Authorship Attribution using Variable Length Part-of-Speech Patterns

Yao Jean Marc Pokou, Philippe Fournier-Viger and Chadia Moghrabi
Dept. of Computer Science, Université de Moncton, Moncton, NB, Canada

Keywords: Authorship Attribution, Stylometry, Part-of-Speech Tags, Variable Length Sequential Patterns.

Abstract: Identifying the author of a book or document is an interesting research topic having numerous real-life applications. A number of algorithms have been proposed for the automatic authorship attribution of texts. However, it remains an important challenge to find distinct and quantifiable features for accurately identifying or narrowing the range of likely authors of a text. In this paper we propose a novel approach for authorship attribution, which relies on the discovery of variable-length sequential patterns of parts of speech to build signatures representing each author's writing style. An experimental evaluation using 10 authors and 30 books, consisting of $2,615,856$ words, from Project Gutenberg was carried. Results show that the proposed approach can accurately classify texts most of the time using a very small number of variable-length patterns. The proposed approach is also shown to perform better using variable-length patterns than with fixed-length patterns (bigrams or trigrams).

1 INTRODUCTION

Throughout history many well known books and documents were published under pseudonyms leaving a doubt on the identity of their authors. The challenges related to identifying their authorship remains a topical issue (Koppel et al., 2007). In addition, identifying an author could be useful in detecting plagiarism, clearing controversies over a disputed authorship and in some cases it helps ensuring justice by confirming or clearing a suspect during a forensic investigation. In response to such problems, a multidisciplinary research field known as Stylometry was introduced. It consists of finding new markers that exploit any stylistic attributes of the text in order to achieve a better author identification. In fact, a variety of the idiosyncratic features extracted from text are measurable. Medenhall laid the foundation of Authorship Attribution by highlighting the uniqueness of the word-length frequency curve for every author (Mendenhall, 1887). His studies showed a connection between the work of William Shakespeare and Christopher Marlow. Yule proposed other statistical features like the average sentences and word lengths (Yule, 1939). But the most spectacular study was the one conducted by Mosteller and Wallace, who applied Bayesian statistical analysis methods on the 12 Federalist papers anonymously published in 1787-1788 (Mosteller and Wallace, 1964). Most recent studies on authorship attribution focused on various types of features (Stamatatos et al., 2000) such as:

Syntactic: Syntactic features of text are considered reliable because they are unconsciously used by authors. They include frequencies of ngrams, character ngrams, function words, etc. (Koppel et al., 2013). Baayen et al. used syntactic features for the first time (Baayen et al., 1996). Their results show that, with a well implemented and fully-automated parser, syntactic features can outdo other markers.

Semantic: Semantic features take advantage of words meanings and their likeness. For example, Hannon and Clark defined a synonym-based classifier that quantifies how often an author uses a specific word instead of its synonyms (Clark and Hannon, 2007).

Because each feature may have limitations, state-of-the-art systems for authorship attribution often combine a wide-range of features to achieve higher accuracy. For example, the JStylo system offers more than 50 configurable features such as word frequencies, letter frequencies and length of paragraphs (e.g. (McDonald et al., 2012)). Although many features have been proposed, it still remains an important challenge to find new features that can characterize each author, for accurately identifying or narrowing the range of likely authors of a text (Stamatatos et al., 2000). In particular, syntactic markers have been less studied because of their language-dependant aspect.

Pokou, Y., Fournier-Viger, P. and Moghrabi, C.
Authorship Attribution using Variable Length Part-of-Speech Patterns.
DOI: 10.5220/0005710103540361
In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) - Volume 2*, pages 354-361
ISBN: 978-989-758-172-4

However, if they are used with accurate and robust Natural Language Processing Tools (NLP) and combined with others features, they are quite promising for identifying authors (Gamon, 2004).

