News Authorship Identification with Deep Learning

Liuyu Zhou Huafei Wang
Electrical Engineering Aeronautics & Astronautics
Stanford University
lyzhou@stanford.edu huafei@stanford.edu

Abstract

Authorship identification identifies the most possible author from a group of candidate authors for academic articles, news, emails and forum messages. It can be applied to find the original author of an uncited article, to detect plagiarism and to classify spam / non- spam messages. In this project, we tackled this classification task in author level, article level, sentence level and word level with various deep and non-deep classification algorithms and GloVe word vectors are used as the pre-trained word vectors. Among all the algorithms, sentence-level Recurrent Neural Network (RNN) achieves the best performance since it captures the context information as well as word / sentence sequence information from the training dataset.

1 Introduction

Authorship identification (authorship attribution) determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author. Author Identification study is useful to identify the most plausible authors and to find evidences to support the conclusion. It can be applied in many tasks such as authorship characterization, detecting plagiarism, cybercriminal analysis and classifying spam / non-spam messages. This is a highly interdisciplinary area as it takes advantage of machine learning, information retrieval, and natural language processing.

Authorship identification problem has been studied in the last few decades with various of methods and techniques. Most previous studies such as [1,2,3,4] used stylometric analysis techniques for analyzing and attributing authorship of literary texts. Stylistic features are used in the stylometric analysis which include attributes or writing-style markers that are the most effective discriminators of authorship. The vast array of stylistic features includes lexical, syntactic, structural, content-specific, and idiosyncratic style markers [1].

Over 1,000 different features have been used in previous authorship analysis research, with no consensus on a best set of style markers. This could be attributable to certain feature categories being more effective at capturing style variations in different contexts. This necessitates the use of larger feature sets comprised of several categories of features (e.g., punctuation, word-length distributions, etc.) spanning various feature groups (i.e., lexical, syntactic, etc.) [1].

As a result, the performance of authorship identification tasks depends highly on the chosen features and the quality of these features. Therefore reliable and efficient techniques are needed to extract these features.

For learning, we have both supervised learning and unsupervised learning at our disposal. Supervised learning are those that require author-class labels for classification, while unsupervised techniques make classification with no prior knowledge of author classes. In this paper, we only focus on the supervised learning paradigm.

Supervised learning methods used in previous studies include regularized least squares, support vector machine (SVM), decision tree, feed-forward neural network and etc. on various datasets (Twitter, RCV1, PAN 2012 and etc.). The accuracy varies significantly when different approaches and datasets are used from low 20s to high 90s.

This paper researches the news authorship identification problem by exploring different machine learning algorithms on different levels

- Article-level linguistic features represented by average word length and etc. with non-deep machine learning algorithms (e.g. support vector machine).
- Word-level features represented by word vectors with Global Vectors for Word Representation (GloVe) with different classifiers (e.g. nearest neighbor).
- Recurrent Neural Network (RNN) with pre-trained GloVe word vectors. A sentence vector is generated as the input in each step. The number of steps is the number of sentences in the article.

GloVe pre-trained word vectors are used since GloVe captures the word context information (such as word similarity). We mainly focus on Recurrent Neural Network (RNN) on GloVe word vectors since we believe that RNN can also capture the word / sentence sequence information which can help us to better classifiy the authorship of articles. The experiment results confirm this assumption.

We focus on news articles since news articles provide not only stylometry features as in other types of texts, but also context information since journalists tend to focus on a narrow range of topics and thus context can also be leveraged for the identification purpose.

2 Approach

2.1 Dataset

A subset of the Reuters RCV1 news article dataset is used to develop our multi-level machine learning algorithms. The dataset contains 5000 news articles for 50 different journalism authors (100 texts per author) and pre-splits the dataset 50-50 for training and testing. The dataset is obtained from the Center for Machine Learning and Intelligent Systems hosted by the Unviersity of California, Irvine.

2.2 Baseline

High-level linguistic features of articles are used as attributes in our baseline model. The following stylometry features are used: Average Word Length, Average Sentence Length, Hapax Legomenon Ratio (fraction of unique words).

Traditional machine learning methods including Support Vector Machines, Naive Bayes, Random Forest and etc. are implemented utilizing the above features.

2.3 GloVe

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space [5].

In this project, GloVe pre-trained word vectors are used to encode the training articles and find the most representative vector that can represent the author. Then the test articles are encoded using the same GloVe word vectors to obtain the most representative vector that can represent the article. Different GloVe word vector dimensions are used to evaluate their influence on the performance of the algorithm. F1 Score is used for evaluation.

