Unsupervised Decomposition of a Multi-Author Document Based on Naive-Bayesian Model

Khaled Aldebei   Xiangjian He
School of Computing and Communications
Faculty of Engineering and IT
University of Technology, Sydney, Australia
Khaled.w.aldebei@student.uts.edu.au
Xiangjian.He@uts.edu.au

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

Abstract

This paper proposes a new unsupervised method for decomposing a multi-author document into authorial components. We assume that we do not know anything about the document and the authors, except the number of the authors of that document. The key idea is to exploit the difference in the posterior probability of the Naive-Bayesian model to increase the precision of the clustering assignment and the accuracy of the classification process of our method. Experimental results show that the proposed method outperforms two state-of-the-art methods.

1 Introduction

The traditional studies on text segmentation, as shown in Choi (2000), Brants et al. (2002), Misra et al. (2009) and Hennig and Labor (2009), focus on dividing the text into signification components such as words, sentences and topics rather than authors. Natural Language Processing techniques (NLP) and various machine learning schemas have been applied for these approaches. Due to the availability of online communication facilities, the cooperation between authors to produce a document becomes much easier. The co-authored documents include Web pages, books, academic papers and blog posts. There are almost no approaches that have concentrated on developing techniques for segmentation of a multi-author document according to the authorship. The existing approaches, as those in Schaalje et al. (2013), Segarra et al. (2014) and Layton et al. (2013) that are most related to our research in this paper, deal with documents written by a single author only. Although the work in Koppel et al. (2011) has considered the segmentation of a document according to multi-authorship, this approach requires manual translations and concordance to be available beforehand. Hence, their method can only be applied on particular types of documents such as Bible books. Akiva and Koppel (2013) investigated this limitation and presented a generic unsupervised method. They evaluated their method using two different types of features. The first one is the occurrence of the 500 most common words in the document. The second one is the synonym set, which is only valid on special types of documents like Bible books. Their method relies on the distance measurement to increase the precision and accuracy of the clustering and classification process. The performance of this method is degraded when the number of authors increases to more than two.

The contributions of this paper are as follows.

- A procedure for segment elicitation is developed and it is applied in the clustering assignment process. It is for the first time to develop such a procedure relying upon the differences in the posterior probabilities.
- A probability indication procedure is developed to improve the accuracy of sentence classification. It selects the significant and trusted sentences from a document and involves them to reclassify all sentences in the document. Our approach does not require any information about the document and the authors other than the number of authors of the document.
- Our proposed method is not restricted to any type of documents. It is still workable even when the topics in a document are not detectable.

The organization of this paper is as follows. Section 2 demonstrates the proposed framework. Section 3 uses an example to clarify our method. Results are conducted in Section 4. Finally, Section 5 presents the conclusion and future work.

*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics
and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 501–505,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

2 Proposed Framework

Given a multi-author document written by l authors, it is assumed that every author has written consecutive sequences of sentences, and every sentence is completely written by only one of the l authors. The value of l is pre-defined.

Our approach goes through the following steps:

- *Step 1* Divide the document into segments of fixed length.
- *Step 2* Represent the resulted segments as vectors using an appropriate feature set which can differentiate the writing styles among authors.
- *Step 3* Cluster the resulted vectors into l clusters using an appropriate clustering algorithm targeting on achieving high *recall* rates.
- *Step 4* Re-vectorize the segments using a different feature set to more accurately discriminate the segments in each cluster.
- *Step 5* Apply the "*Segment Elicitation Procedure*" to select the best segments from each cluster to increase the *precision* rates.
- *Step 6* Re-vectorize all selected segments using another feature set that can capture the differences among the writing styles of all sentences in a document.
- *Step 7* Train the classifier using the Naive-Bayesian model.
- *Step 8* Classify each sentence using the learned classifier.
- *Step 9* Apply the "*Probability Indication Procedure*" to increase the *accuracy* of the classification results using five criteria.