In such context, this paper studies the possibility of using more complex linguistic information carried by parts of speech (POS) as a novel feature for authorship attribution. The hypothesis is that an analysis of POS ngrams appearing in texts, not limited to bigrams and trigrams, and combined with a thorough analysis of texts, could accurately characterize each author's style. The contributions of this paper are the following. We define a novel feature for identifying authors based on variable length POS ngrams. A signature is computed for each author as the intersection of the top k most frequent POS ngrams found in his texts, that are less frequent in texts by others authors. The system takes as input a training corpus of texts with known authors. Each text is tagged using the Standford NLP tagger (Toutanova et al., 2003), and individual signatures are generated. Finally, an algorithm is proposed to identify the author of a text by finding the most similar signature. An experimental evaluation using 30 books and 10 authors involving $2,615,856$ words from the Gutenberg project shows that our approach can infer authors with great accuracy.

In the rest of this paper, section 2 gives an overview of related work. Section 3 describes the proposed approach. Section 4 presents an experimental evaluation. Finally, section 5 draws the conclusions.

2 RELATED WORK

In literature, a style is a toolset used by an author to convey his ideas or describe events or entities. It is known that orators and well-known authors have a tendency to use the same words and stylistic features through their work, hence giving others the ability to identify them. Authorship Attribution (AA) is the process of inferring the identity of an author by analyzing his writings and extracting his unique markers.

The use of AA techniques and methods goes back to the 19th century with the studies of Shakespeare's work. However, Baayen and Van Halteren are the forerunners of the use of syntactic features in representing the style of an author. They rewrote the frequencies rules for Authorship Attribution based on two syntactically annotated samples taken from the Nijmegen corpus (Baayen et al., 1996).

Similarly, Stamatatos et al. exclusively used natural language processing tools for isolating a set of three level style markers: the token level (the input text is considered as a sequence of tokens grouped in sentences), the phrase level (where the frequencies of single POS tags is considered), and the analysis-level. The latter comprises style markers that represent the way in which the input text has been analyzed by the sentence Chunk Boundaries Detector (SCBD). They used SCBD for analyzing a corpus of 300 texts by 10 authors, written in Greek. Stamatatos et al. achieved around 80% accuracy by excluding some less significant markers (Stamatatos et al., 2001).

It was also demonstrated by Gamon et al. that combining syntactic information and lexical features can be used to accurately identify authors (Gamon, 2004). They used various features such as the frequencies of ngrams, function words, and part-of-speech trigrams. Parts of speech were obtained using the NLPWin system, which considers a set of eight parts of speech. Gamon et al. varied the number of trigrams and obtained the best results with 505 trigrams. An accuracy of over 95% was obtained. However, an important limitation of this study is that it was evaluated using only three texts, written by three authors (Gamon, 2004).

Argamon et al. defined a feature consisting of the frequencies of 500 function words and 685 part-of-speech trigrams for text categorization. Their method was tested on collections of articles from four newspapers and magazines obtained from the Nexus database [1]. Argamon et al. used a five-fold validation and achieved an average of 71.4% accuracy in identifying the publication of an article using only part-of-speech trigrams. However, when combined with function words, parts-of-speech trigrams produced an average of 79.13% for the same task (Argamon-Engelson et al., 1998).

More recently, Sidorov et al. introduced syntactic ngrams (sngrams) as a feature for authorship attribution. It is important to note that 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. Sidorov et al. showed that there can be various types of sngrams according to the types of elements that form them. More than three types of sngrams are defined such as: (1) sngrams of POS tags, (2) sngrams of syntactic relations and (3) sngrams of words. In their experiment, Sidorov et al. compared the use of sngrams with ngrams of words, parts of speech, and characters. They used from 400 to 11,000 ngrams/sngrams of fixed length varying from 2 to 5. A corpus of 39 documents by three authors extracted from Project Gutenberg was used. Classification was performed using SVM, J48 and Naive Bayes imple-

[1] http://nexus.nrf.ac.za/

mentations provided by the WEKA machine learning library. The best results were obtained by SVM with sngrams (Sidorov et al., 2014). A limitation of this work is that it was evaluated with only three authors. Besides, the length of ngrams was predetermined.

Variations of ngrams have also been considered in the literature. For example, García-Hernández et al. designed an algorithm to discover skip-grams. 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. A criticism of skip-grams is that their number can be very large and they are discovered using complex algorithms (Sidorov et al., 2014). To reduce the number of skip-grams, a cut-off frequency threshold can be used (García-Hernández et al., 2010). Another variation is sequential rules of function words extracted from sentences (Boukhaled and Ganascia, 2015).

