i.e. $(PSGa_i)^k_{n,x} = \bigcup_{t \in T_i} (PSGt)^k_{n,x}$*

For example, consider the paragraph written by the author Louisa May Alcott, presented in the previous subsection. The set $(PSG_{Alcott})^5_{1,3}$ of the $k = 5$ most frequent part-of-speech skip-grams having a length between $n = 1$ and $x = 3$ in this text for $maxgap = 1$, is IN-NN, IN-DT-NN, IN-DT, IN and DT, each having a relative frequency of 100.0 %. In this example, the tags IN, DT and NN respectively represents Preposition or subordinating conjunction, Determiner, and Noun, singular or mass.

In this example, it can be seen that allowing a gap of 1 tag between two adjacent tags in a skip-gram allows the discovery of **IN-NN** which is present in the three sentences. This skip-gram has therefore a frequency of 100%. It is important to note that traditional POS n-grams such as POS bigrams would not find the pattern **IN-NN** in these sentences. The occurrences of the skip-gram **IN-NN** are highlighted here: *VBG RP DT JJ NN PRP VBD RP IN DT NNS CC VBD NN JJ IN DT VBN NN **IN** DT **NN** NN. PRP VBD RB CC DT JJ VBD TO VB **IN** DT **NN**. RB PRP VBD RB VBN **IN** DT **NN** CC VBG DT JJ NN PRP VBD PRP TO DT NN NN*

Then, the signature of the author a_i is extracted by performing the intersection of the part-of-speech skip-grams appearing in his/her texts.

Definition 3 *Let a_i be an author and T_i be the set of texts written by a_i. The signature s_{a_i} of a_i is the intersection[2] of the POS skip-grams of his/her texts, formally defined as:*

$$(s_{a_i})^k_{n,x} = \bigcap_{t \in T_i} (PSGt)^k_{n,x}$$

This work supposes that the POS skip-grams 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. This set is formally defined as:

Definition 4 (common POS skip-grams excluding an author) *The* Common POS skip-grams of all authors excluding an author a_i *is the union of all the PSG of these authors, i.e.*

$$(CPSGa_i)^k_{n,x} = \bigcup_{a \in A \wedge a \neq a_i} (PSGa)^k_{n,x}$$

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

The overall process for extracting each author's signature takes as input a set of authors with their texts, plus the parameters n, x and k, and outputs the revised signatures for each author. How to best set the parameters n, x and k to obtain optimal accuracy for authorship attribution will be discussed in the results section.

Phase 3: Author Classification

The third phase of the proposed approach consists of using the extracted signatures for classifying anonymous texts. A classification algorithm is thus developed to identify the author a_u of an anonymous text t_u that was not used for training. This is performed by Algorithm 1.

The algorithm takes as input an anonymous text t_u, the sets $s'^{,k}_{n,x}$, and the parameters $n, x, maxgap$, and k. The algorithm first extracts the part-of-speech skip-grams 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 his/her similarity value is 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 such Euclidian distance, Pearson correlation and cosine similarity. In this work, the Pearson correlation is chosen as it provided better results in initial experiments.

Results

A set of experiments was performed to assess the effectiveness of the proposed approach for authorship attribution based on the usage of POS skip-grams. Each text from the corpus was preprocessed. Then, for learning and assessing the performance of the proposed approach, leave-one-out cross-validation was used. Thus, for each text, the designed system was trained using the 29 other texts. The common POS skip-grams of the 29 other texts were created and used

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

Algorithm 1: Establishing Authorship Candidates.

input : an anonymous text t_u, the sets $s'^{k}_{n,x}$, the parameters n, x, $maxgap$ and k
output: a list L ranking the most likely authors of t_u

Extract $(PSG_{t_u})^k_{n,x}$ from t_u;
$L = \emptyset$;
foreach $(s'_{a_i})^k_{n,x} \in s'^{k}_{n,x}$ **do**
    $similarity = similarity((PSG_{t_u})^k_{n,x}, (s'_{a_i})^k_{n,x})$;
    Insert $(a_i, similarity)$ in L;
end
return L *sorted by decreasing similarity values*

to create the signatures of the 10 authors. The validation consisted of comparing the signature of this text with the 10 author signatures to rank them from the most likely to the least likely one. This whole process was performed for the 30 texts. Note that 80% of each text was used for learning and the remaining 20% for testing.

Influence of parameters on overall results

Recall that our proposed approach takes four parameters as input, i.e. the minimum and maximum length of part-of-speech skip-grams n and x, the number of patterns to be extracted k in each text, and $maxgap$. The influence of these parameters on authorship attribution success was first evaluated. For our experiment, parameter k was set to 50, 100, 250, and 500, and the parameter $maxgap$ was set to 1, 2 and 3. For each value of k, the length of part-of-speech skip-grams was varied from $n = 2$ to $x = 5$.

