SUDMAD: Sequential and Unsupervised Decomposition of a Multi-Author Document Based on a Hidden Markov Model

Khaled Aldebei
Global Big Data Technologies Centre
University of Technology Sydney, Australia
Fujian Provincial Key Laboratory of
Information Processing and Intelligent Control
Minjiang University, Fuzhou, Fujian, 350121, China
Khaled.Aldebei@uts.edu.au

Xiangjian He (Corresponding author & equal first author)
Global Big Data Technologies Centre
University of Technology Sydney, Australia
School of Software and and Microelectronics
Northwestern Polytechnical University, China
Xiangjian.He@uts.edu.au

Wenjing Jia
Global Big Data Technologies Centre
University of Technology Sydney, Australia
Wenjing.Jia@uts.edu.au

Weichang Yeh
Department of Industrial Engineering and Engineering Management
National Tsing Hua University, Taiwan
Wcyeh@ie.nthu.edu.tw

Abstract

Decomposing a document written by more than one author into sentences based on authorship is of great significance due to the increasing demand for plagiarism detection, forensic analysis, civil law (i.e., disputed copyright issues) and intelligence issues that involve disputed anonymous documents. Among existing studies for document decomposition some were limited by specific languages, according to topics or restricted to a document of two authors, and their accuracies have big rooms for improvement. In this paper, we consider the contextual correlation hidden among sentences and propose an algorithm for Sequential and Unsupervised Decomposition of a Multi-Author Document (SUDMAD) written in any language disregarding to topics, through the construction of a Hidden Markov Model (HMM) reflecting authors' writing styles. To build and learn such a model, an unsupervised, statistical approach is first proposed to estimate the initial values of HMM parameters of a preliminary model, which does not require the availability of any information of authors or document's context other than how many authors have contributed to writing the document. To further boost the performance of this approach, a boosted HMM learning procedure is proposed next, where the initial classification results are used to create labelled training data to learn a more accurate HMM. Moreover, the contextual relationship among sentences is further utilized to refine the classification results. Our proposed approach is empirically evaluated on three benchmark datasets which are widely used for authorship analysis of documents. Comparisons with recent state-the-art approaches are also presented to demonstrate the significance of our new ideas and the superior performance of our approach.

Introduction

Authorship analysis is the process of analysing the authors of a disputed anonymous document, which uses a statistical study of linguistic called stylometry (Baayen et al., 2002) to identify the background of authors of the questioned text document. The task of authorship analysis is considered as a very old research topic. The first endeavor for identifying the writing style of a text document was in the 19th century with the study of Mendenhall (1887) on the Shakespeare's plays. Several studies in the 20th century have also focused on analysing a text document by exploiting measurements of some stylometric features in order to determine the author's writing style of the document (Zipf, 1932).

In recent years, authorship analysis has received increasing attention and been considered as an important problem in many fields including information retrieval and computational linguistics. This importance springs from the fact that the large amount of disputed information of documents on Internet needs to be analysed and investigated.

Many different scenarios have been considered for studying the authorship analysis. For example, the work of Koppel & Winter (2014) focused on the authorship verification problem, also called similarity detection problem. It aimed to determine whether two documents were written by the same author without the attention of the real author. In this case, there is no need to have a set of candidate authors. Another important scenario, which has been studied extensively in the last few years, is the authorship attribution (Stamatatos, 2009; Savoy, 2016). The idea is that, given text samples of a number of candidate authors, we are required to determine which of them is the real author of a given disputed text document.

In this article, we address another intriguing application scenario, which is also related to the authorship analysis, called "multi-author document decomposition". The trajectory of this scenario is to decompose a document written by more than one author into components each written by only one author. Although this problem is very important because of applications in plagiarism detection (Stamatatos, 2011), forensic analysis (Orebaugh et al., 2014), civil law (i.e., disputed copyright issues) (Grant, 2007) and intelligence issues (Layton et al., 2010), studies on this area have been extremely limited so far. The work in Koppel et al. (2011) has considered a new unsupervised approach for decomposing a multi-author document into authorial parts. They created artificially merged documents by using only one dataset containing 5 biblical books, which were written in Hebrew by 5 authors. However, this approach is limited to a specific type of documents only (i.e., Hebrew language documents), and it has been tested using only documents formed by two authors. Akiva & Koppel (2012) presented an unsupervised approach for identifying distinct authorial components of a multi-author document. Unlike the approach described in Koppel et al. (2011), this approach has been tested on documents written by 2, 3 and 4 authors respectively, and also it is a language-independent approach. However, the overall accuracy of this approach is not high enough. One year later, this approach was further improved in Akiva & Koppel (2013) by taking advantages of distance-based methods. However, when the number of authors increased to more than 2, the accuracy degraded significantly. For the same purpose, the approach was examined in Giannella (2015) and an improved approach called BayesAD was proposed, where the number of authors of the document can be either known or unknown. However, only documents with very few turns among authors were tested in the work, and its performance heavily relied on the parameter setting. Recently, we presented a new approach

in Aldebei et al. (2015) for unsupervised multi-author document decomposition based on the Naive-Bayesian model. It requires the estimate of a threshold value in order to produce good results.

In Daks & Clark (2016), the authors have proposed an unsupervised approach for segmenting documents according to their authorships. However, they have assumed that each document has been written by only a single author.

Some researchers have focused on the task of text intrinsic plagiarism detection. The task, which has been directly addressed in PAN 2011 competition (Oberreuter et al., 2011; Kestemont et al., 2011; Rao et al., 2011), aims to determine whether a given suspicious document contains plagiarized text or not when no reference documents are provided. Furthermore, it detects plagiarized text in case that the document has a plagiarism. Most algorithms in intrinsic plagiarism detection attempt to detect plagiarized passages by analysing style changes within the document. Unlike the task of this article, in intrinsic plagiarism detection, usually most sentences of the document are written by one author (i.e., the main author) with limited percentage of the document written by other authors of which the number is not known. Whereas in the task that our work targets, each author has written long successive sentences in a document.

Some other researchers, such as Brooke et al. (2012), have presented a model for automatically segmenting a stylistically inconsistent text, i.e., identifying the points in a "multi-personal" poem The Waste Land (1922) by T. S. Eliot, where the style changes. The work in Brooke et al. (2013) has also considered an unsupervised approach to distinguish voices in the same poem.

Typically, classical learning models are considered for constructing a classifier that can accurately predict the labels of new data given some training data. The main assumption made with regard to these models is that the data are independently and identically distributed (iid) from an unknown probability distribution. In our work, instead of assuming that the data are iid, we propose a novel idea to make use of the sequence of the data, i.e., the contextual relationship between the sentences. These sequences provide valuable sequential correlations. Sequential patterns are of great practical importance for many computational linguistic applications (Bishop, 2006), where they have been employed to enhance the prediction accuracy of classifiers.

In our work, we propose to segment a multi-author document into components according to authorship. We consider the contextual information hidden among series of sentences and propose to use the Hidden Markov Model (HMM) to explore the sequential patterns in the document. Note that, the initial idea of this work has recently been published in the ACL conference (Aldebei, He, Jia, & Yang, 2016). A simple HMM was constructed to find a useful sequential correlation between consecutive sentences of the document, which has achieved very encouraging results. In this paper, we further extend our work and propose a two-stage HMM model in order to utilize the sequential patterns among sentences more comprehensively. Apart from more details and experiments that are included to disclose the benefit of this idea, this paper distinguishes from our previous work with (Aldebei, He, Jia, & Yang, 2016) significantly in the following three new contributions:

- We propose to utilize the useful sequential correlations among the consecutive sentences in order to determine the authorial components and construct a two-stage Hidden Markov Model, called "SequentialUD" - Sequential Unsupervised Decomposition, to model the relationships between authorships and sentences.
- To further boost the performance of this approach, a boosted HMM learning procedure is proposed. The initial classification results obtained using the statistically learned and preliminary HMM are used to create a labelled training dataset to learn a more accurate HMM.
- Moreover, the contextual relationships among sentences are further utilized to refine the classification results and a refined version of the SequentialUD is proposed.

In summary, the new approach proposed in this paper further exhibits the benefits of exploring the sequential patterns of sentences for analysing document's authorships. This approach is completely unsupervised and does not require the availability of any information of authors or document's context. It is effective even when the topics in the document are not distinguishable among authors. When the number of authors increases, the performance of this approach is still very satisfactory. To the best of our knowledge, there have been no similar ideas reported in the literature.