Ngrams of parts of speech have also been used for other problems related to authorship attribution such as predicting the personality of the author of a text. For example, Litvinova et al., used the frequencies of 227 possible part-of-speech bigrams as a marker for predicting personality (Litvinova et al., 2015).

Unlike previous work using part-of-speech ngrams, the approach presented in this paper uses variable length part-of-speech ngrams. Another distinctive characteristic of the proposed work is that it finds only the k most frequent part-of-speech ngrams of variable length in each text (where k is set by the user), rather than using a large set of ngrams 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 have used several hundred or thousands of ngrams.

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 $T_i = \{t_1, t_2, \dots t_z\}$ in the corpus. The proposed approach is composed of three modules described in the following subsections.

3.1 The Preprocessor Module

The preprocessor module prepares texts from the corpus so that they can be used for generating author signatures. The module performs two steps.

Table 1: An example of text transformation.

#	Original Sentence	Transformed Sentence into POS sequences
1	Now Alexander was born the heir to the throne of one of the Grecian kingdoms.	RB NNP VBD VBN DT NN TO DT NN IN CD IN DT JJ NNS
2	He possessed, in a very remarkable degree, the energy, and enterprise, and military skill so characteristic of the Greeks and Romans.	PRP VBD IN DT RB JJ NN DT **NN CC** NN CC JJ NN RB JJ IN DT NNPS CC NNPS
3	He organized armies, crossed the boundary between Europe and Asia, and spent the twelve years of his career in a most triumphant military incursion into the very center and seat...	PRP VBD NNS VBD DT NN IN NNP CC NNP CC VBD DT CD NNS IN PRP$ NN IN DT RBS JJ JJ NN IN DT JJ **NN CC** NN...

Removing Noise from Texts. The first step consists of removing noise from texts by removing all information that does not carry an author's style. For example, for a book this noise can be the preface, the index, the table of contents and sometimes illustrations added by the editor or publisher. The goal is to keep the original work of the author free of atypical elements. In addition, each text in the corpus is stripped of punctuations and is splitted into sentences using the Natural Language Processing Library Rita, developed by Howe (Howe, 2009).

Transforming Texts into Sequences of Part-of-Speech Symbols (Tags). The second step consists of tagging every text using the Standford NLP Tagger. This results in texts where each word is annotated with a part-of-speech (POS) tag, from a set of 36 possible part-of-speech tags. 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 information about parts of speech is maintained. Thus, each text becomes a set of sentences, where each sentence is a sequence of POS tags. For example, consider three sentences from *"History of Julius Caesar"* by *Jacob Abbott*, shown in Table 1. The second and third columns show the original and transformed sentences, respectively.

3.2 The Signature Extraction Module

After preprossessing every text from the corpus, each consisting of many sentences, they can be seen as sequences of symbols (POS tags). The signature extraction module takes these sequences as input and produces a signature representing the writing style of each author from the corpus. This process is performed as follows.

Finding POS Patterns in each Text. The first step is to find patterns of POS tags in each text. The hypothesis is that each text may contain patterns of POS tags unconsciously left by its author, representing his writing style, and could be used to identify that author accurately. In other words, Authorship Attribution is considered as the problem of discovering the right set of discriminative patterns that are recurrent in an author's work. In the field of data mining, several algorithms have been proposed for discovering patterns in sequences of symbols. According to Han and Kamber (Han et al., 2011), there are four main kinds of patterns that can be mined from sequences. These are trends, similar sequences, sequential patterns, and periodic patterns. This work chose to mine sequential patterns (Agrawal and Srikant, 1995; Fournier-Viger et al., 2013), as the main interest is finding subsequences of POS tags appearing frequently in multiple sentences of a text. Recently, Mwamikazi et al. used a similar approach for mining patterns of answers in adaptive questionnaires (Mwamikazi et al., 2014).

The task of discovering sequential patterns in a text is defined as follows. Let *POS* denotes the set of POS tags. A sequence or pattern of POS is an ordered list of symbols (tags) $\langle p_1, p_2, ...p_v\rangle$, where $p_i \in POS$ $(1 \leq i \leq v)$. A sequence $seq_a = \langle p_1, p_2, ...p_v\rangle$ is said to be contained in a sequence $seq_b = \langle q_1, q_2, ...q_w\rangle$ if there exist integers $1 \leq i1 < i2 < ... < iv \leq w$ such that $p_1 = q_{i1}$, $p_2 = q_{i2}$ ... $p_v = q_{iv}$. 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 sequence $\langle NN, CC\rangle$ is 2 in the text of Table 1 (this pattern appears in the second and third sentence).

