Unsupervised Decomposition of Multi-Author Document: *Exploiting the difference of Syntactic writing styles*

Kautsya Kanu and **Sayantan Sengupta**
Indian Institute of Technology Delhi

Abstract

This paper proposes an improvement over a recent $paper^{[1]}$. We have worked on two aspects, In the first aspect, we try to capture writing style of author by n-gram model of words, POS Tags and PQ Gram model of syntactic parsing over used basic uni-gram model. In the second aspect, we added some layers of refinements in existing baseline model and introduce new term "similarity index" to distinguish between pure and mixed segments before unsupervised labeling. Similarity index uses overall and sudden change of writing style by PQ Gram model and words used using n-gram model between lexicalised/unlexicalised sentences in segments for refinement. In this paper, we investigate the role of feature selection that captures the syntactic patterns specific to an author and its overall effect in the final accuracy of the baseline system. More specifically, we insert a layer of refinement to the baseline system and define a threshold based on the similarity measure among the sentences to consider the purity of the segments to be given as input to the GMM.The key idea of our approach is to provide the GMM clustering with the "good segments" so that the clustering precision is maximised which is then used as labels to train a classifier. We also try different features set like bigrams and trigrams of POS tags and an PQ Grams based feature on unlexicalised PCFG to capture the distinct writing styles which is then given as an input to a GMM trained by iterative EM algorithm to generate good clusters of the segments of the merged document.

Introduction

Here we propose a new unsupervised method for decomposing a multi-author document in to authorial components. We have assumed that we have no prior information about the authors and the documents, except the number of authors of the document. The key idea is to exploit the differences of the grammatical writing styles of the authors and use this information to build paragraph clusters. This is a difficult problem in many levels. Its easy to decompose based on topics and contexts, which is often known as text segmentation in literature. So it gets difficult to distinguish if multiple authors have written on the same topic. Quantifying the difference of the grammatical writing styles of authors is another big challenge. As there is no prior information/access to the authors written texts, supervised classification approaches cant be applied directly. On top of this, the number of author is not known in general of a random a document/article in general (in case of plagiarism). So fixing the number of clusters is another big task. So considering the above constraints, this paper focus more on the feature selection part of the texts which is the most important part of the whole unsupervised clustering, as good features will lead to more precise clustering of the correct sentences to their respective clusters. The traditional studies on text segmentation, as shown in Choi (2000), Brants et al. (2002), Misra et al. (2009) and Henning and Labor (2009), focus on dividing the the text into significant components such as words, sentences and topics rather than authors. There are almost no approaches, as those in Schaalje et. al. (2013), Segarra et al (2014) and Layton et al. (2013) deal with documents written by a single author only. 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 document can only be applied on particular types of documents such as Bible books. Akiva and Koppel (2013) tried to come up with a solution. Their method relies on distance measurement to increase the precision and accuracy of the clustering and classification process. The performance is degraded when the number of authors increases to more than two.

Modified Baseline

After modifying the latest state of the art technique used is described below: Given a multi-author document written by *l* authors, it is assumed that every sentence is completely written by only one of the authors. The approach goes through the following steps:

- Divide the document into segments of fixed length.
- Represent each sentence inside a segment as vectors using n-grams of words and pq grams as feature set.
- Separate pure and mixed segments by analyzing sudden change in writing style or words used between sentences inside a segment using "similarity index" of segment.
- Represent the resulted pure segments as vectors using an appropriate feature set(Words, POS Tags, PQ Grams) in

whole merged document which can differentiate the writing styles among authors.

- Cluster the resulted vectors into l clusters using an appropriate clustering algorithm targeting on high recall rates(GMM with iterative EM algorithm).
- Re-vectorize the segments using a different feature set to more accurately discriminate the segments in each cluster.
- Apply the segment Elicitation procedure, which identifies the vital segments from each clusters to improve the precision rates.
- Re-vectorize all selected segments using another feature set that can capture the differences in the writing styles of all the sentences in a document.
- Train the classifier using a Naive Bayesian model.
- Classify each sentence using the learned classifier.