The following section presents the framework of our proposed SequentialUD approach. The detailed procedure of estimating the initial parameters and learning the preliminary HMM using our proposed statistical approach are first given. The preliminary HMM is then used for the initial sentence decoding. In the section of "Learning the Boosted HMM", the predicted labels are then used to create a labeled dataset from the unlabelled input, which is used to learn the final, boosted HMM. Eventually, sentence classification results are produced. A refinement procedure based on a modified probability indication procedure is proposed to further improve the purity, detailed in Section "Refinement with ModPIP". Then, the experiments are presented with conclusions given in the end.

Framework of the Proposed SequentialUD Approach

The problem of multi-author document decomposition can be more formally presented as follows. Suppose that there are N (a known number greater than 1) authors who have participated in creating a document C, each author has written long successive sentences in the document and each sentence is written by only one author. The goal is to decompose the sentences in the document into components according to their authorship, so that all sentences in a component are written by only one author.

The framework of the proposed approach is shown in Figure 1. The modules enclosed by dashed lines represent the two stages of the proposed SequentialUD approach, i.e., Estimating the Preliminary HMM, and Learning the Boosted HMM. Optionally, the classification results can be refined to further improve its purity by performing ModPIP, resulting in a refined version of the SequentialUD approach.

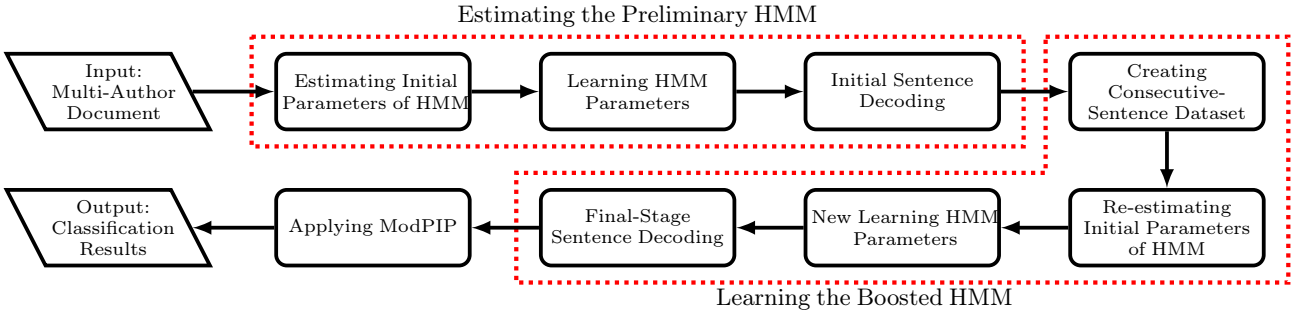

FIG. 1: The framework of the proposed SequentialUD and its refined version.

As seen in Figure 1, our proposed SequentialUD approach has two main stages. In the first stage, given unlabelled input data, we first propose a statistical approach to estimate the initial parameters of a preliminary HMM, which enables the Baum-Welch Algorithm to learn the preliminary HMM. Once the preliminary HMM is learned, it is used to estimate the best sequence of authors for sentences in document C using the Viterbi Algorithm. With these initial prediction results, the approach then proceeds into Stage 2, where the problem now becomes a supervised learning problem to learn a boosted HMM. The predicted labels resulted from the first stage are now used to create a new, labelled training dataset, which is then used to learn a more accurate HMM. In the end, the Viterbi Algorithm is used again to find a more accurate sequence of authors for the sequence of all sentences of document C. As an optional step, the classification results can be further refined by taking use of the contextual information.

Estimating a Preliminary HMM from Unlabelled Input Data

To make use of the contextual information for document decomposition, we utilize the Hidden Markov Model (HMM), a widely-used effective technique for sequential learning models, and take benefit from the powerful HMM tools to improve the classification purity result. In this section, for the integrity of this paper, we first briefly introduce the HMM. Then, we focus on how we formulate our document decomposition problem into the HMM and address the parameter initialization problem with no labelled data.

Hidden Markov Model

HMM is a statistical probabilistic model for sequential data consisting of a sequence of observable data and a hidden variable, which is not directly observable, for each observed data. The observable data are called "observations" and the hidden variables are called "hidden states". The hidden states in HMM form a Markov chain and the probability distribution of the observation depends on the underlying state.

Let us denote the T observations as $O = \{o_1, o_2, \ldots, o_T\}$ and the hidden states as $Q = \{q_1, q_2, \ldots, q_T\}$, where q_t is the hidden state of the t^{th} observation o_t. Each observation, which is assumed to be a discrete symbol, has one of the possible values from the set of observations $W = \{w_1, w_2, \ldots, w_M\}$ and each hidden state has one of the values from the set of states $S = \{s_1, s_2, \ldots, s_N\}$. Here, M and N represent the number of distinct observations and the number of distinct states in the model, respectively. Figure 2 illustrates the graphical structure of the HMM.

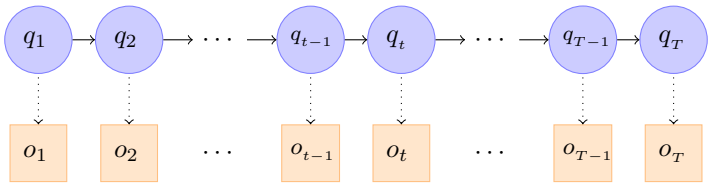

FIG. 2: A graphical model of the HMM with T hidden states $(q_1, \ldots, q_T)$ and T observations $(o_1, \ldots, o_T)$.

The conditional probability $p(q_t|q_{t-1})$ is called the "transition probability". The transition probabilities of all possible state values can be formed in an $N \times N$ transition matrix, denoted by $\boldsymbol{A}$. Each probability is given by $A_{ij} = p(q_t = s_j|q_{t-1} = s_i)$, where s_i, $s_j \in S$, $0 \leq A_{ij} \leq 1$ and $\sum_j A_{ij} = 1$.

The initial state q_1 is defined as a marginal distribution $p(q_1)$. All initial states are represented by a $1 \times N$ vector, denoted by $\boldsymbol{\pi}$. Each probability is given by $\pi(i) = p(q_1 = s_i)$, where $s_i \in S$ and $\sum_i \pi(i) = 1$.

The conditional probability $p(o_t|q_t)$ is called the "emission probability". In this article, because we assume that the observations are discrete symbols where each observation has one of the M possible values, the emission probabilities of all observations given their states are formed in an $N \times M$ emission matrix, denoted by $\boldsymbol{B}$. Each conditional probability is given by $b_i(k) = p(o_t = w_k|q_t = s_i)$, where $w_k \in W$, $s_i \in S$.

An HMM can be defined by the above three probabilities, denoted as $\boldsymbol{\theta}$, with $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$, for brevity.

In this article, in order to prevent computational underflow, all probabilities are computed as logarithms.

Estimating Initial Parameters of HMM

We consider HMM for document decomposition problem, where each observation represents one sentence and the hidden states represent the authors of the document. The goal is to decompose the document based on the writing style which is determined by the hidden state, i.e., authorship. The size of the observation and hidden state sequence, denoted by T, is the number of sentences in the document. In our model, due to that the number of distinct observations is not clearly observable, and the chance of having more than one sentence with the same syntactic structure is very low, we consider the number of unique observations (i.e., M) is also equal to the number of sentences in the document (i.e., T). Specifically, $T = M = |C|$ where $|C|$ is the number of sentences in document C. The number of unique states is equal to the number of authors of the document, which is denoted by N. The purpose of this model is to find the most probable sequence of authors that could have generated a given series of sentences in a document.

As illustrated in the previous subsection, an HMM can be specified by three parameters, $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$. We learn this model by maximizing the likelihood function of HMM in order to find a best estimation of $\boldsymbol{\theta}$ so that the probability of the observations maximizes, as in

$$\theta = \arg\max_{\theta} \left(p(O|\theta)\right) \tag{1}$$

Normally, the learning process starts with some initial values of $\boldsymbol{\theta}$. For unsupervised learning problems (like the one we are dealing with), the initial values of $\boldsymbol{\theta}$ are not directly observed and therefore need to be manually set. The selection of $\boldsymbol{\theta}$ has a significant impact on the overall efficiency of the model as it directly affects the convergence rate of the learning process, as well as whether the learning process can converge on global maximum (Hoang & Hu, 2004).