The goal of sequential pattern mining is to find all frequent sequences (*sequential patterns*) in a text, that is patterns having a frequency (support) no less than a threshold *minsup* set by the user (Agrawal and Srikant, 1995; Fournier-Viger et al., 2013). In this work, the task of sequential pattern mining is adapted in three ways. First, disallow gaps between POS symbols in a pattern to ensure that they appear contiguously in sequences. Thus, a sequence $seq_a = \langle p_1, p_2, ...p_v\rangle$ is now said to be contained in a sequence $seq_b = \langle q_1, q_2, ...q_w\rangle$ if there exists an integer $1 \leq i \leq w$ such that $p_1 = q_i$, $p_2 = q_{i+1}$ ... $p_v = q_{i+v-1}$. Second, instead of using a fixed threshold *minsup* for discovering patterns, the *k* most frequent sequential patterns are discovered from a text (the top-k sequential patterns), where *k* is a parameter set by the user. Third, only patterns having a minimum length *n* and a maximum length *x* are searched for. Thus, in this approach, mining sequential patterns has three parameters: the number of patterns to be found *k*, the minimum length *n*, and the maximum length *x*.

For each text *t*, the *k* most frequent POS sequential patterns are extracted. In the following, the term *patterns* of *t*, abbreviated as $(POSPt)^k_{n,x}$ is used to refer to those patterns, 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 patterns appearing in any of the texts written by the author are found.

Definition 1. *The* POS patterns of an author a_i *is denoted as* $(POSPa_i)^k_{n,x}$ *and defined as the union of the POS patterns found in all of his texts, i.e.* $(POSPa_i)^k_{n,x} = \bigcup_{t \in T_i} (POSPt)^k_{n,x}$

For example, consider that the author J. Abbott has written a single text, illustrated in Table 1. Consider $(POSPa_{Abbott})^5_{1,3}$, the part-of-speech patterns of this author, where patterns have a length between $n = 1$ and $x = 3$, and $k = 5$. These patterns are NN (Noun, singular or mass), JJ (Adjective), DT (Determiner), PRP-VBD (Personal pronoun - Verb, past tense), and NNP (Proper noun, singular), which respectively have a frequency of 100.0%, 100.0%, 100.0%, 66.6%, and 66.6%. We can see that Noun (NN), Adjectives (JJ), and Determinant (DT) occur in every sentence (thus, have a relative frequency of 100%).

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

Definition 2. *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 of the POS patterns of his texts. The signature is formally defined as:*

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

For example, the signature $(s_{a_i})^5_{1,3}$ that has been computed using three books written by Jacob Abbott is the patterns PRP-VBD and NNP, both having a frequency of 66.6 %. Note that the relative frequency of each pattern is calculated as the relative frequency over all texts containing the pattern.

This work supposes that 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. This set is formally defined as:

Definition 3 (common part-of-speech patterns excluding an author)**.** *The* Common POS patterns of all authors excluding an author a_i *is the union of all the POSP of these authors, that is*

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

For example, consider the three authors: J. Abbott, C. Trail and L. M. Child. Table 2 shows the common part-of-speech patterns excluding the author Child ($(POSP_{Child})^5_{1,3}$) for two texts used in the experimental evaluation of this paper (the history of Julius Caesar by Jacob Abbott and A Tale of The Rice Lake Plains by Catharine Traill). The first column of the table shows the part-of-speech tags and the next columns indicate their relative frequencies. Note that the relative frequency of each pattern in CPOS is calculated as the relative frequency over all texts containing the pattern.

Table 2: Shared POS patterns between Abbott and Traill.