Data Set

The data sets we have used to evaluate our model is:

- 690 blogs written by Gary Becker and Richard Posner.
- 1,182 New York Times articles written by Maureen Dowd, Gail Collins, Thomas Freidman and Paul Krugman.
  Each data set has its own set of challenges, since each author has written a lot of different topics and some topics are taken by both authors.

The resulting table is shown below:

Table 1: Table Title

Dataset	Accuracy	sentences	Authors
Becker-Posner	0.82	26922	2
GC-TF-PK	0.67	11984	3
MD-TF-PK	0.70	13422	3
MD-GC-PK	0.66	13448	3
MD-GC-TF-PK	0.61	15584	4

Limitations of the Baseline System

We can see that no deep NLP features are used for the task. A bag of words model is a weak model to be able to discriminate between the authors. Also, the accuracy of the final stage of classification depends on the chunk (V) of sentences picked from the individual authors to form the merged document. Changing that parameter (V) from 200 to 50 reduces the final accuracy from 82% to 49%. Training on segments and testing on sentences is not such a good idea as the whole bottleneck for achieving high accuracy is the clustering algorithm.Also the cases of mixed segments(sentences comprising of both the authors) pose a problem during the clustering process which affects the precision and recall badly.

Figure 1:

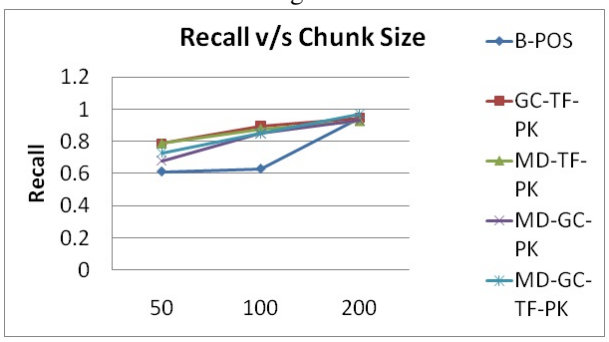
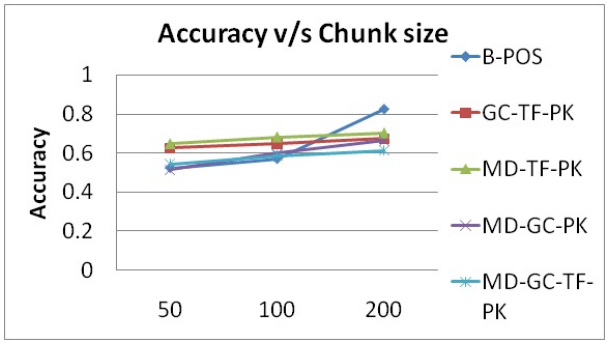


Figure 2:



Preliminary: PQ Grams

Similar to n-grams that represent the subparts of given length n of a string, p-q grams extract substructures of an ordered labelled tree. The size of p-q gram is determined by stem (p) and base (q). P defines how many nodes are included vertically, and q defines the number of nodes to be considered horizontally. For example, a valid p-q gram with p=2 and q=3 starting from PP at the left side of the tree (S2) shown in the above figure would be [PP-NP-DT-JJ-NNS]. The p-q gram index then consists of all possible p-q grams of a tree. In order to obtain all p-q grams, the base is shifted left and right additionally. If less than p nodes exists horizontally, the corresponding place in the pq-gram is filled with *, indicating a missing node.