In our work, we propose a statistical approach and make use of the contextual information of sequential data to initiate the HMM parameter set $\boldsymbol{\theta}$. Next, we give the details of initializing these parameters in the order of transition matrix $\boldsymbol{A}$, prior $\boldsymbol{\pi}$, and emission probability $\boldsymbol{B}$. These are detailed as follows.

Estimating Transition Matrix $\boldsymbol{A}$

1. We first create a sequence of *segments*, where each segment is a series of v successive sentences from the document and does not overlap with any other segments. Intuitively, the segment length v relates to the length of the document, as well as the mean Author Run Length (simplified as meanARL), which represents the mean number of successive sentences in the document written by the same author. In the section of Experiments, detailed analysis is provided to find the most appropriate value of v for a given document. We then collect the statistic of the segments. Note that, working on segments instead of sentences allows us to capture the sequential patterns of sentences. Formally, let us denote the series of segments as $SEG = \{Seg_1, Seg_2, \ldots, Seg_e\}$. For a

document of size $|C|$, this produces e segments, where $e = Ceiling(|C| \,/\, v)$. Notice that, each segment may be either a pure segment, where its sentences are produced by a single author, or a mixed segment, where its sentences are produced by more than one author.

2. For each segment, we then extract a feature vector based on the concept of "Bag of Words". To do this, first a word list is created for the document, where distinct words (i.e., the words occurred three or more times in the document) are added into a word list, denoted by $BagOfWords1 = \{word_1, word_2, \ldots, word_{D_1}\}$, where D_1 is the length of the list (i.e., the total number of the words in the list). In this paper, a word is defined as a consecutive sequence of letters and digits. Then, each segment is represented as a D_1-dimension binary vector using the word list $BagOfWords1$, where each dimension takes a value of 1 or 0, with 1 indicating the corresponding word in the list appears in the segment and 0 indicating not. Thus, the segments SEG can be represented as a sequence of e D_1-dimension binary feature vectors, denoted by $X = \{x_i, i = 1, 2, \ldots, e\}$. More details can be found in the Experiments section.

3. With the binary feature vectors X, we then cluster them into different groups, each representing a unique writing style. The Gaussian Mixture Model (GMM) (McLachlan & Basford, 1988) is adopted for clustering after comparing with classical clustering methods such as K-means. Since there are N authors who have contributed writing the document C, the GMMs have N Gaussian components, each representing a different author's writing style. Each vector x_i, $i \in 1, 2, \ldots, e$, is clustered into one of the N Gaussian components.

4. Based on the Gaussian component that a vector x_i is assigned to during the above clustering process, each vector x_i is given a label. Apparently, the label of vector x_i, denoted by $h(x_i)$, takes one label from a set of N elements, i.e., $h(x_i) = n$, where $n \in \{1, 2, \ldots, N\}$.

   Note. The approaches of Akiva & Koppel (2012), Akiva & Koppel (2013) and Aldebei et al. (2015) also start from segmenting the original document into segments and then represent them as feature vectors in order to cluster them. However, the purpose of these steps in their approaches is different from the purpose in the proposed approach.

5. Then, with the labels $h(x_i)$, $i = 1, 2, \ldots, e$ of all the segments, the transition probability of moving from state n_1 to state n_2, denoted by $A_{n_1 n_2}$, can be computed using Eq. 2 as

$$A_{n_1 n_2} = \frac{Count\ (h(x_i) = n_2, h(x_{i-1}) = n_1) + 1}{Count\ (h(x_{i-1}) = n_1) + N}, \quad i = 2, \ldots, e, \tag{2}$$

   where $n_1, n_2 \in \{1, 2, \ldots, N\}$.

   Finding the transition probabilities of all possible state values (i.e., $n_1 = \{1, 2, \ldots, N\}$, $n_2 = \{1, 2, \ldots, N\}$) will produce the $N \times N$ transition matrix $\boldsymbol{A}$. Here, we employ the "add-1" smoothing technique (Manning & Schütze, 1999) in order to prevent zero values of transition probabilities.

Estimating the Prior $\boldsymbol{\pi}$

We then move on to estimate the initial probability $\pi(n)$, i.e., the prior probability of each author. With each segment $(x_i, i \in \{1, 2, \ldots, e\}$, where e is the number of feature vectors) being labelled as $h(x_i)$, the initial probability of each state, denoted by $\pi(n)$, can be simply measured as a fraction of the occurrences of each state $h(x)$ as: $\pi(n) = Count\ (h(x) = n) \,/\, e$, where $n \in \{1, 2, \ldots, N\}$.

Finding the initial probabilities of all possible state values (i.e., $n = \{1, 2, \ldots, N\}$) will produce a $1 \times N$ vector, which is denoted by $\boldsymbol{\pi}$.

Estimating the Emission Probabilities $\boldsymbol{B}$

The emission probabilities $\boldsymbol{B}$ address the relation between observations and states, i.e., given the authorship ("state"), the probability of observing each sentence ("observation").

1. The sequence of segments SEG, each consisting v successive sentences, is employed in order to find the initial value of $\boldsymbol{B}$. In order to do that, a new feature list, denoted as $BagOfWords2 = \{word_1, word_2, \ldots, word_{D_2}\}$, where D_2 is the length of the list, is created. The words that have occurred at least two times in the document are considered for this feature. The list of words are used for representing the sequence of segments (SEG) as a sequence of binary feature vectors, $X' = \{x'_i, i = 1, 2, \ldots, e\}$. Each vector has D_2 elements. Note that each feature in the vector represents one word of the $BagOfWords2$ list.

The process of creating the sequence X' is similar to the process of creating the sequence X. The key difference is that we use the $BagOfWords2$ list of D_2 features instead of the $BagOfWords1$ of D_1 features. Note that, including words that have occurred for at least two times instead of three times into the word list, allows better chance to have more listed words appear in a sentence, which contains much less words than a segment does.

Each vector x'_i takes the same label of vector x_i, i.e., $h(x'_i) = h(x_i) = n$, where $i = 1, 2, \ldots, e.$ and $n \in \{1, 2, \ldots, N\}$.

2. Given the sequence of feature vectors, X', and the set of all possible values of labels, the probability of each feature in X' given a label n ($n \in \{1, 2, \ldots, N\}$) is computed using the conditional probability shown in Eq. 3:

$$p(j|n) = \frac{Count_j^n + 1}{Count^n + D_2}, \quad j = 1, 2, \ldots, D_2. \tag{3}$$

where j represents a feature, $Count_j^n$ represents the count of observed feature j in the vectors that have a label n, $Count^n$ represents the count of all observed features in the vectors that have a label n, and D_2 is the number of features.

Note that we again employ the "add-1" smoothing technique in Eq. 3 in order to prevent a zero probability.

3. Each sentence of document C is represented as a D_2-dimension binary feature vector using the word list $BagOfWords2$, where each dimension takes a value of 1 or 0, indicating the presence of the corresponding word in the sentence. Thus, the sentences can be represented as a sequence of T D_2-dimension binary feature vectors, denoted by $O = \{o_i, i = 1, 2, \ldots, T\}$.

Using Eq. 3, the computation of the conditional probability of each feature given each possible value of labels (i.e., $n = 1, 2, \ldots, N$) will lead us to compute the initial value of the emission probability of an observation given each state of the HMM, as shown in Eq. 4.

$$p(o|n) = \prod_{j=1}^{D_2} p(j|n)^{o^j}, \quad n = 1, 2, \ldots, N, \tag{4}$$

where o represents an observation, j represents a feature, o^j represents the value of feature j in observation o, and D_2 is the number of features.

The initial estimated probabilities of $\boldsymbol{\theta}$ will be used in the next subsection for learning the HMM in order to find a best estimation of $\boldsymbol{\theta}$.

Learning the Preliminary HMM

In this subsection, we work on the HMM to learn $\boldsymbol{\theta}$ (i.e., $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\pi}$) based on Eq. 1.

Formally, the HMM, which consists of a sequence of hidden states and independent observations as seen in Figure 2, is formed as follows:

Assume that there are T sentences in document C (remember $T = |C|$), denoted by $\{Sen_i, i = 1, 2, \ldots, T\}$, where i represents the position of a sentence in the document (for example Sen_1 and Sen_T denote the first sentence and last sentence of document C, respectively).

Each hidden state represents the most likely author of the corresponding sentences. Therefore, there are T hidden states, denoted by $Q = \{q_1, q_2, \ldots, q_T\}$. Each state takes only one possible value from a set denoted by $S = \{1, 2, \ldots, N\}$. For generality, we substitute the set $S = \{1, 2, \ldots, N\}$ by a set $S = \{s_1, s_2, \ldots, s_N\}$.