Patterns	Jacob Abbott (rel. frequency)	Catharine Traill (rel. frequency)
DT	90.5	89.9
IN	89.8	89.3
JJ	82.5	70.7
NN	93.3	91.2
VBD	77.1	87.2

When the signature of each author $a_1, a_2, ...a_m$ has been extracted, the collection of author signatures $s^k_{n,x} = \{s_1, s_2, ...s_m\}$ are saved, with the corresponding set of CPOS denoted as:

$$c^k_{n,x} = \{(CPOSa_1)^k_{n,x}, (CPOSa_2)^k_{n,x}, ...(CPOSa_m)^k_{n,x}\}$$

The algorithm for extracting each author signature and the corresponding CPOS takes as input a set of authors with their texts, plus the parameters n, x and k, and outputs the signatures and CPOS 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 experimental evaluation section.

3.3 The Authorship Attribution Module

After the signatures have been generated by the signature extraction module, the Authorship Attribution (AA) module can use them 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 module takes as input an anonymous text t_u, the sets $s^k_{n,x}$ and $c^k_{n,x}$, and the parameters n, x and k. It 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 his 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 such as Euclidian distance, Pearson correlation and cosine similarity. In this work, the Pearson correlation is chosen as it provided better results in initial experiments.

4 EXPERIMENTAL EVALUATION

A set of experiments is performed to assess the effectiveness of the proposed approach for authorship attribution based on the usage of sequential patterns of parts of speech having various lengths.

In these experiments, a corpus was created, inspired by the one used by Clark and Hannon, to evaluate their proposed synonym-based authorship attribution method (Clark and Hannon, 2007). Similarly, in this work, a corpus was extracted from Project Gutenberg [2]. The corpus consists of a set of 10 contemporary English novelists of the XIX century. For each author, we selected novels and books and we discarded other kinds like poems and dictionaries where authors follow specific set of rules. The resulting corpus consists of 30 books written by 10 different authors, where every author has exactly 3 books. The corpus has a total of 2,615,856 words and books have 3,330 sentences on average. Detailed statistics for each author are presented in Table 3.

Each text was preprocessed using the Preprocessor Module. Then, to assess the performance of the proposed approach, leave-One-Out-Cross-Validation (LOOCV) was used. Thus, for each text, the designed system was trained using the 29 other texts. The common part-of-speech patterns of the 29 other texts were created and used to create the signatures of the 10 authors using the Signature Extraction Module. The testing consisted of comparing the signatures of the remaining text with the 10 author signatures to rank the authors from the most likely author to the least

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

Table 3: Corpus Statistics.

Authors	Total words	Total Sentences
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
Thornton W Burgess	55,916	2,950
Totals	$2,615,856$	99,903

likely author. This whole process was performed for the 30 texts. A variety of text sizes for learning sets and testing sets were examined. Holding out 20% of sentences for the testing set gave the best results.

4.1 Influence of Parameters n, x and k on Overall Results

Recall that our proposed approach takes three parameters as input, i.e. the minimum and maximum length of POS patterns n and x, 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, 250, 500, and 1000. For each value of k, the length of the POS patterns was varied from $n = 1$ to $x = 5$. For each combination of parameters, we measured the *success ratio*, defined as the number of correct predictions divided by the number of predictions.

Tables 4 and 5 respectively show the results obtained for $k = 50$ and 1000, for various values of n and x. Furthermore, in these tables, the results are also presented by ranks. The row 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 (success ratio). 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.

From these results, we can make several observations. First, the best overall results are achieved by $n = 1, x = 3$, and $k = 50$. For these parameters, the author of an anonymous text is correctly identified 70% of the time, 90% as one of the two most likely authors and 93% as one of the three most likely authors.

Second, it is interesting to observe that increasing the number of patterns 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 $k = 50$ and still characterize well the writing style of authors. This is in contrast with previous works that have used a large amount of ngrams. For example, Argamon et al. have suggested computing the frequencies of 685 trigrams (Argamon-Engelson et al., 1998) and Sidorov et al. computed the frequencies of 400 to 11,000 ngrams/sngrams.

Third, it can be observed that increasing the maximum length x of the POS patterns generally improves the success ratio, although there are a few exceptions.

Table 4: Top-K, for k=50.