To assess the performance of the proposed scheme, we perform our experiments on an artificially merged document. The generation of this merged document begins with randomly choosing an author from an authors list. Then, we pick up the first r previously-unselected sentences from a document of that author, and merge them with the first r previously-unselected sentences from the documents of other randomly selected authors. Keep doing like this until all sentences from all authors' documents are selected. The value of r on each switch is an integer value chosen randomly from a uniform distribution varying from 1 to V.

3 Ezekiel-Job Document as Example

For interpretative intent, we will exploit the bible books of Ezekiel and Job to create a merged document. The book of Ezekiel contains 1,273 sentences and book of Job contains 1,018 sentences. We use this example of a merged document to clarify each step of our proposed framework shown in Section 2. We also use this merged document to work out the values of parameters used in our approach. We set V to be equal to 200. In the merged document, there are 2,291 sentences in total and there are hence 20 transitions from Ezekiel sentences to Job sentences and from Job's to Ezekiel's.

In *Step 1*, we divide the merged document into segments. Each segment has 30 sentences. As a result, we get 77 segments, of which 34 are written by Ezekiel, 27 are written by Job and 16 are mixed. In *Step 2*, we represent each segment using a binary vector that reflects all words that appear at least three times in the document. In *Step 3*, we cluster these segments by using a Gaussian Mixture Model (GMMs) into 2 multivariate Gaussian densities. The GMMs are trained using the iterative Expectation-Maximization (EM) algorithm (Bilmes and others, 1998). We find that all 34 Ezekiel segments are clustered in Cluster 1, and all 27 Job segments are clustered in Cluster 2. Mixed segments are divided equally between the two clusters (Note that, the *recalls* of both cluster are 100%, and the precisions are 81% and 77% in Cluster 1 and Cluster 2, respectively). In *Step 4*, all of the segments in both clusters are re-vectorized using the binary representation of the 1500 most frequently-appeared words in the document.

In the *Step 5*, a Segment Elicitation Procedure is proposed. The key idea is to choose only the segments from a cluster that can best represent the writing style of the cluster. We call these selected segments *vital segments*. The vital segments have the following two features. First, they can represent the expressive style of a specific cluster. Second, they can distinguish the writing style of that cluster from other clusters. Henceforth, we consider all of the segments as labelled, based on the results of the clustering assignment (Step 3). To find the vital segments of each class (noting that, the term 'cluster' is now substituted with 'class'), we consider the differences in the posterior probabilities of each segment according to the other classes. Expressly, for each segment in a class, we compute the differences between the posterior probability of that segment in its class and the maximum posterior probability of that segment in other classes. Then, we select *s%* of them which have the biggest differences as vital segments of that class. To prevent the underflow point, we compute the posterior probability by adding the logarithms of probabilities instead of multiplying the probabilities. Furthermore, we assume that the features in the segments are mutually independent. In the Ezekiel-Job document, Cluster 1 is the Ezekiel class and Cluster 2 is the Job class. We set *s* to be 80, so we get 34 vital segments for the Ezekiel class and 28 vital segments for the Job class. Of the 34 vital segments in Ezekiel class, 30 are truly written by Ezekiel, and of the 28 vital segments in Job class, 25 are truly written by Job. As a result, the precisions of Ezekiel class and Job class are increased to 88.2% and 89.3%, respectively. The vital segments for two classes are used to train the supervised classifier which can best classify each sentence to the correct author's class. Therefore, in *Step 6*, the vital segments are represented in terms of the frequencies of all words that have appeared at least three times in the whole document.

In *Step 7*, the Naive-Bayesian model is applied to learn a classifier. In *Step 8*, this classifier is used to classify the sentences in the merged document to either Ezekiel class or Job class.We find that 93.1% of all sentences of Ezekiel and Job classes are correctly classified.

