Unsupervised Multi-Author Document Decomposition Based on Hidden Markov Model

Khaled Aldebei **Xiangjian He** **Wenjing Jia**
Global Big Data Technologies Centre
University of Technology Sydney
Australia
{Khaled.Aldebei,Xiangjian.He,Wenjing.Jia}@uts.edu.au

Jie Yang
Lab of Pattern Analysis
and Machine Intelligence
Shanghai Jiaotong University
China
Jieyang@sjtu.edu.cn

Abstract

This paper proposes an unsupervised approach for segmenting a multi-author document into authorial components. The key novelty is that we utilize the sequential patterns hidden among document elements when determining their authorships. For this purpose, we adopt Hidden Markov Model (HMM) and construct a sequential probabilistic model to capture the dependencies of sequential sentences and their authorships. An unsupervised learning method is developed to initialize the HMM parameters. Experimental results on benchmark datasets have demonstrated the significant benefit of our idea and our approach has outperformed the state-of-the-arts on all tests. As an example of its applications, the proposed approach is applied for attributing authorship of a document and has also shown promising results.

1 Introduction

Authorship analysis is a process of inspecting documents in order to extract authorial information about these documents. It is considered as a general concept that embraces several types of authorship subjects, including *authorship verification*, *plagiarism detection* and *author attribution*. Authorship verification (Brocardo et al., 2013; Potha and Stamatatos, 2014) decides whether a given document is written by a specific author. Plagiarism detection (Stein et al., 2011; Kestemont et al., 2011) seeks to expose the similarity between two texts. However, it is unable to determine if they are written by the same author. In author attribution (Juola, 2006; Savoy, 2015), a real author of an anonymous document is predicted using labeled documents of a set of candidate authors.

Another significant subject in authorship analysis, which has received comparatively less attention from research community, is *authorship-based document decomposition* (ABDD). This subject is to group the sentences of a multi-author document to different classes, of which each contains the sentences written by only one author. Many applications can take advantage of such a subject, especially those in forensic investigation, which aim to determine the authorship of sentences in a multi-author document. Furthermore, this kind of subject is beneficial for detecting plagiarism in a document and defining contributions of authors in a multi-author document for commercial purpose. ABDD can also be applied to identify which source (regarded as an 'author' in this paper) a part of a document is copied from when the document is formed by taking contents from various sources.

In despite of the benefits of ABDD, there has been little research reported on this subject. Koppel et al. (2011) are the first researchers who implemented an unsupervised approach for ABDD. However, their approach is restricted to Hebrew documents only. The authors of Akiva and Koppel (2013) addressed the drawbacks of the above approach by proposing a generic unsupervised approach for ABDD. Their approach utilized distance measurements to increase the precision and accuracy of clustering and classification phases, respectively. The accuracy of their approach is highly dependent on the number of au-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 706–714,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

thors. When the number of authors increases, the accuracy of the approach is significantly dropped. Giannella (2015) presented an improved approach for ABDD when the number of authors of the document is known or unknown. In his approach, a Bayesian segmentation algorithm is applied, which is followed by a segment clustering algorithm. However, the author tested his approach by using only documents with a few transitions among authors. Furthermore, the accuracy of the approach is very sensitive to the setting of its parameters. In Aldebei et al. (2015), the authors presented an unsupervised approach ABDD by exploiting the differences in the posterior probabilities of a Naive-Bayesian model in order to increase the precision and the classification accuracy, and to be less dependent on the number of authors in comparing with the approach in Akiva and Koppel (2013). Their work was tested on documents with up to 400 transitions among authors and the accuracy of their approach was not sensitive to the setting of parameters, in contrast with the approach in Giannella (2015). However, the performance of their approach greatly depends on a threshold, of which the optimal value for an individual document is not easy to find.

Some other works have focused on segmenting a document into components according to their topics. For applications where the topics of documents are unavailable, these topic-based solutions will fail. In this paper, the ABDD approach is independent of documents' topics.