Success ratio in %					
n,x	$1,1$	$1,2$	$1,3$	$1,4$	$1,5$
R_1	40.0	73.3	70.0	63.3	63.3
R_2	76.6	83.3	90.0	83.3	83.3
R_3	83.4	90.0	93.3	86.6	86.6
R_4	90.1	93.3	96.6	96.6	96.6
R_5	96.8	96.6	96.6	96.6	96.6
R_6	96.8	96.6	96.6	96.60	96.6
R_7	100.0	100.0	100.0	100.0	100.0
R_8	100.0	100.0	100.0	100.0	100.0
R_9	100.0	100.0	100.0	100.0	100.0
R_{10}	100.0	100.0	100.0	100.0	100.0

Table 5: Top-K, for k=1000.

Success ratio in %					
n,x	$1,1$	$1,2$	$1,3$	$1,4$	$1,5$
R_1	40.0	60.0	66.7	66.7	66.7
R_2	76.7	73.3	76.7	83.4	86.7
R_3	83.4	86.6	90.0	90.1	90.0
R_4	90.1	93.3	90.0	93.4	96.7
R_5	96.8	96.6	93.3	93.4	96.7
R_6	96.8	96.6	93.3	96.7	96.7
R_7	100.0	96.6	96.6	100.0	100.0
R_8	100.0	96.6	100.0	100.0	100.0
R_9	100.0	96.6	100.0	100.0	100.0
R_{10}	100.0	100.0	100.0	100.0	100.0

4.2 Influence of Parameters n, x and k on AA for each Author

The previous subsection assessed the influence of parameters n, x and k on the ranking of authors for all anonymous texts. This section analyzes the results for each author separately. Recall that each author has three texts in the corpus. Table 6 and 7 show the number of texts correctly attributed to each author (R_1).

It can be observed that for most authors, at least two out of three texts are correctly attributed. For example, for $n = 1$, $x = 3$ and $k = 50$, five authors have two texts correctly classified, three have all texts correctly identified, and two have only one.

Furthermore, it is interesting that some authors are harder to classify. For instance, the proposed approach never identifies more than one 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 its authorship was attributed to Addams after his death. A plausible explanation for the difficulty of identifying the author of this text is that Addams may have attempted to hide his writings to preserve his anonymity. For the second text, the likely explanation is that it is an autobiography rather than a fiction novel, and thus the writing style may be different. As a result, the signature of Addams may be less coherent, which may result in low success for authorship attribution of his texts.

There is also some authors that were easily identified. For example, all texts by Jacob Abbott are correctly classified for all tested parameter values. The reason is that Jacob Abbott has a more distinctive writing style in terms of part-of-speech patterns.

Table 6: Top-K, for k=50.

Success ratio per author					
Authors	$1,1$	$1,2$	$1,3$	$1,4$	$1,5$
Catharine Traill	0/3	2/3	2/3	2/3	2/3
Emerson Hough	1/3	2/3	1/3	1/3	1/3
Henry Addams	0/3	1/3	1/3	1/3	1/3
Herman Melville	0/3	2/3	2/3	1/3	1/3
Jacob Abbott	3/3	3/3	3/3	3/3	3/3
Louisa May Alcott	2/3	2/3	2/3	2/3	2/3
Lydia Maria Child	1/3	2/3	2/3	1/3	1/3
Margaret Fuller	2/3	3/3	3/3	3/3	3/3
Stephen Crane	0/3	2/3	2/3	2/3	2/3
Thornton W Burgess	3/3	3/3	3/3	3/3	3/3

Table 7: Top-K, for k=1000.

Success ratio per author					
Authors	$1,1$	$1,2$	$1,3$	$1,4$	$1,5$
Catharine Traill	0/3	2/3	0/3	0/3	0/3
Emerson Hough	1/3	2/3	2/3	2/3	2/3
Henry Addams	0/3	1/3	1/3	1/3	1/3
Herman Melville	0/3	0/3	2/3	2/3	2/3
Jacob Abbott	3/3	3/3	3/3	3/3	3/3
Louisa May Alcott	2/3	2/3	2/3	2/3	2/3
Lydia Maria Child	1/3	2/3	2/3	2/3	2/3
Margaret Fuller	2/3	2/3	3/3	3/3	3/3
Stephen Crane	0/3	1/3	2/3	2/3	2/3
Thornton W Burgess	3/3	3/3	3/3	3/3	3/3

Table 8: Bi-grams and tri-grams top-K, for k=50.