The estimation of $\boldsymbol{\theta}$, which can explain the observations more effectively, is performed by using the Baum-Welch algorithm (Dempster et al., 1977), which is considered as a special case of the Expectation Maximization (EM) algorithm. The process starts with using the initial values of $\boldsymbol{\theta}$, which were estimated in the previous subsection, and computes the probabilities of being in each state at each time. This is done by using the forward-backward algorithm (Rabiner & Juang, 1986; Bishop, 2006). After that, the estimated probabilities are used to obtain a better estimate of $\boldsymbol{\theta}$. Using the improved (hopefully) $\boldsymbol{\theta}$, the forward-backward algorithm is applied again, and the cycle repeats until the convergence of either the $\boldsymbol{\theta}$ or the estimated probabilities occurs.

The learned $\boldsymbol{\theta}$ will be used in the next subsection in order to find the best sequence of authors that represents the sequence of sentences of document C.

Initial Sentence Decoding

In our problem, we are interested in finding the most likely sequence of states (i.e., authors) that generates the corresponding sequence of observations (i.e., sentences), as shown in Eq. 5.

$$Q^* = \arg\max_{Q} p(Q|O). \tag{5}$$

The Viterbi algorithm (Viterbi, 1967; Forney Jr, 1973), also known as max-sum algorithm, is used to find efficiently the most likely sequence of states for the given observations.

After all, by using the Viterbi algorithm, the best sequence of authors, $Q^* = \{q_1, q_2, \ldots, q_T\}$, that represents the corresponding sentences in document C is determined.

Learning the Boosted HMM

As we have mentioned earlier, the initial values of $\boldsymbol{\theta}$ have a significant impact on the learning process of HMM so that it affects the performance of the decoding process. For unlabelled data, we have proposed a statistical approach to better estimate the initial values of $\boldsymbol{\theta}$ by using segments and learned a preliminary HMM. The HMM has been used to classify each sentence. In this section, the resulted, labelled sentences obtained in the previous section can be used to re-calculate the initial values of $\boldsymbol{\theta}$, which can then be used to learn a more accurate, boosted HMM, to further improve the performance of the decoding.

Creating Consecutive-Sentence Dataset

A procedure, called "Consecutive-Sentence Dataset", is proposed to create a new labelled dataset which can be employed to re-estimate the initial values of $\boldsymbol{\theta}$ and re-construct the HMM. The procedure aims to provide a dataset with a high rate of correctly labelled data. It strives to provide a dataset with more labelled data for calculating $\boldsymbol{\theta}$, by using sentences rather than segments. This procedure works as follows. Given the labels of all of the sentences of document C, sequences of minimum five consecutive sentences having the same label are inserted into the new dataset for that label.

Our experiments have shown that the purity results are not sensitive to the setting of the minimum number of consecutive sentences as long as it does not exceed the mean author run length (i.e., meanARL) in the document.

Eventually, the new dataset, denoted by $CSD = \{(Sentence_1, q'_1), (Sentence_2, q'_2), \ldots, (Sentence_{T'}, q'_{T'})\}$ is created, where $q' = s$ with $s \in S$, and T' represents the number of sentences in CSD.

Re-Estimating and Learning the HMM parameters, and Final-Stage Sentence Decoding

We use the new dataset CSD to re-estimate the new initial values of $\boldsymbol{\theta}$. The computations of the initial values of $\boldsymbol{A}$ and $\boldsymbol{\pi}$ are similar to the computations which have been applied in the previous section, and we replace the set of all labels $h(x_i)$, $i = 1, 2, \ldots, e$ with the set of all states q'_i, $i = 1, 2, \ldots, T'$.

The initial values in $\boldsymbol{B}$ are also re-calculated using the new dataset. However, due to the fact that the new dataset is a sequence of sentences, rather than segments, it is desirable to increase the number of features used in representing the sentences to capture the relation between the observations (i.e., sentences) and the states (i.e., authors). Therefore, a new feature list, denoted by $BagOfWords3 = \{word_1, word_2, \ldots, word_{D_3}\}$, where D_3 is the length of the list, which contains all distinct words that occur at least one time in the document C, is created. By using this list, all sentences in CSD are represented as binary feature vectors, denoted by $X'' = \{x''_i, i = 1, 2, \ldots, T'\}$.

The probability computation of each feature in X'' given a label n ($n \in \{1, 2, \ldots, N\}$) is similar to the computation which has been applied in the previous section. The only difference is that we replace the sequence of vectors, X', of D_2 features by the sequence of vectors, X'', of D_3 features.

Then, the new initial values in $\boldsymbol{\theta}$ (i.e., $\boldsymbol{A}$, $\boldsymbol{\pi}$ and $\boldsymbol{B}$) are utilized for learning the HMM again. The process of learning the HMM is the same as the process discussed in the previous section. The only difference is that we replace the $BagOfWords2$ list of D_2 features by the $BagOfWords3$ list of D_3 features for representing the observation sequence, O.

Lastly, the final-stage sentence decoding process is applied in order to find the most likely sequence of authors corresponding to all sentences in the document C. Here, the same algorithm illustrated in the previous section is used to perform the decoding process of this step.

Thus far, the SequentialUD approach, which consists of seven steps shown in Figure 1, is done.

Refinement with ModPIP

The work in Aldebei et al. (2015) proposed a probability indication procedure (PIP) in order to enhance the purity of sentence classification process. The procedure consists of five criteria. It proceeds by selecting trusted sentences from a document and using them to re-classify each sentence of the document into the author's class. The procedure has been implemented using the Naive-Bayesian model.

Following this idea, in this work, a modified version of PIP, named by ModPIP, is proposed to refine the classification results and further improve the sentence classification purity results. Since we treat the sentences of a document as sequential data, the ModPIP is developed based on a sequential model. This is detailed as below.

1. A sentence in the document C, which has been assigned a specific state value, is recorded as a *trusted sentence* if and only if the posterior probability of its state value given the observed sequence of all sentences is greater than the posterior probabilities of all other state values given the observed sequence of all sentences, by more than a threshold R. The state values of the trusted sentences will be fixed.

2. If the first trusted sentence in the document C is not the first sentence in the document, then all sentences starting from the first sentence in the document till the sentence located before the trusted sentence are given the same state value of the trusted sentence.

3. If the last trusted sentence in the document C is not the last sentence in the document, then all sentences starting from the sentence located after the trusted sentence till the last sentence in the document are given the same state value of the trusted sentence.

4. If a group of non-trusted consecutive sentences is surrounded between two trusted sentences that have the same state value, then all the sentences in the group are given the state value of the two trusted sentences.

5. If a group of non-trusted consecutive sentences is surrounded between two trusted sentences that have different state values, then the best split point in the group is picked out in order to divide the group into two subgroups. All the sentences in the first subgroup, which comes before the split point, are given the same state value of the trusted sentence which comes before them. All the sentences in the second subgroup, which comes after the split point, are given the same state value of the trusted sentence which comes after them. The best separation point is the one that gives the maximum summation value of all posterior probabilities of the assigned state values of the sentences in the group given all observed sentences in the document.

The posterior probability of a single state given the observed sequence of all sentences, $p(q|O)$, which is used in the first and fifth criteria, is computed using the forward-backward algorithm.

Experiments

In this section, the performance of the proposed approach (i.e., SequentialUD and its refined version) is evaluated and compared with state-of-the-arts on three benchmark datasets used as benchmarks in Koppel et al. (2011), Akiva & Koppel (2012), Akiva & Koppel (2013), Aldebei et al. (2015) and Giannella (2015). Furthermore, to test its performance on more realistic cases, scientific articles are also utilized. In this article, the state-of-the-arts results used for comparisons are directly taken from their articles.

In this paper, experimental results are evaluated by using *Purity* (Zhao & Karypis, 2001; Manning et al., 2008; Amigó et al., 2009). The purity measure focuses on the frequency of the most common category in each class. Assuming that $L = \{L_1, L_2, \cdots, L_N\}$ is the set of classes to be evaluated, $U = \{U_1, U_2, \cdots, U_N\}$ is the set of categories, N is the number of classes (or categories) to be evaluated, and T is the number of observations, the purity is computed by taking a weighted average of maximal precision values, as shown in Eq. 6:

$$Purity = \sum_{i=1}^{N} \left(\frac{|L_i|}{T} \max_j P(L_i, U_j)\right) \times 100\%, \tag{6}$$

where P represents the precision of a class L_i for a given category U_j and is defined as:

$$P(L_i, U_j) = \frac{|L_i \cap U_j|}{|L_i|}. \tag{7}$$

Datasets

The first corpus tested is a group of five biblical books written by five authors (see Table 1). These books are related to two genres of literature, wisdom and prophetic. Note that, we adopted biblical books for three reasons. First, this corpus is highly motivated, since various researchers have been working on authorship analysis of biblical literature for centuries. Second, written in Hebrew language, this corpus gives an opportunity to test non-English documents. Third, because the five bible books are related to two literatures, it allows to evaluate the effectiveness of our approach in documents created by merging two books of the same literature.