All of the existing works have assumed that the observations (i.e., sentences) are independent and identically distributed (i.i.d.). No consideration has been given to the contextual information between the observations. However, in some cases, the i.i.d. assumption is deemed as a poor one (Rogovschi et al., 2010). In this paper, we will relax this assumption and consider sentences of a document as a sequence of observations. We make use of the contextual information hidden between sentences in order to identify the authorship of each sentence in a document. In other words, the authorships of the "previous" and "subsequent" sentences have relationships with the authorship of the current sentence. Therefore, in this paper, a well-known sequential model, Hidden Markov Model (HMM), is used for modelling the sequential patterns of the document in order to describe the authorship relationships.

The contributions of this article are summarized as follows.

1. We capture the dependencies between consecutive elements in a document to identify different authorial components and construct an HMM for classification. It is for the first time the sequential patterns hidden among document elements is considered for such a problem.

2. To build and learn the HMM model, an unsupervised learning method is first proposed to estimate its initial parameters, and it does not require any information of authors or document's context other than how many authors have contributed to write the document.

3. Different from the approach in Aldebei et al. (2015), the proposed unsupervised approach no longer relies on any predetermined threshold for ABDD.

4. Comprehensive experiments are conducted to demonstrate the superior performance of our ideas on both widely-used artificial benchmark datasets and an authentic scientific document. As an example of its applications, the proposed approach is also applied for attributing authorship on a popular dataset. The proposed approach can not only correctly determine the author of a disputed document but also provide a way for measuring the confidence level of the authorship decision for the first time.

The rest of this article is organised as follows. Section 2 reviews the HMM. Section 3 presents the details of our proposed approach, including the processes for initialization and learning of HMM parameters, and the Viterbi decoding process for classification. Experiments are conducted in Section 4, followed by the conclusion in Section 5.

2 Overview of HMM

In this paper, we adopt the widely used sequential model, the Hidden Markov Model (HMM) (Eddy, 1996), to classify sentences of a multi-author document according to their authorship. The HMM is a probabilistic model which describes the statistical dependency between a sequence of observations $O = \{o_1, o_2, \cdots, o_T\}$ and a sequence of hidden states $Q = \{q_1, q_2, \cdots, q_T\}$. The observations can either be discrete variables, where each o_i takes a value from a set of M symbols $W = \{w_1, \cdots, w_M\}$, or be continuous variables. On the other hand, each q_i takes one possible value from a set of N symbols, $S = \{s_1, \cdots, s_N\}$.

The behaviour of the HMM can be determined by three parameters shown as follows.

1. Initial state probabilities $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_N\}$, where $\pi_n = p(q_1 = s_n)$ and $s_n \in S$, for $n = 1, 2, \cdots, N$.

2. Emission probabilities $\boldsymbol{B}$, where each emission probability $b_n(o_t) = p(o_t|q_t = s_n)$, for $t = 1, 2, \cdots, T$ and $n = 1, 2, \cdots, N$.

3. State transition probabilities $\boldsymbol{A}$. It is assumed that the state transition probability has a time-homogeneous property, i.e., it is independent of the time t. Therefore, a probability $p(q_t = s_l|q_{t-1} = s_n)$ can be represented as a_{nl}, for $t = 1, 2, \cdots, T$ and $l, n = 1, 2, \cdots, N$.

3 The Proposed Approach

The ABDD proposed in this paper can be formulated as follows. Given a multi-author document C, written by N co-authors, it is assumed that each sentence in the document is written by one of the N co-authors. Furthermore, each co-author has written long successive sequences of sentences in the document. The number of authors N is known beforehand, while typically no information about the document contexts and co-authors is available. Our objective is to define the sentences of the document that are written by each co-author.

Our approach consists of three steps shown as follows.

1. Estimate the initial values of the HMM parameters $\{\boldsymbol{\pi}, \boldsymbol{B}, \boldsymbol{A}\}$ with a novel unsupervised learning method.