Success ratio in %		
n,x	$2,2$	$3,3$
R_1	70.0	53.3
R_2	73.3	73.3
R_3	93.3	80.0
R_4	93.3	90.0
R_5	96.6	96.7
R_6	96.6	96.7
R_7	100.0	96.7
R_8	100.0	100.00
R_9	100.0	100.0
R_{10}	100.0	100.0

Table 9: Bi-grams and tri-grams top-K, for k=1000.

Success ratio in %		
n,x	$2,2$	$3,3$
R_1	60.0	63.3
R_2	83.3	80.0
R_3	90..0	83.3
R_4	90.0	86.6
R_5	93.3	89.9
R_6	96.6	93.2
R_7	96.6	96.5
R_8	96.6	96.6
R_9	96.6	96.5
R_{10}	100.0	100.0

4.3 Influence of using Patterns of Variable Length vs Fixed Length

The proposed approach uses POS patterns of variable length (between n and x). However, previous studies for authorship attribution using ngrams of words or POS mostly focused on part-of-speech sequences of fixed lengths such as bigrams and trigrams (Argamon-Engelson et al., 1998; Koppel and Schler, 2003). In an effort to facilitate comparison with these previous works, this section presents results obtained with the proposed approach using only bigrams ($n = x = 2$) and trigrams ($n = x = 3$). Tables 8 and 9 show the results for $k = 50$ and 1000.

The best results with fixed length patterns are obtained for bigrams, and $k = 50$. For these parameters, the author of an anonymous text is correctly identified 70% of the time, 73% as one of the two most likely authors and 93% as one of the three most likely authors. This remains less favorable than the best results using variable-length patterns, where the best obtained results were 70%, 90% and 93%.

5 CONCLUSIONS

This paper explored the possibility of using the top-k part-of-speech sequential patterns of variable length as a feature for authorship attribution. The proposed approach discovers sequential patterns of parts of speech to build signatures representing each author's writing style. It then uses them to perform automatic authorship attribution. An experimental evaluation using 30 books and 10 authors from Project Gutenberg was carried. Results show that authors can be accurately classified with more than 70% accuracy using a very small number of variable-length patterns (e.g. $k = 50$). The proposed approach was also shown to perform better using a small amount of variable-length patterns than with many fixed-length patterns such as POS bigrams and trigrams. Our future work experiments with blog texts, which have a very different general style.

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.

REFERENCES

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Proc. of the Eleventh Intern. Conf. on Data Engineering, 1995.*, pages 3–14. IEEE.

Argamon-Engelson, S., Koppel, M., and Avneri, G. (1998). Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, pages 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.

Boukhaled, M. A. and Ganascia, J.-G. (2015). Using function words for authorship attribution: Bag-of-words vs. sequential rules. *Natural Language Processing and Cognitive Science: Proc. 2014*, page 115.

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

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

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proc. of the 20th international conference on Computational Linguistics*, page 611. ACL.

García-Hernández, R. A., Martínez-Trinidad, J. F., and Carrasco-Ochoa, J. A. (2010). Finding maximal sequential patterns in text document collections and single documents. *Informatica*, 34(1).

Han, J., Kamber, M., and Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.

Howe, D. C. (2009). Rita: creativity support for computational literature. In *Proc. of the 7th ACM conference on Creativity and cognition*, pages 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, page 72.

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

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(2):1261–1276.

Litvinova, T., Seredin, P., and Litvinova, O. (2015). Using part-of-speech sequences frequencies in a text to predict author personality: a corpus study. *Indian Journal of Science and Technology*, 8(S9):93–97.

McDonald, A. W., Afroz, S., Caliskan, A., Stolerman, A., and Greenstadt, R. (2012). Use fewer instances of the letter i: Toward writing style anonymization. In *Privacy Enhancing Technologies*, pages 299–318.

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

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

Mwamikazi, E., Fournier-Viger, P., Moghrabi, C., and Baudouin, R. (2014). A dynamic questionnaire to further reduce questions in learning style assessment. In *Artificial Intelligence Applications and Innovations*, pages 224–235. Springer.

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.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of the 2003 Conf. of the North American Chapter of the ACL on Human Language Technology-Vol. 1*, pages 173–180. ACL.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3-4):363–390.