TABLE 1: Statistics regarding the five Bible Books.

	Author Name	Chapter Numbers	Literature Genre	Number of Sentences
1	Proverbs (Prov)	1-31	Wisdom	915
2	Jeremiah (Jer)	1-52	Prophetic	1,364
3	Ezekiel (Eze)	1-48	Prophetic	1,273
4	Isaiah (Isa)	1-35	Prophetic	676
5	Job	3-41	Wisdom	1,018

The second corpus, referred to as "The Becker-Posner Blog", is a group of 690 blogs written by the Nobel Prize winning economist Gary Becker and the legal scholar and federal judge Richard Posner. The Becker-Posner Blog was started in 2004 to discuss current issues of law, economics and policy in a dialogic format. It provides a good basis for inspecting the performance of various approaches on documents where the topics among authors are not differentiated. The work in Giannella (2015) manually created six single-topic documents from the Becker-Posner blogs in order to evaluate the performance of his work (see Table 2), where each document has sentences representing only one single topic. In this work, we use these documents because each of these documents has only one single topic, all these documents are short and the total number of consecutive sentences of each author in these documents is relatively small. Therefore, this corpus makes the task of distinguishing the sentences in a document, according to authorship, rather than topics, be more challenging.

TABLE 2: Statistics of the six single-topic documents created from the Becker-Posner Blogs.

Topics	Author order and number of sentences per author
Traffic Congestion (TC)	Becker(57), Posner(33), Becker(20)
Senate Filibuster (SF)	Posner(39), Becker(26), Posner(28), Becker(24)
Microfinance (Mic)	Posner(51), Becker(37), Posner(44), Becker(33)
Tort Reform (TR)	Posner(29), Becker(31), Posner(24)
Profiling (Pro)	Becker(35), Posner(19), Becker(21)
Tenure (Ten)	Posner(73), Becker(36), Posner(33), Becker(19)

We test our approach on the third corpus, a group of 1,182 *New York Times* articles. These articles, having diverse topics, were written by four columnists (see Table 3). We use this corpus in order to evaluate the performance of the proposed approach on documents that are written by more than two authors (i.e., three or four authors).

TABLE 3: Statistics of the *New York Times* articles.

Columnist Name	Number of opinion articles	Number of sentences
Thomas Friedman (TF)	279	11,230
Maureeen Dowd (MD)	299	11,660
Paul Krugman (PK)	331	12,634
Gail Collins (GC)	273	11,327

Furthermore, in order to show the efficiency of the proposed approach on more realistic cases, we have randomly selected some scientific articles, which are cited in the References section, covering the same topics. The articles of each topic are mixed in one article and the proposed approach is then applied in order to recover the author of each sentence in the mixed article. Due to the difficulty of finding articles written by single authors and covering same topics, as well as due to the fact that in most cases, there is one main author of an article whose writing style can be found throughout of the article, we consider each article produced by more than one author as an article produced by only one author. In each selected article, we have ignored all metadata (e.g., titles, author names, references, equations, tables and citations). We randomly select two articles on plagiarism detection topic. The articles are Rao

et al. (2011) and Kestemont et al. (2011). The lengths of these articles are 66 and 111 sentences, respectively. We also randomly select three articles on authorship attribution topic. The articles are Baayen et al. (2002), Layton et al. (2010) and Savoy (2016). The lengths of these articles are 91, 197 and 304 sentences, respectively. Furthermore, we randomly select four articles on authorship-based document decomposition topic (i.e., the same topic addresses in this article). The articles are Koppel et al. (2011), Giannella (2015), Daks & Clark (2016) and Aldebei, He, & Yang (2016). The lengths of these articles are 257, 215,104 and 229 sentences, respectively. The four articles have also used the same data sets in their approaches. Note that, all articles of each topic are randomly selected, in which each article is produced by different authors.

To also work on other types of non-artificial document, we have tested the proposed approach on a very early draft of a scientific paper produced by two Ph.D students (Students A and B) in our research team. To use this document, we have ignored all the figures as well as all metadata (e.g., titles, author names, references and citations). The paper consists of 313 sentences and has 6 sections (including the Abstract and Conclusion). Each author has written 3 sections. Student A has written 131 sentences and Student B has written 182 sentences.

Experimental Results

The performance of the proposed approach (i.e., SequentialUD and its refined version) is examined through a set of experiments on different documents. In the first four experiments, artificially created documents are prepared by employing the same procedure used in Koppel et al. (2011), Akiva & Koppel (2012), Akiva & Koppel (2013) and Aldebei et al. (2015) for fairness of comparison, which is summarised as follows.

Suppose that there are N authors, each having written a group of documents. The document of N authors is composed by recursively picking up a random number (m) of unselected successive sentences from a document of a randomly chosen author and merging them together until all sentences in all documents of N authors are selected. During each iteration, the value of m is randomly chosen from a uniform distribution ranging from 1 to V. We also follow the approaches described in Koppel et al. (2011), Akiva & Koppel (2012), Akiva & Koppel (2013) and Aldebei et al. (2015) and assign 200 to V. In our experiments, we empirically assign 15 to the threshold R for the refined SequentialUD approach.

In order to determine the optimal segment length v, which is employed to estimate the transition matrix A of the preliminary HMM, we group documents from the datasets based on the number of their sentences into two categories, i.e., Long Documents (containing 500 or more sentences) and Short Documents (containing fewer than 500 sentences). Furthermore, based on our observations, the segment length v also depends on the meanARL. To depict the impact of document length and meanARL on v, for each category we randomly pick up one document and apply our SequentialUD approach on its sentences with different values of meanARL. We employ the same procedure described above and use the sentences of the document to create merged documents with different values of meanARL. The meanARL of a document is determined by setting the value of V, which results in a mean of around $0.5V$ successive sentences from the same author on the document. A document resulted from merging the biblical books of Ezekiel and Proverbs (containing 2188 sentences) and the two-student scientific document (containing 313 sentences) have been selected to determine the best segment length for Long and Short Documents, respectively. Our experiments on the Eze-Prov document (a Long Document) show that the proposed approach yields higher purity results when v is less than meanARL and 60. Recall that, in our work each author is assumed to have written long successive sentences in a document (i.e., a larger meanARL) and most Long Documents used in our experiments are created with a meanARL of around 100 (i.e., $V = 200$). Our experiments also show that the highest purity result in the Eze-Prov document with the meanARL of around 100 is achieved when v is 30. Therefore, we assign 30 to v for all Long Documents in our experiments. However, for the scientific document (a Short Document), our experiments show that the proposed approach achieves higher purity results when v is less than the meanARL and 40. Furthermore, our experiments show that the most highest purity results on the scientific document are achieved when v is 10. Therefore, we assign 10 to v for all Short Documents in our experiments. To ensure that the number of segments used in the clustering process (Step 3 of estimating transition matrix A) is always larger than the number of clusters, the segment length v for Short Document is set to $\min(10, F(\text{No. Sentences}/\text{No. Authors}) - 1)$, where F represents the commonly known floor function.