2. Learn the values of the HMM parameters using the $Baum - Welch$ algorithm (Baum, 1972; Bilmes and others, 1998).

3. Apply the *Viterbi* algorithm (Forney Jr, 1973) to find the most likely authorship of each sentence.

3.1 Initialization

In our approach, we assume that we do not know anything about the document C and the authors, except the number of co-authors of the document (i.e., N). This approach applies an HMM in order to classify each sentence in document C into a class corresponding to its co-author. The step (see Sub-section 3.2) for learning of HMM parameters $\{\boldsymbol{\pi}, \boldsymbol{B}, \boldsymbol{A}\}$ is heavily dependent on the initial values of these parameters (Wu, 1983; Xu and Jordan, 1996; Huda et al., 2006). Therefore, a good initial estimation of the HMM parameters can help achieve a higher classification accuracy.

We take advantage of the sequential information of data and propose an unsupervised approach to estimate the initial values of the HMM parameters. The detailed steps of this approach are shown as follows.

1. The document C is divided into *segments*. Each segment has 30 successive sentences, where the i^{th} segment comprises the i^{th} 30 successive sentences of the document. This will produce s segments, where s = Ceiling($|C|/30$) with $|C|$ representing the total number of sentences in the document. The number of sentences in each segment (i.e., 30) is chosen in such a way that each segment is long enough for representing a particular author's writing style, and also the division of the document gives an adequate number of segments in order to be used later for estimating the initial values of HMM parameters.

2. We select the words appearing in the document for more than two times. This produces a set of D words. For each segment, create a D-dimensional vector where the i^{th} element in the vector is one (zero) if the i^{th} element in the selected word set does (not) appear in the segment. Therefore, s binary D-dimensional vectors are generated, and the set of these vectors is denoted by $X = \{x_1, \cdots, x_s\}$.

3. A multivariate Gaussian Mixture Models (GMMs) (McLachlan and Peel, 2004) is used to cluster the D-dimensional vectors X into N components denoted by $\{s_1, s_2, \cdots, s_N\}$. Note that the number of components is equal to the number of co-authors of the document. Based on the GMMs, each vector, x_i, gets a label representing the Gaussian component that this vector x_i is assigned to, for $i = 1, 2, \cdots, s$.

4. Again, we represent each segment as a binary vector using a new feature set containing all words appearing in the document for at least once. Assuming the number of elements in the new feature set is D', s binary D'-dimensional vectors are generated, and the set of these vectors is denoted by $X' = \{x'_1, \cdots, x'_s\}$. Each vector x'_i will have the same label of vector x_i, for $i = 1, 2, \cdots, s$.

5. We construct a Hidden Markov model with a sequence of observations O' and its corresponding sequence of hidden states Q'. In this model, O' represents the resulted segment vectors X' of the previous step. Formally, observation o'_i, is the i^{th} binary D'-dimensional vector x'_i, that represents the i^{th} segment of document C. In contrast, Q' represents the corresponding authors of the observation sequence O'. Each q'_i symbolizes the most likely author of observation o'_i. According to Steps 3 and 4 of this sub-section, each x'_i representing o'_i takes one label from a set of N elements, and the label represents its state, for $i = 1, 2, \cdots, s$.

By assigning the most likely states to all hidden states (i.e., $q'_i, i = 1, 2, \cdots, s$), the state transition probabilities $\boldsymbol{A}$ are estimated.

As long as there is only one sequence of states in our model, the initial probability of each state is defined as the fraction of times that the state appears in the sequence Q', so $\pi_n = \frac{Count(q'=s_n)}{Count(q')}$, for $n = 1, 2, \cdots, N$.

6. Given the sequence X', and the set of all possible values of labels, the conditional probability of feature f_k in X' given a label s_n, $p(f_k|s_n)$, is computed, for $k = 1, 2, \cdots, D'$ and $n = 1, 2, \cdots, N$.

7. The document C is partitioned into sentences. Let $z = |C|$ represent the number of sentences in the document. We represent each sentence as a binary feature vector using the same feature set used in Step 4. Therefore, z binary D'-dimensional vectors, denoted by $O = \{o_1, \cdots, o_z\}$, are generated. By using the conditional probabilities resulted in Step 6, the initial values of $\boldsymbol{B}$ are computed as $p(o_i|s_n) = \prod_{k=1}^{D'} o_i^{f_k} p(f_k|s_n)$, where $o_i^{f_k}$ represents the value of feature f_k in sentence vector o_i, for $i = 1, 2, \cdots, z$ and $n = 1, 2, \cdots, N$.