Tables 1 and 2 respectively show the results obtained for $maxgap = 1$ and 2, for the various values of n, x and k. (Results for k = 500 and $maxgap$ = 3 are not included due to space limitations.) Furthermore, in each subtable, the results are also presented by ranks. The column R_z represents the number of texts where the author was predicted as one of the z most likely authors, divided by the total number of texts. This measure is called the *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, with z is varied from 1 to 10.

From these results, we can make several observations. First, the best overall results are achieved by $n = 1$ and $x = 2$ for $maxgap = 1$ and $k = 250$. For these parameters, 66.67% of the texts when anonymously considered have their author correctly recognized (rank R_1), 80.0% of texts are attributed to the two most likely authors (R_2), and 93.34% to one of the three most likely authors (R_3). The next best results are for k = 100, with the same n and x, where the success rates are 63.33%, 80.0%, and 93,33%, respectively.

Second, it is interesting to observe that increasing the number of patterns beyond 250 generally does not provide better results. This is interesting because it means that signatures can be extracted using a very small number of patterns such as k varying from 50 to 250 and still characterize well

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

(a) k = 50.

Success ratio in %				
n, x	$1, 2$	$1, 3$	$1, 4$	$1, 5$
R_1	56.67	50.0	60.0	60.0
R_2	73.34	76.67	76.67	80.0
R_3	90.01	90.0	93.34	93.33
R_4	93.34	90.0	93.34	93.33
R_{5-6}	93.34	96.67	96.67	96.66
R_{7-10}	100.0	100.0	100.0	100.0

(b) k = 100.

Success ratio in %				
n, x	$1, 2$	$1, 3$	$1, 4$	$1, 5$
R_1	63.33	60.0	63.33	63.33
R_2	80.0	83.33	76.66	76.66
R_3	93.33	93.33	89.99	93.33
R_4	96.66	96.66	93.32	96.66
R_{5-6}	96.66	96.66	96.64	96.66
R_{7-10}	100.0	100.0	100.0	100.0

(c) k = 250.

Success ratio in %				
n, x	$1, 2$	$1, 3$	$1, 4$	$1, 5$
R_1	66.67	66.67	66.67	66.67
R_2	80.0	76.67	76.67	76.67
R_3	90.0	93.34	93.34	93.34
R_4	90.0	93.34	93.34	93.34
R_{5-6}	96.67	96.67	96.67	96.67
R_{7-10}	100.0	100.0	100.0	100.0

the writing style of authors. This is in contrast with previous works that have used a large amount of n-grams. For example, Argamon et al. have suggested computing the frequencies of 685 trigrams (Argamon-Engelson, Koppel, and Avneri 1998) and Sidorov et al. computed the frequencies of 400 to 11,000 n-grams/sn-grams.

Third, it can be observed that a smaller gap ($maxgap = 1$) is generally better than using a larger gap.

To put these results into perspective, this paper also compares the results with part-of-speech bigrams and trigrams (i.e. skip-grams with $maxgap = 0$), used in previous work (Argamon-Engelson, Koppel, and Avneri 1998; Koppel and Schler 2003). Table 3 shows the overall results for bigrams and trigrams. It can be seen that bigrams with $k = 250$ achieves the best results, which is quite close to the best results obtained with skip-grams. However, the results with skip-grams can be considered better since the success rate for predicting the authors correctly is 66.67% with skip-grams in R_1, 80.0% in R_2, and 90.0% in R_3 as opposed to 66.7%, 76,67%, and 86.67% using part-of-speech bigrams.

Influence of parameters on authorship attribution for each author

This section analyzes the results for each author separately for $maxgap = 1$. Recall that each author has three texts

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

(a) k = 50.

Success ratio in %				
n, x	$1, 2$	$1, 3$	$1, 4$	$1, 5$
R_1	50.0	56.67	56.67	56.67
R_2	70.0	76.67	73.34	73.34
R_3	73.33	83.34	83.34	83.34
R_4	83.33	83.34	83.34	83.34
R_5	90.0	90.01	93.34	93.34
R_6	93.33	90.01	96.67	93.34
R_7	96.66	96.68	96.67	96.67
R_{8-10}	100.0	100.0	100.0	100.0

(b) k = 100.

Success ratio in %				
n, x	$1, 2$	$1, 3$	$1, 4$	$1, 5$
R_1	56.67	53.33	43.33	46.67
R_2	76.67	73.33	76.66	76.67
R_3	93.34	86.66	89.99	86.67
R_4	96.67	93.33	96.66	93.34
R_5	96.67	96.66	96.66	93.34
R_6	96.67	96.66	96.66	96.67
R_7	96.67	100.0	100.0	100.0
R_{8-10}	100.0	100.0	100.0	100.0

(c) k = 250.

