Mixture of Experts Authorship Attribution
Notebook for PAN at CLEF 2012

Michael Ryan, John Noecker Jr

Evaluating Variations in Language Laboratory – Duquesne University
{mryan, jnoecker}@jgaap.com

Abstract. For problems A, B, C, D, I, and J we used three Authorship Attribution techniques; a distance based nearest neighbor, a svm, and method that used a distanced based NN approach to classify sections of a document and classifying based on who wrote majority of the document. These three techniques were then considered experts and each given a vote to determine the author of each document. For problems E and F we clustered paragraphs based on named entity use and then preformed authorship attribution on the non-clustered paragraphs.

1 Introduction

We will describe two different techniques we applied to generate solutions for the PAN 2012 Author Identification task, specifically the Authorship Attribution subtask. For our analysis, we used a number of our current best performing techniques on each problem, then combined them according to a weighted voting system. For problems E and F, we clustered paragraphs based on shared named entities, then used the technique from problems A, B, C, D, I and J to attribute previously unattributed paragraphs to the appropriate author

2 Materials and Methods

2.1 Corpus

We used the PAN 2012 Authorship Attribution corpus, a subtask of the Author Identification task. The corpus is available, in several parts, from the PAN 2012 website, currently hosted at *pan.webis.de*. It consists of 8 different problem sets. Problems A, C and I are standard, closed-class authorship identification tasks. Problems B, D and J are open-class identification tasks drawn from the same three groups of candidate authors.

Problems E and F were authorship diarization tasks, wherein an unknown number of candidate authors' writings were commingled, by paragraph, into a single document. The task was then to identify how many authors were present in the sample, and to which author each paragraph belonged.

2.2 Authorship Attribution

The authorship attribution task included problems A, B, C, D, I and J. As previously mentioned, A, C and I were closed-class problems, while B, D and J were the corresponding open-class problems, with the training data drawn from the same candidate author set.

We used the Java Graphical Authorship Attribution Program (JGAAP) for the analysis. JGAAP follows a multi-step paradigm for closed-class authorship attribution, as follows:

1. *Canonicizers* – Preprocessing is done to the documents, for instance to remove punctuation or remove letter case distinctions.
2. *Events* – Documents are broken down into features, such as characters or words.
3. *Analysis* – Various nearest-neighbor distance-based analysis or classification methods (e.g. SVMs) can be run on the data, with JGAAP returning the most likely candidate author for each document based on the result.

Typically we combined the results of the following three methods:

2.2.1 Centroid L1

This method used no canonicizers and used character 3-gram events. The character 3-grams use a sliding window where each event is a group of 3 adjacent characters. For the analysis stage, we used a centroid-based author model with a nearest-neighbor L_1 distance classifier.

2.2.2 Weka SMO

This method also used no canonicizers and character 3-gram events. Here, we used a Weka sequential minimal optimization trained support vector machine classifier.

2.2.3 Repeated Microdocument Analysis

For this method, we used Normalize Whitespace and Unify Case canonicizers, with character 11-grams as events. Each document was split into chunks that were roughly 3,000 characters in size, both for training and testing documents. The model was generated using a centroid-based intersection distance on the split training data. Each test document split was then classified, and the overall document was assigned authorship by a "voting" system. For open-class problems, an answer of "none of the above" was given if there was no clear majority author.

2.3 Authorship Diarization

For problems E and F, a different approach was used, which we will describe as "authorship diarization". For these problems, each document contained samples of writing by an unknown number of different authors, where each paragraph came from a single author. The goal here was to group the paragraphs by their author

For this task, we clustered paragraphs based on their shared named entity usage. In this way, we identified a first set of candidate authors (For instance, we might generate a set of documents with characters named "Bob" and "Sue", another set with characters named "Zebulon" and Xyzyyz" and a third set with "John" and "Michael"). Documents without shared named entity usage were analyzed later.

After the possible candidate authors had been identified, the repeated microdocument analysis technique from 2.2.3 was used both to consolidate authors and to assign previously unassigned paragraphs to their appropriate author (e.g. perhaps Authors 1 and 3 are actually the same author but from different novels – in this step we would consolidate both into a single author). Instead of using 3,000 character chunks, however, we kept the paragraphs whole.

This step was done repeatedly, training on the entire identified corpus and testing on all yet-unidentified paragraphs. At each step, the paragraph identified with the highest confidence was then tagged and added to the training step for the next iteration. This process continued until all paragraphs were identified.

3 Results

Table 1, below, gives the total number of correctly identified documents, the number of total documents, and the overall accuracy for each problem.

The official PAN results give two scores, an "overall percentage" and a "documents correct" score. The overall percentage is given as the mean accuracy across the problem sets. The documents correct score is the total number of documents correctly identified in all of the problems. Table 1 thus gives the documents correct score.

Problem	Number Correct	Total Number	Accuracy
A	6	6	100%
B	7	10	70%
C	7	8	87.5%

D	10	17	58.8%
E	83	90	92.2%
F	77	80	96.3%
I	12	14	85.7%
J	12	16	75.0%
Total	214	241	88.8%

Table 1: Results

4 Discussion

Observe that the overall by-document accuracy was 88.8%. This was the highest document accuracy in the PAN 2012 results. The mean accuracy was 83.2%, which places the results third in the mean accuracy portion of the competition.

One particularly interesting thing to consider is that our system is not tuned for open-class attribution problems. A stopgap solution was put together for PAN 2012, but this was the first time we have attempted to apply these techniques to an open-class problem. Removing the "none of the above" problems from the set results in a document accuracy of 91.6% and a mean accuracy of 88.5%.

We thus propose that the most important future work here is to better adapt the methodology to work on open-class problems. We have such a system in the works, based on a larger number of "experts" in the analysis voting, and hope to have refined it for the next iteration of the PAN competition.