In this approach, we use *add-one smoothing* (Martin and Jurafsky, 2000) for avoiding zero probabilities of $\boldsymbol{A}$ and $\boldsymbol{B}$. Furthermore, we take the logarithm function of the probability in order to simplify its calculations.

The initial values of the $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\pi}$ are now available. In next sub-section, the learning process of these parameter values is performed.

3.2 Learning HMM

After estimating the initial values for the parameters of HMM, we now find the parameter values that maximize likelihood of the observed data sequence (i.e., sentence sequence). The learning process of the HMM parameter values is performed as follows.

1. Construct a Hidden Markov model with a sequence of observations, O, and a corresponding sequence of hidden states, Q. In this model, O represents the resulted sentence vectors (Step 7 in the previous Sub-section). Formally, the observation o_i, is the i^{th} binary D'-dimensional vector and it represents the i^{th} sentence of document C. In contrast, Q represents the corresponding authors of observation sequence O. Each q_i symbolizes the most likelihood author of observation o_i, for $i = 1, 2, \cdots, z$

2. The Baum-Welch algorithm is applied to learn the HMM parameter values. The algorithm, also known as the $forward-backward$ algorithm (Rabiner, 1989), has two steps, i.e., *E-step* and *M-step*. The *E-step* finds the expected author sequence (Q) of the observation sequence (O), and the *M-step* updates the HMM parameter values according to the state assignments. The learning procedure starts with the initial values of HMM parameters, and then the cycle of these two steps continues until a convergence is achieved in $\boldsymbol{\pi}$, $\boldsymbol{B}$ and $\boldsymbol{A}$.

The learned HMM parameter values will be used in the next sub-section in order to find the best sequence of authors for the given sentences.

3.3 Viterbi Decoding

For a Hidden Markov model, there are more than one sequence of states in generating the observation sequence. The Viterbi decoding algorithm (Forney Jr, 1973) is used to determine the best sequence of states for generating observation sequence. Therefore, by using the Hidden Markov model that is constructed in previous sub-section and the learned HMM parameter values, the Viterbi decoding algorithm is applied to find the best sequence of authors for the given sentences.

4 Experiments

In this section, we demonstrate the performance of our proposed approach by conducting experiments on benchmark datasets as well as one authentic document. Furthermore, an application on authorship attribution is presented using another popular dataset.

4.1 Datasets

Three benchmark corpora widely used for authorship analysis are used to evaluate our approach. Furthermore, an authentic document is also examined.

The first corpus consists of five Biblical books written by Ezekiel, Isaiah, Jeremiah, Proverbs and Job, respectively. All of these books are written in Hebrew. The five books belong to two types of literature genres. The first three books are related to prophecy literature and the other two books are related to a wisdom literature.

The second corpus consists of blogs written by the Nobel Prize-winning economist Gary S. Becker and the renowned jurist and legal scholar Richard A. Posner. This corpus, which is titled "The Becker-Posner Blogs" (`www.becker-posner-blog.com`), contains 690 blogs. On average, each blog has 39 sentences talking about particular topic. The Becker-Posner Blogs dataset, which is considered as a very important dataset for authorship analysis, provides a good benchmark for testing the proposed approach in a document where the topics of authors are not distinguishable. For more challenging documents, Giannella (2015) has manually selected six single-topic documents from Becker-Posner blogs. Each document is a combination of Becker and Posner blogs that are talking about only one topic. The six merged documents with their topics and number of sentences of each alternative author are shown in Table 1.

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

Table 1: The 6 merged single-topic documents of Becker-Posner blogs.

The third corpus is a group of New York Times articles of four columnists. The articles are subjected to different topics. In our experiments, all possible multi-author documents of articles of these columnists are created. Therefore, this corpus permits us to examine the performance of our approach in documents written by more than two authors.