In our approach, in order to reduce the influence of topics on final results, only those words appearing at least three times in the document are used as features of $BagOfWords1$ to depict the writing style of the segments (see Step 2 of the subsection "Estimating Transition Matrix $\boldsymbol{A}$"). However, note that these words may not necessarily be purely topic-independent words. Based on our observation on different documents, the words selected into the feature set are mostly function words and words that are independent of topics. Increasing the frequency threshold does help to exclude these topic-specific words but at the meantime it also decreases the recall rates of pure segments on the clustering process (Step 3 of the same subsection), and this results in insufficient data being produced for the

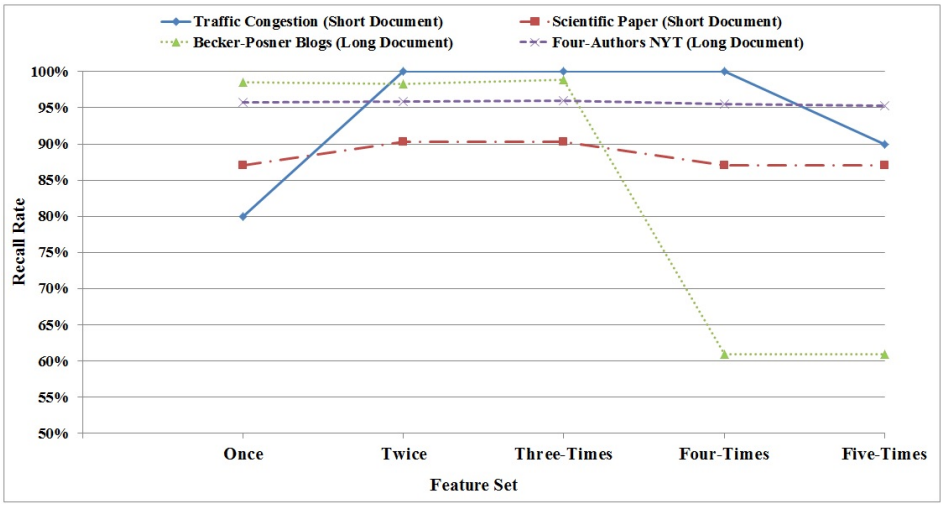

FIG. 3: The recall rates of the clustering process obtained using words that have occurred at least once, twice, three times, four times and five times in four documents as features of BagOfWords1.

estimation process. Also note that we use recall rates to evaluate clustering results, because our interest here is to evaluate the capability of the clustering process for retrieving pure segments. Furthermore, we observe that choosing words appearing for at least three times as features of $BagOfWords1$ in a Short or Long Document produces generally higher recall rates on the clustering process. Figure 3 shows the recall rates of the clustering process using words that have occurred at least once, twice, three times, four times and five times respectively in four documents as features of $BagOfWords1$. The documents of Becker-Posner Blogs and four-author columnists of New York Times articles have been used as Long Documents. The documents of Traffic Congestion and scientific paper have been used as Short Documents. Obviously, as shown in Figure 3, using all words that have occurred three or more times has achieved higher recall rates on the clustering process for all four documents.

Next, the results of the proposed approach SequentialUD and its refined version are compared with those obtained by five state-of-the-art approaches.

Results on the Biblical Books Dataset

For the first set of experiments, the five biblical books of five authors are utilized to produce a set of total 10 merged documents of two authors by using the procedure mentioned before. The 10 documents are related to either different genres or the same genre (see Table 1). In Tables 4 and 5, we report the purity results obtained by applying the proposed approach SequentialUD and its refined version on the documents composed by merging two biblical books of different genres (Table 4) and the documents composed by merging two biblical books of the same genre (Table 5), respectively. In both cases, the results of the SequentialUD and its refined version are compared with the approaches of Koppel et al. (2011), Akiva & Koppel (2013), Akiva & Koppel (2013)-SynonymSet and Aldebei et al. (2015). The purity results of the approach in Akiva & Koppel (2012) are used further for comparison in Table 5.

TABLE 4: Purity comparison on documents composed by merging two biblical books of *different genres*. Approaches in comparison: 1- Koppel et al. (2011), 2- Akiva & Koppel (2013), 3- Akiva & Koppel (2013)-SynonymSet, 4- Aldebei et al. (2015), 5- Our SequentialUD and 6- Our refined SequentialUD.

Doc.	**1**	**2**	**3**	**4**	**5**	**6**
Jer-Prov	72.7%	97.0%	75.0%	99.0%	99.6%	99.8%
Isa-Job	82.2%	98.7%	89.1%	98.7%	99.6%	99.6%
Eze-Prov	76.6%	98.7%	90.8%	97.9%	99.2%	99.4%
Isa-Prov	70.4%	95.0%	85.0%	97.9%	98.7%	99.2%
Eze-Job	85.9%	98.7%	95.0%	99.0%	99.7%	99.7%
Jer-Job	87.3%	98.0%	93.1%	97.8%	99.1%	99.2%
Overall	***79.2%***	***97.7%***	***88.0%***	***98.4%***	***99.3%***	***99.5%***

TABLE 5: Purity comparison on documents composed by merging two biblical books of the *same genre*. Approaches in comparison: 1- Koppel et al. (2011), 2- Akiva & Koppel (2012), 3- Akiva & Koppel (2013), 4- Akiva & Koppel (2013)-SynonymSet, 5- Aldebei et al. (2015), 6- Our SequentialUD and 7- Our refined SequentialUD.

Doc.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Job-Prov	84.5%	84.9%	93.9%	82.0%	95.2%	98.6%	99.2%
Jer-Eze	82.0%	87.6%	96.6%	95.9%	97.0%	97.7%	98.2%
Isa-Jer	71.8%	63.4%	66.7%	82.7%	71.0%	73.1%	73.3%
Isa-Eze	78.9%	76.0%	80.0%	88.0%	82.7%	83.6%	83.8%
Overall	***79.3%***	***78.0%***	***84.3%***	***87.2%***	***86.5%***	***88.3%***	***88.6%***

From the purity results presented in the tables, we can observe that the results obtained with our proposed approach SequentialUD and its refined version are quite promising with a purity of over 99.5% achieved on some documents. We can also see that the overall purities of our proposed approach are remarkably better than those obtained using other approaches. In some cases (e.g., for the Jer-Prov document mentioned in Table 4), SequentialUD produces a 37% larger purity result than Koppel et al. (2011) and 33% larger purity result than Akiva & Koppel (2013)-SynonymSet. Note that, two of them, i.e., Koppel et al., 2011 and Akiva and Koppel, 2013-SynonymSet, are specially developed for biblical books only, and not applicable for other documents.

Results on Becker-Posner Blogs Dataset (Controlling for Topic)

For the second set of experiments, we apply the proposed approach on the merged documents composed from the Becker-Posner blogs corpus.

On the first part of our experiments using this corpus, we work on a document created by merging all Becker blogs and Posner blogs. The merged document has 26,922 sentences and 246 turns between the two authors. It does not have any topic indication that can be used to differentiate between authors. As shown in Table 6, the purity results achieved by applying our proposed approach on this document are significantly higher.

In fact, it is important to know the effectiveness of applying the procedures of our SequentialUD approach, i.e., the preliminary HMM, the Boosted HMM and the ModPIP refinement. Table 6 shows the intermediary and final purity results achieved by applying our SequentialUD approach on Becker-Posner blogs. In this table, "4-First-Stage HMM" is the purity obtained by applying the first-state preliminary HMM, and "5-Our SequentialUD" is the purity obtained after further applying the Boosted HMM, and "6-Our Refined SequentialUD" is the result obtained after applying the ModPIP refinement in the end. From these results, it can be seen clearly that: 1) The purity achieved using our preliminary HMM is already very effective and has outperformed the other three approaches; 2) Our BoostedHMM and ModPIP refinement have further improved the purity results, each by 0.6%.

TABLE 6: Purity comparison on a document of Becker-Posner Blogs. Approaches compared: 1- Akiva & Koppel (2012), 2- Akiva & Koppel (2013), 3- Aldebei et al. (2015), 4- First-Stage HMM, 5-Our SequentialUD and 6- Our Refined SequentialUD.

Approaches	**1**	**2**	**3**	**4**	**5**	**6**
Purity Results	94.0%	94.9%	96.6%	96.7%	97.3%	97.9%

On the second part of our experiments regarding this corpus, the six single-topic documents (see Table 2) manually created by Giannella (2015) from the Becker-Posner blogs are used to test the performance of the proposed approach, where each document has sentences representing only one single topic. Figure 4 illustrates the purity results obtained using our proposed SequentialUD approach and its refined version, compared with that of the approach in Giannella (2015). As shown in the figure, both versions of our approach have yielded better purity results (up to 42.5% in the "Traffic Congestion (TC)" document) in all six documents.

Results on New York Times Articles Dataset ($N \geq 2$)

In these experiments, we employ the *New York Times* articles of four columnists to create a set of merged documents, and the merged documents have two, three or four authors.