Success ratio in %				
n, x	$1, 2$	$1, 3$	$1, 4$	$1, 5$
R_1	63.33	56.67	56.67	63.33
R_2	86.66	76.67	73.34	80.0
R_3	93.33	96.67	96.67	96.67
R_4	93.33	96.67	96.67	96.67
R_5	93.33	96.67	96.67	96.67
R_6	93.33	96.67	96.67	96.67
R_7	100.0	100.0	100.0	96.67
R_{8-10}	100.0	100.0	100.0	100.0

Table 3: Bi-grams and tri-grams top-K, for k=50,100 and 250 with $maxgap$ = 0.

Success ratio in %						
	k = 50		k = 100		k = 250	
n, x	$2, 2$	$3, 3$	$2, 2$	$3, 3$	$2, 2$	$3, 3$
R_1	56.67	43.33	63.33	63.33	66.67	70.0
R_2	73.34	50.0	80.0	76.66	76.67	76.67
R_3	83.34	70.0	90.0	89.99	86.67	86.67
R_4	93.34	73.33	90.0	93.32	90.0	86.67
R_5	96.67	80.0	93.33	93.32	90.0	86.67
R_6	96.67	80.0	96.66	93.32	90.0	86.67
R_7	100.0	86.67	100.0	100.0	93.33	90.0
R_8	100.0	93.34	100.0	100.0	93.33	93.33
R_{9-10}	100.0	100.0	100.0	100.0	100.0	100.0

in the corpus. Table 4 shows the number of texts correctly attributed to each author (R_1). It can be observed that for most authors, generally two out of three texts are correctly attributed. For example, for $n = 1$, $x = 2$ and $k = 250$, four authors have two texts correctly classified and two have all three texts correctly identified.

Furthermore, some authors are harder to classify. For instance, the proposed approach rarely identifies more than two of the three texts written by Henry Addams. Those texts are: "Democracy, an American Novel", "The Education of Henry Addams" and "Mont-Saint-Michel and Chartres". The first text is a political novel that was written anonymously in 1881 and was attributed to Addams after his death. A plausible explanation is that Addams may have attempted to hide his writings to preserve his anonymity. As a result, the signature of Addams may be less coherent, with low success for AA of his texts. However, increasing k to 250 improved his success rate to two out of three.

Likewise, the success ratio for Catharine Traill is equal to zero with 0.9 accuracy (Table 5). She reported daily routine life between Canadians and Natives. This could explain that her POS patterns were subtracted by the common POS. Moreover, the study done by Pokou et al. used the same data set obtained the same success rates (Pokou, Founier-Viger, and Moghrabi 2016) for ngrams.

Some authors were easily identified, such as Jacob Abbott whose texts were correctly classified for all tested parameters. The reason is thus that Jacob Abbott has a more distinctive writing style in terms of POS skip-grams.

Table 4: Success ratio per author using skip-grams with $maxgap$ = 1

(a) k=100.

Authors	n,x			
	$1, 2$	$1, 3$	$1, 4$	$1, 5$
Catharine Traill	0/3	0/3	0/3	0/3
Emerson Hough	1/3	1/3	1/3	1/3
Henry Addams	2/3	1/3	1/3	1/3
Herman Melville	2/3	2/3	2/3	2/3
Jacob Abbott	3/3	2/3	3/3	3/3
Louisa May Alcott	2/3	2/3	2/3	2/3
Lydia Maria Child	2/3	3/3	3/3	3/3
Margaret Fuller	3/3	3/3	3/3	3/3
Stephen Crane	1/3	1/3	1/3	1/3
Thornton WBurgess	3/3	3/3	3/3	3/3

(b) k=250.

Authors	n,x			
	$1, 2$	$1, 3$	$1, 4$	$1, 5$
Catharine Traill	0/3	0/3	0/3	0/3
Emerson Hough	2/3	2/3	2/3	2/3
Henry Addams	2/3	2/3	2/3	2/3
Herman Melville	2/3	2/3	2/3	2/3
Jacob Abbott	3/3	3/3	3/3	3/3
Louisa May Alcott	2/3	2/3	2/3	2/3
Lydia Maria Child	2/3	2/3	2/3	2/3
Margaret Fuller	3/3	3/3	3/3	3/3
Stephen Crane	1/3	1/3	1/3	1/3
Thornton WBurgess	3/3	3/3	3/3	3/3

Table 5: Accuracy per author using skip-grams with $maxgap$ = 1

(a) k=100.