The fourth corpus is a very early draft of a scientific article co-authored by two PhD students each being assigned a task to write some full sections of the paper. We employ this corpus in order to evaluate the performance of our approach on an authentic document. For this purpose, we have disregarded its titles, author names, references, figures and tables. After that, we get 313 sentences which are written by two authors, where Author 1 has written 131 sentences and Author 2 has written 182 sentences.

4.2 Results on Document Decomposition

The performance of the proposed approach is evaluated through a set of comparisons with four state-of-the-art approaches on the four aforementioned datasets.

The experiments on the first three datasets, excluding the six single-topic documents, are applied using a set of artificially merged multi-author documents. These documents are created by using the same method that has been used by Aldebei et al. (2015). This method aims to combine a group of documents of N authors into a single merged document. Each of these documents is written by only one author. The merged document process starts by selecting a random author from an author set. Then, the first r successive and unchosen sentences from the documents of the selected author are gleaned, and are merged with the first r successive and unchosen sentences from the documents of another randomly selected author. This process is repeated till all sentences of authors' documents are gleaned. The value of r of each transition is selected randomly from a uniform distribution varying from 1 to V. Furthermore, we follow Aldebei et al. (2015) method and assign the value of 200 to V.

Bible Books

We utilize the bible books of five authors and create artificial documents by merging books of any two possible authors. This produces 10 multi-author documents of which four have the same type of literature and six have different type of literature. Table 2 shows the comparisons of classification accuracies of these 10 documents by using our approach and the approaches developed by Koppel et al. (2011), Akiva and Koppel (2013)-500CommonWords, Akiva and Koppel (2013)-SynonymSet and Aldebei et al. (2015).

	Doc.	**1**	**2**	**3**	**4**	**5**
Different	Eze-Job	85.8%	98.9%	95.0%	99.0%	99.4%
	Eze-Prov	77.0%	99.0%	91.0%	98.0%	98.8%
	Isa-Prov	71.0%	95.0%	85.0%	98.0%	98.7%
	Isa-Job	83.0%	98.8%	89.0%	99.0%	99.4%
	Jer-Job	87.2%	98.2%	93.0%	98.0%	98.5%
	Jer-Prov	72.2%	97.0%	75.0%	99.0%	99.5%
	Overall	***79.4%***	***97.8%***	***88.0%***	***98.5%***	***99.1%***
Same	Job-Prov	85.0%	94.0%	82.0%	95.0%	98.2%
	Isa-Jer	72.0%	66.9%	82.9%	71.0%	72.1%
	Isa-Eze	79.0%	80.0%	88.0%	83.0%	83.2%
	Jer-Eze	82.0%	97.0%	96.0%	97.0%	97.3%
	Overall	***79.5%***	***84.5%***	***87.2%***	***86.5%***	***87.7%***

Table 2: Classification accuracies of merged documents of different literature or the same literature bible books using the approaches of 1- Koppel et al. (2011), 2- Akiva and Koppel (2013)-500CommonWords, 3- Akiva and Koppel (2013)-SynonymSet, 4- Aldebei et al. (2015) and 5- our approach.

As shown in Table 2, the results of our approach are very promising. The overall classification accuracies of documents of the same literature or different literature are better than the other four state-of-the-art approaches.

In our approach, we have proposed an unsupervised method to estimate the initial values of the HMM parameters (i.e., $\boldsymbol{\pi}$, $\boldsymbol{B}$ and $\boldsymbol{A}$) using segments. Actually, the initial values of the HMM parameters are sensitive factors to the convergence and accuracy of the learning process. Most of the previous works using HMM have estimated these values by clustering the original data, i.e., they have clustered sentences rather than segments. Figure 1 compares the results of using segments with the results of using sentences for estimating the initial parameters of HMM in the proposed approach for the 10 merged Bible documents in terms of the accuracy results and number of iterations till convergence, respectively. From Figures 1, one can notice that the accuracy results obtained by using segments for estimating the initial HMM parameters are significantly higher than using sentences for all merged documents. Furthermore, the number of iterations required for convergence for each merged document using segments is significantly smaller than using sentences.