In the first set of experiments regarding this corpus, all possible documents of two authors are composed and six documents are produced. We find that the purity results of the six documents range from 94.1% to 97.0% and from

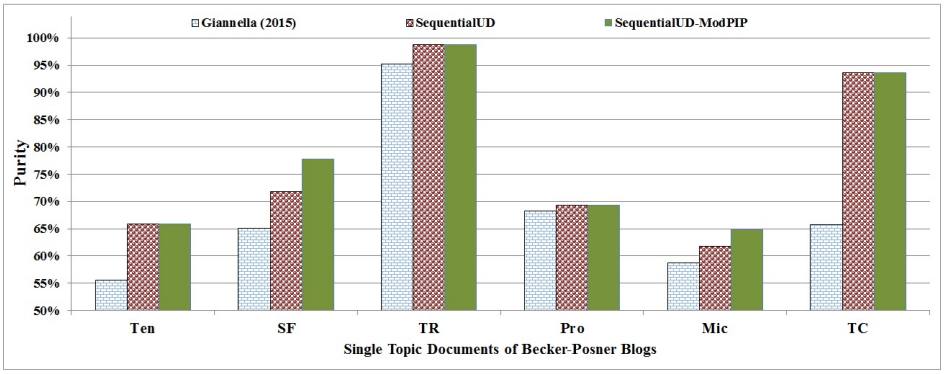

FIG. 4: Purity results of the approaches proposed by Giannella (2015), our SequentialUD and our refined SequentialUD using the six single-topic documents of Becker-Posner blogs.

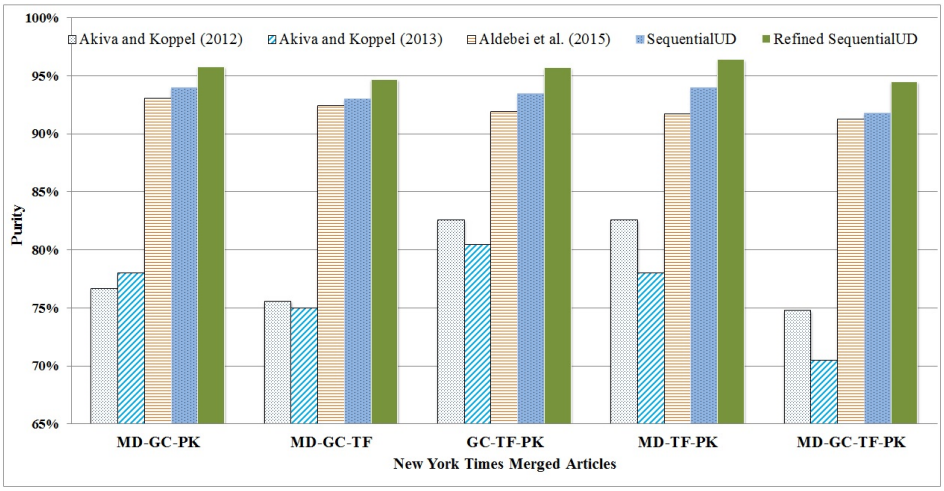

FIG. 5: Purity results of the approaches proposed by Akiva & Koppel (2012), Akiva & Koppel (2013), Aldebei et al. (2015), SequentialUD and refined SequentialUD using documents composed by merging articles of three and four *New York Times* columnists.

95.5% to 98.0% using SequentialUd and refined SequentialUD, respectively. We also find that the purity results of our approach have exceeded the results obtained using the Aldebei et al. (2015) in all of the six documents. Furthermore, these results outperform the purity of 88.0%, acquired by Akiva & Koppel (2012) and Akiva & Koppel (2013).

In the second set of our experiments regarding this corpus, all possible documents of three or four authors are composed. This results in four documents of three authors and one document of four authors. Figure 5 shows the purity results of applying our SequentialUD and refined SequentialUD on the five aforementioned documents (i.e., four documents having three authors and one document having four authors), comparing with the approaches of Akiva & Koppel (2012), Akiva & Koppel (2013) and Aldebei et al. (2015).

As shown in Figure 5, the purities achieved by our SequentialUD approach and its refined version are significantly higher no matter if a document is written by three or four authors. In addition, it should be noted that the proposed approach consistently outperforms the other three state-of-art approaches in all of the five documents. Figure 5 also shows that, in the experiments involving the four-author document (i.e., MD-GC-TF-PK), the refined SequentialUD produces a 34% higher purity than the approach in Akiva & Koppel (2013). Once again, comparing the purity results obtained in all of the five documents using the refined and non-refined version of our approach, one can see that, applying our ModPIP on the BoostedHMM has further improved the performance by 2.2% on average. This clearly demonstrates the effectiveness of the ModPIP procedure.

In order to examine the proposed approach in shorter documents, another set of experiments regarding the *New York Times* corpus is applied. In this set of experiments, merged documents of two, three and four columnists composed of only n different randomly selected articles of each columnist are created. For each resultant merged document, the SequentialUD approach and its refined version are applied and the purity results are computed. We repeat this process 50 times and then the mean purity results over the 50 trials for SequentialUD and its refined

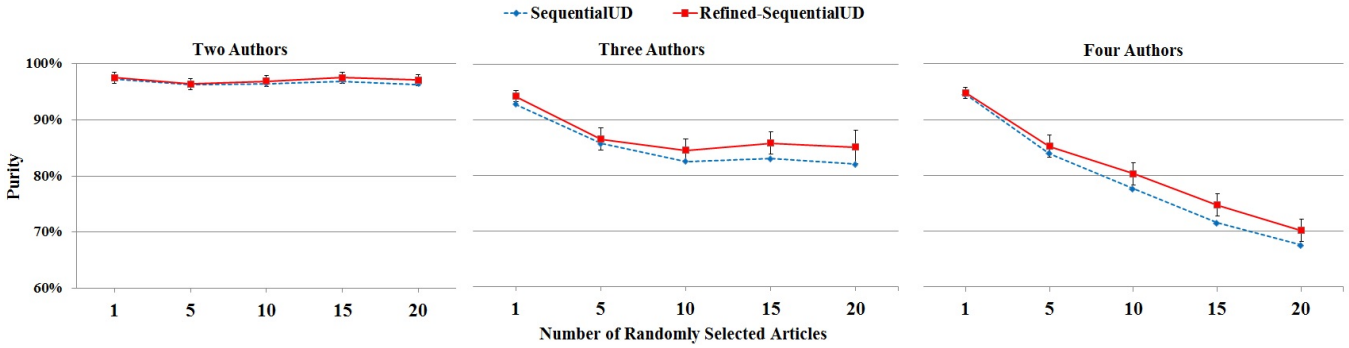

FIG. 6: Purity results of SequentialUD approach and its refined version on merged documents created by merging 1, 5, 10, 15 and 20 randomly selected articles of two, three and four authors. The error bars depict 0.95 confidence interval for the refined SequentialUD approach. In many cases the confidence intervals are quite small and are not easily seen in the figure.

are computed. The 0.95 confidence intervals over the 50 purity results for the refined SequentialUD version are also computed. Figure 6 shows the purity results of SequentialUD approach and its refined version on merged documents created by merging 1, 5, 10, 15 and 20 randomly selected articles of two, three and four authors. Figure 6 also shows the 0.95 confidence intervals for the refined SequentialUD version.

As shown in Figure 6, the purity results obtained with our proposed approach on *New York Times* short documents are quite promising.

In Giannella (2015), the author has examined his approach of document decomposition on short documents created by merging a few sentences of the four columnists of New York Times articles. These documents are created using a procedure which is different from the one used in this article. The procedure aims to create a merged document containing a specific number of runs of successive sentences of each columnist. Giannella has performed the procedure for 100 trials. In each trial, a multi-author document is created, his approach on document decomposition is performed, and a matching accuracy is then computed. After that, the mean and the 0.95 confidence intervals over the 100 accuracies of the approach are computed. The experiments applied in this article (excluding the six single topic documents) have assumed that there is a long sequence of consecutive sentences for each author. It is interesting to see how our proposed approach can be applied in short documents with short consecutive sentences. Therefore, we create short documents of four columnists of *New York Times* articles using the same procedure of Giannella (2015). Each created merged document contains exactly two runs of each columnist. During each run of each columnist, the number of selected successive sentences of the columnist in that run is randomly chosen from an exponential distribution with meanARL (when the chosen number is not an integer, we round it to the nearest integer). Figure 7 shows the purity results of SequentialUD approach and its refined version in short documents, which are composed by merging articles of four New York Times columnists using the same procedure of Giannella (2015), when the mean author run length (i.e., meanARL) is varied.