Authors	n,x			
	$1,2$	$1,3$	$1,4$	$1,5$
Catharine Traill	0.900	.900	.900	.900
Emerson Hough	.833	.833	.833	.833
Henry Addams	.967	.933	.933	.933
Herman Melville	.967	.900	.867	.900
Jacob Abbott	1.00	.967	1.00	1.00
Louisa May Alcott	.933	.967	.967	.967
Lydia Maria Child	.933	.933	.967	.967
Margaret Fuller	.900	.933	.967	.933
Stephen Crane	.867	.867	.867	.867
Thornton WBurgess	.967	.967	.967	.967

(b) k=250.

Authors	n,x			
	$1,2$	$1,3$	$1,4$	$1,5$
Catharine Traill	.900	.900	.900	.900
Emerson Hough	.867	.867	.867	.867
Henry Addams	.967	.967	.967	.967
Herman Melville	.967	.967	.933	.933
Jacob Abbott	1.00	1.00	1.00	1.00
Louisa May Alcott	.933	.933	.933	.933
Lydia Maria Child	.933	.933	.933	.933
Margaret Fuller	.900	.900	.933	.933
Stephen Crane	.867	.867	.867	.867
Thornton WBurgess	1.00	1.00	1.00	1.00

Conclusions

In this paper, the use of part-of-speech skip-grams was considered to accurately characterize an author's work for authorship attribution. A study using a collection of 30 texts, written by 10 authors, consisting of $99,903$ sentences and $2,615,856$ words, was carried to evaluate the proposed approach. Experimental results have shown that authors can be well classified with more than 66.7% accuracy using a very small number of part-of-speech skip-grams (e.g. $k = 250$) and that using a small gap (e.g. $maxgap = 1$) provided the best results, with an average accuracy between 85,8%, 92,6%, and 93.34% for $k = 50, 100, and 250$ respectively.

For future work, a longer list of authors is in the planning, as well as different styles of texts such as blogs and e-mails.

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

References

Argamon-Engelson, S.; Koppel, M.; and Avneri, G. 1998. Style-based text categorization: What newspaper am i reading. In *Proc. AAAI Workshop on Text Categorization*, 1–4.

Baayen, H.; Van Halteren, H.; and Tweedie, F. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3):121–132.

Chaski, C. E. 2005. Whos at the keyboard? authorship attribution in digital evidence investigations. *Intern. Journal of Digital Evidence* 4(1):1–13.

Clark, J. H., and Hannon, C. J. 2007. A classifier system for author recognition using synonym-based features. In *MICAI 2007: Advances in Artificial Intell.* Springer. 839–849.

Crain, C. 1998. The bards fingerprints. *Lingua Franca* 4:29–39.

Fournier-Viger, P.; Gomariz, A.; Gueniche, T.; Mwamikazi, E.; and Thomas, R. 2013. Tks: Efficient mining of top-k sequential patterns. In *Advanced Data Mining and Applications*. Springer. 109–120.

Gamon, M. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proc. 20th intern. conf. on Computational Linguistics*, 611.

Guthrie, D.; Allison, B.; Liu, W.; Guthrie, L.; and Wilks, Y. 2006. A closer look at skip-gram modelling. In *Proc. of the 5th Intern. Conference on Language Resources and Evaluation (LREC-2006)*, 1–4.

Howe, D. C. 2009. Rita: creativity support for computational literature. In *Proc. of the seventh ACM conference on Creativity and cognition*, 205–210. ACM.

Koppel, M., and Schler, J. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proc. of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, 72.

Mendenhall, T. C. 1887. The characteristic curves of composition. *Science* 237–249.

Morton, A. Q., and Michaelson, S. 1990. *The qsum plot*, volume 3. Univ. of Edinburgh, Dept. of Computer Science.

Mosteller, F., and Wallace, D. 1964. Inference and disputed authorship: The federalist.

Pokou, Y. J. M.; Founier-Viger, P.; and Moghrabi, C. 2016. Authorship attribution using variable length part-of-speech patterns. In *ICAART*, 75. IEEE.

Sidorov, G.; Velasquez, F.; Stamatatos, E.; Gelbukh, A.; and Chanona-Hernández, L. 2014. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications* 41(3):853–860.

Stamatatos, E.; Fakotakis, N.; and Kokkinakis, G. 2000. Automatic text categorization in terms of genre and author. *Computational linguistics* 26(4):471–495.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. 2003 Conf. of North American Chapter of the ACL on Human Language Technology*, 173–180. Assoc. for Computational Linguistics.

Uzuner, Ö.; Katz, B.; and Nahnsen, T. 2005. Using syntactic information to identify plagiarism. In *Proc. of the second workshop on Building Educational Applications Using NLP*, 37–44. Assoc. for Computational Linguistics.