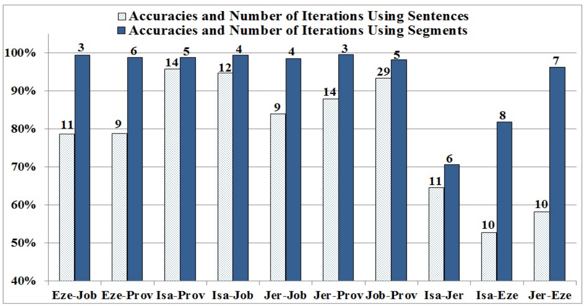

Figure 1: Comparisons between using segments and using sentences in the unsupervised method for estimating the initial values of the HMM of our approach in terms of accuracy (representd as the cylinders) and number of iterations required for convergence (represented as the numbers above cylinders) using the 10 merged Bible documents.

Becker-Posner Blogs (Controlling for Topics)

In our experiments, we represent Becker-Posner blogs in two different terms. The first term is as in Aldebei et al. (2015) and Akiva and Koppel (2013) approaches, where the whole blogs are exploited to create one merged document. The resulted merged document contains 26,922 sentences and more than 240 switches between the two authors. We obtain an accuracy of 96.72% when testing our approach in the merged document. The obtained result of such type of document, which does not have topic indications to differentiate between authors, is delightful. The first set of cylinders labelled "Becker-Posner" in Figure 2 shows the comparisons of classification accuracies of our approach and the approaches of Akiva and Koppel (2013) and Aldebei et al.

(2015) when the whole blogs are used to create one merged document. As shown in Figure 2, our approach yields better classification accuracy than the other two approaches.

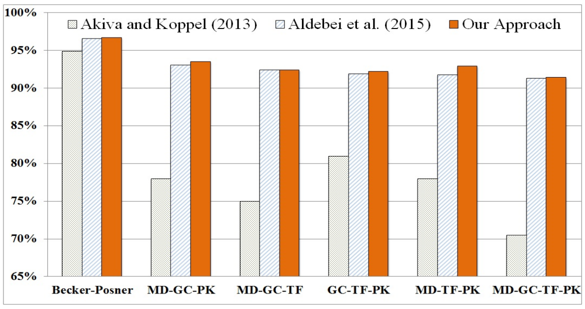

Figure 2: Classification accuracy comparisons between our approach and the approaches presented in Akiva and Koppel (2013) and Aldebei et al. (2015) in Becker-Posner documents, and documents created by three or four *New York Times* columnists (TF = Thomas Friedman, PK = Paul Krugman, MD = Maureeen Dowd, GC = Gail Collins).

The second term is as in the approach of Giannella (2015), where six merged single-topic documents are formed. Due to comparatively shorter lengths of these documents, the number of resulted segments that are used for the unsupervised learning in Sub-section 3.1 is clearly not sufficient. Therefore, instead of splitting each document into segments of 30 sentences length each, we split it into segments of 10 sentences length each. Figure 3 shows the classification accuracies of the six documents using our approach and the approach presented in Giannella (2015). It is observed that our proposed approach has achieved higher classification accuracy than Giannella (2015) in all of the six documents.

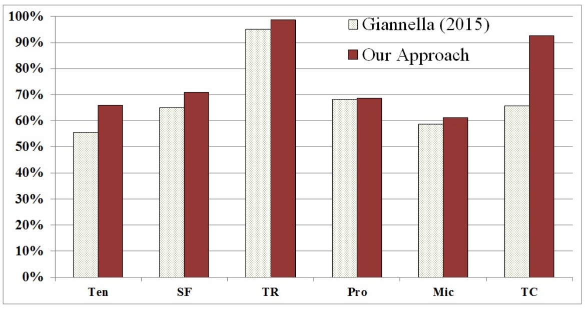

Figure 3: Classification accuracy comparisons between our approach and the approach presented in (Giannella, 2015) in the six single-topic documents of Becker-Posner blogs.