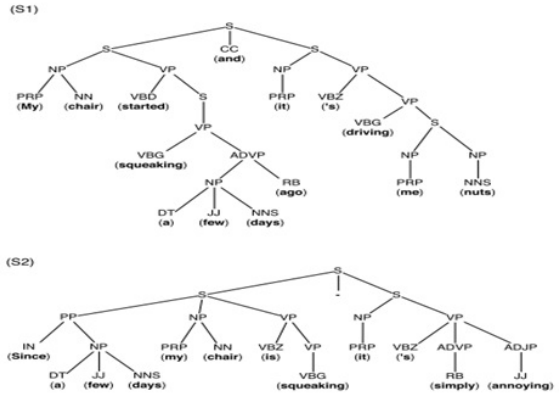
Proposed Methodology

There are two different aspects of this paper. The first aspect is to use a PQ grams based model to featurise the sentences of the segments and study its effect on the final accuracy. The second aspect is to introduce a layer of filtering in to the segments to identify which segments are pure(written completely by one author) to the most extent posible by defining a theshold value which ideally is a similarity score among the sentences in each segments.This will help us in getting better cluster assignments.

In the first aspect, the main idea is to quantify the

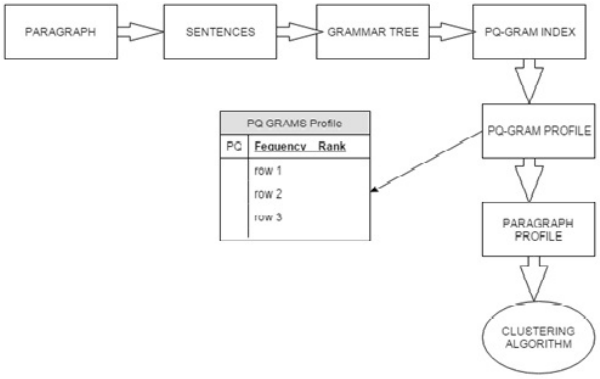
differences of the grammatical writing styles which the earlier baseline model was lacking and use this information to build paragraph clusters. So, by doing this, what kind of sentences can we decompose? An example shown below illustrates this. Consider the two sentences below:
S1: My chair started squeaking a few days ago and its driving me nuts.
S2: Since a few days my chair is squeaking-its simply annoying.

Figure 3: Different parses for similar meaning sentence



The above sentences are semantically similar, although they differ way too much syntactically(as shown in the figure below) and a bag of words model, which just relies on the occurrences of the words/word counts cant distinguish between these two sentences as they have more or less similar kinds of words. The main idea is to quantify those differences by calculating grammar profiles and to use this information to decompose a collaboratively written document.

Figure 4: The overview of our Method



As seen from the flow chart above, paragraphs are extracted from text and each sentence is extracted from the paragraphs. For each sentences, a parse tree is formed using the standard StanfordParser. We call this the Grammar tree. From this Grammar tree, we extract the PQ Gram indices of these sentences. p-q gram index of a sentence is all possible p-q grams of a sentence , whereby multiple occurrences of the same p-q grams are also present multiple times in the index. By combining all p-q gram indices of all sentences, a p-q gram profile is created which contains a list of all p-q grams and their corresponding frequency of appearance in the text. For our experiment, we have used p=2 and q=3. Finally, each paragraph-profile is provided as input for clustering algorithm, which are asked to build clusters based on the p-q grams contained. Also the labels are POS tags of Penn Treebank. We have not used the head words as we want to capture just the structure of the sentence and not the choice of words used by each authors
Another features we tried is the POS tags of each sentences in a bigram, trigram setting build a paragraph profile which gets input to a clustering algorithm (GMM-EM).

In the second aspect

- Calculate the similarity between each sentences in a segment by counting the number of common PQ-grams divided by the multiplication of the total PQ grams of each sentences and obtain a score.

- Do this for all the sentences and sum all the scores to get a similarity score of the segment, repeated over for all the segments.

- We now have a relative measure of the purity of the segments where better segment scores means more pure or in simple words, the mixing is biased towards one of the authors.

- This will give us a indication about which segments to use for clustering algorithm and leave the evenly distributed mixed segments with low similarity scores out of the clustering process.

- Train the GMM on the pure segments and leave the mixed segments out.