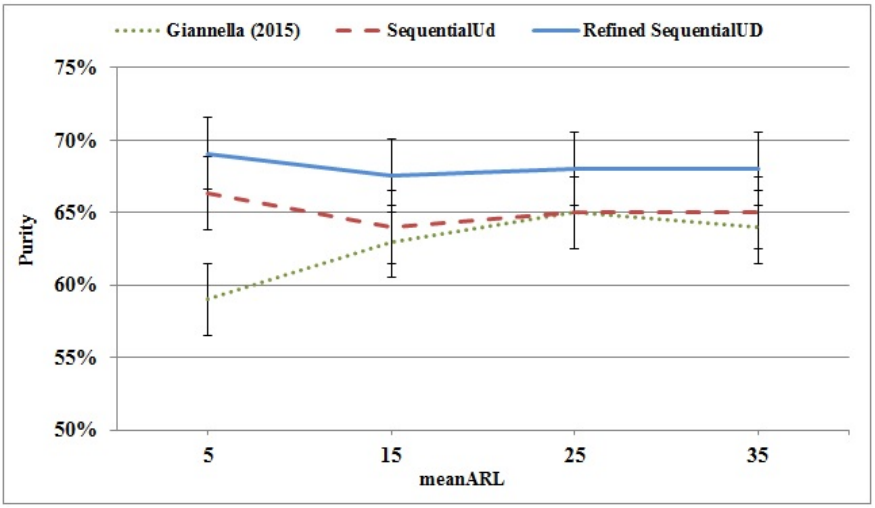

FIG. 7: Comparison of the purity results obtained using the approach in Giannella (2015), SequentialUD approach and refined SequnatialUD approach on short documents with short consecutive sentences composed by merging articles of four *New York Times* columnists using the same procedure of Giannella (2015), when the mean author run length (i.e., meanARL) is varied. The error bars depict 0.95 confidence interval for the three approaches.

As shown in Figure 7, the purity results of the SequentialUD and its refined version outperform the results of Giannella's approach in all values of meanARL (except in the case of meanARL = 25 where SequentialUD achieves a purity equal to that achieved in Giannella's approach).

Results on Randomly Selected Scientific Articles

To show the efficiency of our SequentialUD and its refined version on more realistic cases, we employ randomly selected scientific articles covering the same topics and create a set of merged articles. Each merged article has two, three or four authors and the topics among authors are not differentiated. In the first set of experiments regarding the scientific articles, we create a merged article using two randomly selected scientific articles on plagiarism detection topic (i.e., Rao et al. (2011) and Kestemont et al. (2011)). The merged article consists of 177 sentences written by two authors. Our proposed approach achieves purities of 94.9% by using the SequentialUD approach and 98.3% by using its refined version. In the second set of experiments regarding the scientific articles, we create a merged article using three randomly selected scientific articles on authorship attribution topic (i.e., Baayen et al. (2002), Layton et al. (2010) and Savoy (2016)). The merged article consists of 592 sentences written by three authors. Our proposed approach achieves purities of 92.2% by using the SequentialUD approach and 92.4% by using its refined version. In the last set of experiments regarding the scientific articles, we create a merged article using four randomly selected scientific articles on authorship-based document decomposition topic (i.e., Akiva & Koppel (2013), Giannella (2015), Daks & Clark (2016) and Aldebei, He, & Yang (2016)). The merged article consists of 805 sentences written by four authors. Our proposed approach achieves purities of 93.7% by using the SequentialUD approach and 98.8% by using its refined version. It is clear that the purity results obtained with our approach on merged articles of two, three or four authors are quite promising.

Results on Scientific Document

To also work on other types of non-artificial, scientific document, we have applied our SequentialUD approach and its refined version on a scientific paper initially drafted by two Ph.D students (i.e., Students A and B). Our proposed approach obtains purities of 95.5% by using the SequentialUD approach and 96.8% by using its refined version. It is observed that the purity results obtained using the proposed approach on a non-artificial, scientific document are very promising.

Conclusions

In this work, aiming at segmenting a multi-author document into components according to authorship, we have proposed to utilize the useful sequential correlation among the consecutive sentences in order to determine the authorial components. The well-known sequential model, i.e., Hidden Markov Model, has been adopted to find the best sequence of authors that represents the corresponding sequence of sentences in the document.

Our comparative evaluation results with the state-of-the-arts have demonstrated the strength of our proposed idea in terms of effectively decomposing the document into sentences according to their authorship, regardless of their topics, languages, etc. It has been noticed that the performance of the proposed approach is better when the length of successive sentences of each author is relatively long.

The great strength of our refined SequentialUD approach is the selection of the trusted sentences from the document and using them in re-classifying sentences, such as very short sentences, that do not have sufficiently discriminative features.

However, the proposed approach may not yet be effectively to predict the authors of sentences when the majority of the sentences are very short. For example, each message in a chat on a social network (e.g., *Facebook*) often contains a few words only. Furthermore, the proposed approach requires that the number of contributing authors of a document is known beforehand. Estimating the number of authors may be a direction of future work.

Acknowledgement

This project is partly supported by an Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201701).

References

Akiva, N., & Koppel, M. (2012). Identifying distinct components of a multi-author document. In *2012 European Intelligence and Security Informatics Conference* (pp. 205–209).

Akiva, N., & Koppel, M. (2013). A generic unsupervised method for decomposing multi-author documents. *Journal of the American Society for Information Science and Technology*, *64*(11), 2256–2264.

Aldebei, K., He, X., Jia, W., & Yang, J. (2016). Unsupervised multi-author document decomposition based on hidden Markov model. In *The 54th Annual Meeting of the Association for Computational Linguistics.*

Aldebei, K., He, X., & Yang, J. (2015). Unsupervised decomposition of a multi-author document based on Naive-Bayesian model. *Association for Computational Linguistics, Volume 2: Short Papers*, 501.

Aldebei, K., He, X., & Yang, J. (2016). Unsupervised decomposition of a multi-author document based on hidden Markov model. In *Submitted to of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, *12*(4), 461–486.

Baayen, H., Halteren, H. van, Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. In *6th JADT* (pp. 29–37).

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Brooke, J., Hammond, A., & Hirst, G. (2012). Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. In *Proceedings of the NAACL 12 Workshop on Computational Linguistics for Literature. Montreal, QC* (pp. 26–35).

Brooke, J., Hirst, G., & Hammond, A. (2013). Clustering voices in the waste land. In *Proceedings of the 2nd Workshop on Computational Literature for Literature (CLFL13), Atlanta.*

Daks, A., & Clark, A. (2016). Unsupervised authorial clustering based on syntactic structure. *ACL 2016*, 114.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Forney Jr, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268–278.

Giannella, C. (2015). An improved algorithm for unsupervised decomposition of a multi-author document. *Journal of the Association for Information Science and Technology*.

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language & the Law*, *14*(1).

Hoang, X. D., & Hu, J. (2004). An efficient hidden Markov model training scheme for anomaly intrusion detection of server applications based on system calls. In *Networks, 2004.(ICON 2004). Proceedings. 12th IEEE International Conference on* (Vol. 2, pp. 470–474).

Kestemont, M., Luyckx, K., & Daelemans, W. (2011). Intrinsic plagiarism detection using character trigram distance scores. *Proceedings of the PAN*.

Koppel, M., Akiva, N., Dershowitz, I., & Dershowitz, N. (2011). Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 1356–1364).

Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, *65*(1), 178–187.

Layton, R., Watters, P., & Dazeley, R. (2010). Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second* (pp. 1–8).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 84). Marcel Dekker.

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

Oberreuter, G., LHuillier, G., Rios, S. A., & Velasquez, J. D. (2011). Approaches for intrinsic and external plagiarism detection. *Proceedings of the PAN*.

Orebaugh, A., Kinser, J., & Allnutt, J. (2014). Visualizing instant messaging author writeprints for forensic analysis. In *Proceedings of the Conference on Digital Forensics, Security and Law* (pp. 191–214).

Rabiner, L. R., & Juang, B.-H. (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE*, *3*(1), 4–16.

Rao, S., Gupta, P., Singhal, K., & Majumder, P. (2011). External & Intrinsic Plagiarism Detection: VSM & Discourse Markers based Approach Notebook for PAN at CLEF 2011.

Savoy, J. (2016). Estimating the probability of an authorship attribution. *Journal of the Association for Information Science and Technology*, *6*(67), 1462–1472.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556.

Stamatatos, E. (2011). Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, *62*(12), 2512–2527.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269.

Zhao, Y., & Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis.

Zipf, G. K. (1932). Selected studies of the principle of relative frequency in language.