New York Times Articles ($N > 2$)

We perform our approach on *New York Times* articles. For this corpus, the experiments can be classified into three groups. The first group is for those merged documents that are created by combining articles of any pair of the four authors. The six resulted documents have on average more than 250 switches between authors. The classification accuracies of these documents are between 93.9% and 96.3%. It is notable that the results are very satisfactory for all documents. For comparisons, the classification accuracies of the same documents using the approach presented in Aldebei et al. (2015) range from 93.3% to 96.1%. Furthermore, some of these documents have produced an accuracy lower than 89.0% using the approach of Akiva and Koppel (2013).

The second group is for those merged documents that are created by combining articles of any three of the four authors. The four resulted documents have on average more than 350 switches among the authors. The third group is for the document that are created by combining articles of all four columnists. The resulted merged document has 46,851 sentences and more than 510 switches among authors. Figure 2 shows the accuracies of the five resulted documents regarding the experiments of the last two groups. Furthermore, it shows the comparisons of our approach and the approaches presented in Aldebei et al. (2015) and Akiva and Koppel (2013). It is noteworthy that the accuracies of our approach are better than the other two approaches in all of the five documents.

Authentic Document

In order to demonstrate that our proposed approach is applicable on genuine documents as well, we have applied the approach on first draft of a scientific paper written by two Ph.D. students (Author 1 and Author 2) in our research group. Each student was assigned a task to write some full sections of the paper. Author 1 has contributed 41.9% of the document and Author 2 contributed 58.1%. Table 3 shows the number of correctly assigned sentences of each author and the classification accuracy resulted using the proposed approach. Table 3 also displays the authors' contributions predicted using our approach. As

Author	**Classification Accuracy**	**Predicted Contribution**
1	98.5%	47.6%
2	89.0%	52.4%
Accuracy	**93.0%**	

Table 3: The classification accuracies and predicted contributions of the two authors of the scientific paper using the proposed approach.

shown in Table 3, the proposed approach has achieved an overall accuracy of 93.0% for the authentic document.

4.3 Results on Authorship Attribution

One of the applications that can take advantage of the proposed approach is the authorship attribution (i.e., determining a real author of an anonymous document given a set of labeled documents of candidate authors). The *Federalist Papers* dataset have been employed in order to examine the performance of our approach for this application. This dataset is considered as a benchmark in authorship attribution task and has been used in many studies related to this task (Juola, 2006; Savoy, 2013; Savoy, 2015). The *Federalist Papers* consist of 85 articles published anonymously between 1787 and 1788 by Alexander Hamilton, James Madison and John Jay to persuade the citizens of the State of New York to ratify the Constitution. Of the 85 articles, 51 of them were written by Hamilton, 14 were written by Madison and 5 were written by Jay. Furthermore, 3 more articles were written jointly by Hamilton and Madison. The other 12 articles (i.e., articles 49-58 and 62-63), the famous "anonymous articles", have been alleged to be written by Hamilton or Madison.

To predict a real author of the 12 anonymous articles, we use the first five undisputed articles of both authors, Hamilton and Madison. Note that we ignore the articles of Jay because the anonymous articles are alleged to be written by Hamilton or Madison. The five articles of Hamilton (articles 1 and 6-9) are combined with the five articles of Madison (articles 10, 14 and 37-39) in a single merged document where all the articles of Hamilton are inserted into the first part of the merged document and all the articles of Madison are inserted into the second part of the merged document. The merged document has 10 undisputed articles covering eight different topics (i.e., each author has four different topics). Before applying the authorship attribution on the 12 anonymous articles, we have tested our approach on the resulted merged document and an accuracy of 95.2% is achieved in this document. Note that, the authorial components in this document are not thematically notable.