- We use the posterior of the clustered segments to identify the class of the left out segments.

- Identify vital segments as before and train using Naive Bayes.

- This step of bypassing some of the segments from GMM will enhance the selection of vital segments in the baseline system.

Figure 5:

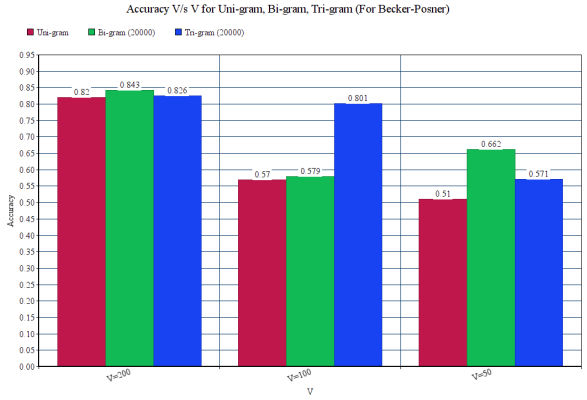
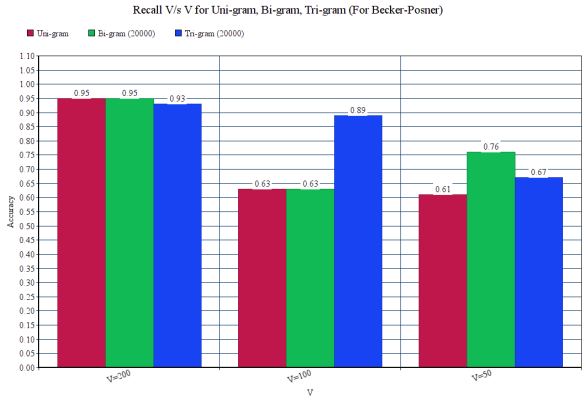


Figure 6:



Results

First Aspect

Table 2: Becker-Posner

Features	Accuracy		
	V=200	V=100	V=50
Baseline	0.82	0.57	0.51
POS-tags(bigram)	0.67	0.64	0.63
POS-tags(trigram)	0.70	0.67	0.65
PQ-GRAMS	0.66	0.63	0.62
PQ-GRAMS+tf-idf	0.69	0.65	0.63

Analysis

When applied to the data set Becker-Posner dataset (26922 sentences), we encountered many long sentences which the parser was not able to parse(out of memory). So we ignored those sentences (only 3) and evaluated the above strategies on this reduced data. Our observation from the above experiment is that, intuitively its a good method to capture the different syntactic aspects of writers. Although, it pushes the dimensionality of the feature space to quite high compared to the baseline.The baseline methods feature size were much smaller and simpler and faster. Also we could see that introducing these features are not exactly increasing the final accuracy of the baseline model.
We could see that the sensitivity of the model with new features even after reducing the chunk size(V) is quite robust. Its not varying drastically as in the case with baseline system. So we can conclude that these set of syntactic features are very stable.

Future Scope

The whole assumption of this investigation was that we know apriori about the number of authors present in the document. Future works can be on the area of determining the number of authors automatically using different clustering algorithms and study the effect of these features on the discriminating properties of those classifiers.

References

(i) A generic unsupervised method for decomposing multi-author documents. Navot Akiva and Moshe Koppel.2013. *(ii) Journal of the American Society for information Science and Technology,* 64: 2256–2264.
Navot Akiva and Moshe Koppel.2013. *Science* 208: 1019–1026.
*(iii) Unsupervised Decomposition of a Multi-Author Document Based on Naive-Bayesian Model.*Khaled Aldebei, Xiangjian He and Jie Yang. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 501505, Beijing, China
(iv) Automatic Decomposition of Multi-Author Documents Using Grammar Analysis,Michael Tschuggnall and Gnther Specht,Databases and Information Systems Institute of Computer Science, University of Innsbruck, Austria