In (*Step 9*), a probability indication procedure is proposed based on the following five criteria. *First*, any sentence in the document is considered as *trusted sentence* if its posterior probability in its class is greater than its posterior probabilities in all other classes by more than threshold *q*. Thereupon, every trusted sentence holds its class. *Second*, if the first sentences in the document are not deemed to be trusted sentences, then they are assigned to the same class of the first trusted sentence that follow them. *Third*, if the last sentences in the document are not deemed to be trusted sentences, then they are assigned to the same class of the last trusted sentence that precede them. *Fourth*, if a group of unassigned sentences is located between two trusted sentences which have the same class, then all of the sentences in that group are assigned to the same class of these trusted sentences. *Fifth*, if a group of unassigned sentences is located between two trusted sentences which have different labels, then the best separated point in that group is detected to separate it into two subgroups, left and right subgroups. The left subgroup is assigned to the same label of the last trusted sentence that precede it and the right subgroup is assigned to the same label of the first trusted sentence that follow it. In the Ezekiel-Job document, by setting the value of *q* to be 5.0, 98.8% of the Ezekiel sentences and 99.1% of the Job sentences are correctly classified. The overall accuracy of all sentences is 99.0%.

4 Results

We use three datasets to test our method and show the adaptability of our method to different types of documents. The first dataset consists of 690 blogs written by Gary Becker and Richard Posner. This dataset containing articles of multiple authors is challenging because it covers a lot of different topics. That means, we cannot depend on the topics to help us distinguish the authors. The second dataset consists of 1,182 *New York Times* articles. These articles have been written by Maureeen Dowd, Gail Collins, Thomas Friedman and Paul Krugman. The third dataset consists of 5 biblical books which are written in Hebrew, a language other than English. These books are written by Isaiah (for Chapters 1-33), Jeremiah, Ezekiel, Job (for Chapters 3-41) and Proverbs. The first 3 are all in the prophetic literature and the other two are in the wisdom literature. In view of this, we conduct our experiments on three different datasets, each dataset has its characteristics which yield us to use it. In our experiments, the merged documents are created in the same way as we have discussed before. We set the value of *V* to be 200, and the number of authors of these documents to be two, three or four ($l = \{2,3,4\}$). We use the same values of the parameters as we have used in the Ezekiel-Job document.

4.1 Becker-Posner

In the first dataset, each author has written for a lot of different topics, and there have been some topics taken by both authors. Therefore, there is no topic indication to distinguish between the two authors. We have achieved an overall accuracy of 96.6% when testing on this dataset. This result is gratifying in this merged document that has more than 246 transitions between sentences writ-

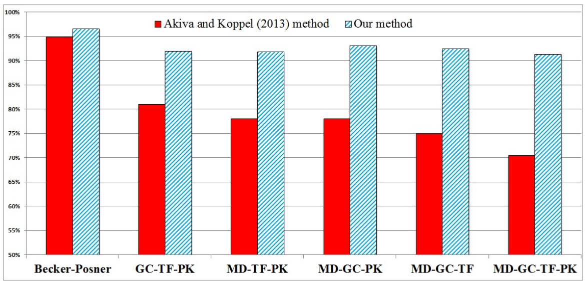

Figure 1: Accuracy comprisons between our method and the method used by Akiva and Koppel (2013) in Becker-Posner document, and in documents created by three or four *New York Times* authors (GC = Gail Collins, PK = Paul Krugman, TF = Thomas Friedman, MD = Maureen Dowd)

ten by the two authors and more than 26,900 sentences. In Figure 1, we show the comparison between our method and the method in Akiva and Koppel (2013).

4.2 *New York Times* Articles

This dataset contains articles written by four authors. First, we test our method using the merged documents created by any pair of the four authors. The results again are noticeable. The classification accuracies range from 93.3% to 96.1%. For comparison,the accuracy can be as low as 88.0% when applying the method in Akiva and Koppel (2013) on some of the merged documents.