For authorship attribution of the 12 anonymous articles, we add one anonymous article each time on the middle of the merged document, i.e., between Hamilton articles part and Madison articles part. Then, we apply our approach on the resulted document, which has 11 articles, to determine to which part the sentences of the anonymous article are classified to be sectences of Hamilton or Madison. As the ground truth for our experiments, all of these 12 articles can be deemed to have been written by Madison becuase the results of all recent state-of-the-art studies testing on these articles on authorship attribution have classified the articles to Madison's. Consistent with the state-of-the-art approaches, these 12 anonymous articles are also correctly classified to be Madison's using the proposed approach. Actually, all sentences of articles 50,52-58 and 62-63 are classified as Madison's sentences, and 81% of the sentences of article 49 and 80% of article 51 are classified as Madison's sentences. These percentages can be deemed as the confidence levels (i.e., 80% conferdence for articles 49, 81% for 51, and 100% confidences for all other articles) in making our conclusion of the authorship contributions.

5 Conclusions

We have developed an unsupervised approach for decomposing a multi-author document based on authorship. Different from the state-of-the-art approaches, we have innovatively made use of the sequential information hidden among document elements. For this purpose, we have used HMM and constructed a sequential probabilistic model, which is used to find the best sequence of authors that represents the sentences of the document. An unsupervised learning method has also been developed to estimate the initial parameter values of HMM. Comparative experiments conducted on benchmark datasets have demonstrated the effectiveness of our ideas with superior performance achieved on both artificial and authentic documents. An application of the proposed approach on authorship attribution has also achieved perfect results of 100% accuracies together with confidence measurement for the first time.

References

[Akiva and Koppel2013] Navot Akiva and Moshe Koppel. 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 et al.2015] Khaled Aldebei, Xiangjian He, and Jie Yang. 2015. Unsupervised decomposition of a multi-author document based on naive-bayesian model. *ACL, Volume 2: Short Papers*, page 501.

[Baum1972] Leonard E Baum. 1972. An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8.

[Bilmes and others1998] Jeff A Bilmes et al. 1998. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126.

[Brocardo et al.2013] Marcelo Luiz Brocardo, Issa Traore, Shatina Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on*, pages 1–6. IEEE.

[Eddy1996] Sean R Eddy. 1996. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

[Forney Jr1973] G David Forney Jr. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

[Giannella2015] Chris Giannella. 2015. An improved algorithm for unsupervised decomposition of a multi-author document. *Journal of the Association for Information Science and Technology.*

[Huda et al.2006] Md Shamsul Huda, Ranadhir Ghosh, and John Yearwood. 2006. A variable initialization approach to the em algorithm for better estimation of the parameters of hidden markov model based acoustic modeling of speech signals. In *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining*, pages 416–430. Springer.

[Juola2006] Patrick Juola. 2006. Authorship attribution. *Foundations and Trends in information Retrieval*, 1(3):233–334.

[Kestemont et al.2011] Mike Kestemont, Kim Luyckx, and Walter Daelemans. 2011. Intrinsic plagiarism detection using character trigram distance scores. *Proceedings of the PAN.*

[Koppel et al.2011] Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 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*, pages 1356–1364. Association for Computational Linguistics.

[Martin and Jurafsky2000] James H Martin and Daniel Jurafsky. 2000. Speech and language processing. *International Edition.*

[McLachlan and Peel2004] Geoffrey McLachlan and David Peel. 2004. *Finite mixture models.* John Wiley & Sons.

[Potha and Stamatatos2014] Nektaria Potha and Efstathios Stamatatos. 2014. A profile-based method for authorship verification. In *Artificial Intelligence: Methods and Applications*, pages 313–326. Springer.

[Rabiner1989] Lawrence R Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Rogovschi et al.2010] Nicoleta Rogovschi, Mustapha Lebbah, and Younes Bennani. 2010. Learning self-organizing mixture markov models. *Journal of Nonlinear Systems and Applications*, 1:63–71.

[Savoy2013] Jacques Savoy. 2013. The federalist papers revisited: A collaborative attribution scheme. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–8.

[Savoy2015] Jacques Savoy. 2015. Estimating the probability of an authorship attribution. *Journal of the Association for Information Science and Technology.*

[Stein et al.2011] Benno Stein, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82.

[Wu1983] CF Jeff Wu. 1983. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.

[Xu and Jordan1996] Lei Xu and Michael I Jordan. 1996. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151.