To prove that our method can also work well when merged documents written by more than two authors, we have created merged documents written by any three of these four authors and formed four merged documents. We have also created a merged document written by all four *New York Times* authors. Then, we apply our method on these documents. In Figure 1, we show the accuracies of our method for classification on these documents. It is obvious that our method achieves high accuracies even when the documents are written by more than two authors. Furthermore, Figure 1 also compares our results with the results achieved by Akiva and Koppel (2013). It shows that our method has given consistent results and better performance than the ones in Akiva and Koppel (2013).

4.3 Bible Books

In these experiments, we use two literature types of biblical books. We create merged documents written by any pair of authors. The resulted documents may belong to either the same literatures or different literatures.

	Documents	**1**	**2**	**3**	**Our method**
Different	Eze-Prov	77%	99%	91%	98%
	Jer-Prov	73%	97%	75%	99%
	Jer-Job	88%	98%	93%	98%
	Isa-Job	83%	99%	89%	99%
	Eze-Job	86%	99%	95%	99%
	Isa-Prov	71%	95%	85%	98%
	Overall	***80%***	***98%***	***88%***	***99%***
Same	Jer-Eze	82%	97%	96%	97%
	Isa-Eze	79%	80%	88%	83%
	Job-Prov	85%	94%	82%	95%
	Isa-Jer	72%	67%	83%	71%
	Overall	***80%***	***85%***	***87%***	***87%***

Table 1: Accuracy performance obtained from documents having different literatures or same literatures using the methods of 1- Koppel et al. (2011), 2- Akiva and Koppel (2013)-BinaryCommonWords, 3- Akiva and Koppel (2013)-Synonyms and our method

In Tables 1, we show the comparisons of accuracies of using our method and the methods presented in Koppel et al. (2011), Akiva and Koppel (2013)-BinaryCommonWords and Akiva and Koppel (2013)-Synonyms.

As can be seen, the accuracies using our method in the documents with different literatures are interesting, and have achieved the accuracies of either 99% or 98% and have performed a lot better than the three state-of-the-art methods. Furthermore, the accuracies using our method on the documents with same literature are encouraging, and our method has achieved approximately the same overall accuracy compared with the method in Akiva and Koppel (2013), and have achieved better overall accuracy compared with the methods in Akiva and Koppel (2013) and Koppel et al. (2011).

5 Conclusion and Future Work

In this paper, we have proposed an unsupervised method for decomposing a multi-author document by authorship.

We have tested our method on three datasets, of which every one has its own characteristics. It is clear that our method has achieved a significantly high accuracies in these datasets, even when there is no topic indication to differentiate sentences between authors, and when the number of authors exceeds 2. Our results tested on these datasets have shown significantly better than those using the methods in Koppel et al. (2011) and Akiva and

Koppel (2013). Furthermore, our method can also compete with the method proposed in (Akiva and Koppel, 2013)-Synonyms, which is only valid for Bible documents.

In our research, our aim is to segment classify sentences in a multi-author document according to the sentences' authors. We assume that the number of authors of that document is known. In our future work, we work to automatically determine the number of authors of a multi-author document. Furthermore, we will explore an adaptive learning method to select the optimal value of the threshold q for the probability indication procedure.

References

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.

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.

Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 211–218. ACM.

Freddy YY Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33. Association for Computational Linguistics.

Leonhard Hennig and DAI Labor. 2009. Topic-based multi-document summarization with probabilistic latent semantic analysis. In *RANLP*, pages 144–149.

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.

Robert Layton, Paul Watters, and Richard Dazeley. 2013. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, 19(01):95–120.

Hemant Misra, François Yvon, Joemon M Jose, and Olivier Cappe. 2009. Text segmentation via topic modeling: an analytical study. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1553–1556. ACM.

G Bruce Schaalje, Natalie J Blades, and Tomohiko Funai. 2013. An open-set size-adjusted bayesian classifier for authorship attribution. *Journal of the American Society for Information Science and Technology*, 64(9):1815–1825.

Santiago Segarra, Mark Eisen, and Alejandro Ribeiro. 2014. Authorship attribution through function word adjacency networks. *arXiv preprint arXiv:1406.4469*.