2.4 Recurrent Neural Network

Recurrent Neural Network (RNN) is also implemented to tackle this task. GloVe pre-trained word vectors are used to encode each word in the article. The standard model - sigmoid hidden layer and softmax projection layer are used, which are as follows.

$$
\begin{aligned}
e^{(t)} &= x^{(t)} L & (1)\\
h^{(t)} &= sigmoid\left(h^{(t-1)} H + e^{(t)} I + b_1\right) & (2)\\
\hat{y}^{(t)} &= softmax\left(h^{(t)} U + b_2\right) & (3)
\end{aligned}
$$

The input vector in each step is a pre-trained constant vector which can represent a word, a sentence, a paragraph or even an article, which will be discussed in detail in the next section.

A more complex RNN model – LSTM / GRU with forget gate, input gate and output gate is also implemented in our model. This approach is adopted since we believe that some words in the article are not representative to the author so thay may pose no effect or even negative effect on the prediction results while some words are truely indicative of the author and they should be assigned higher weights.

Since the size of our dataset is moderate, dropout is applied to the input and output of the RNN model in order to mitigate the undesirable overfitting.

3 Experiment

3.1 Baseline

Several machine learning methods including Support Vector Machines, Naive Bayes, Random Forestd using the stylometry features are implemented. The best accuracy is achieved by Gradient Boosting Classifier with an **12.24%** accuracy.

The low accuracy is expected as only 3 high level summary features are used. As discussed in the Introduction, over 1000 different stylometry features are used in previous studies in order to have a satisfactory classification performance.

In addition, word, sentence and paragraph contexts as well as sequences that may contain author writing style and pattern information are discarded during the summarization and feature extraction process.

This preliminary result serves as a baseline for this project.

3.2 GloVe

The F1 Score and confusion matrix visualizations with the Nearest Neighbor Classifier (L_2 Distance) are shown below. We can observe that the classification with GloVe greatly outperforms the baseline methods because of the different context information associated with each author that is able to be extracted and aggregated from their articles using GloVe.

It can also be observed that the higher the word dimension is, the higher the F_1 score the result has. The highest F_1 score of the Nearest Neighbor Classifier comes with GloVe word vectors with dimension 300, which is around **0.46**. This is reasonable since higher dimension word vector can generally capture more context information of a word.

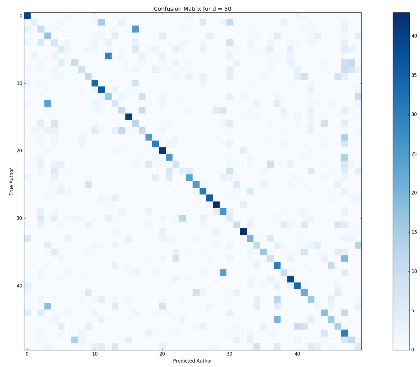

Figure 1: Confusion Matrix with GloVe D = 50

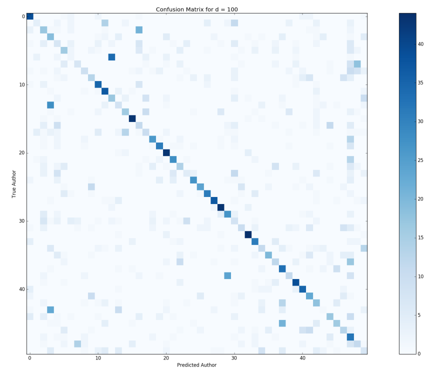

Figure 2: Confusion Matrix with GloVe D = 100

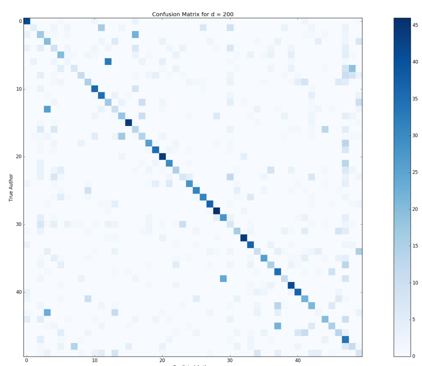

Figure 3: Confusion Matrix with GloVe D = 200

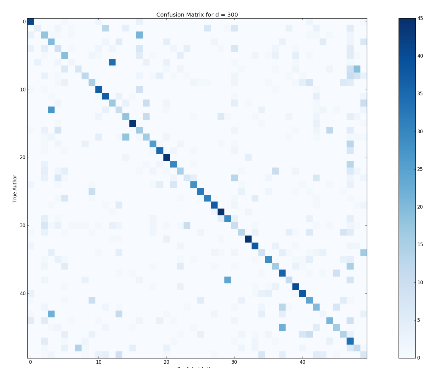

Figure 4: Confusion Matrix with GloVe D = 300

3.3 RNN

The following two approaches of data input are implemented for RNN

- Word level approach. A single word vector is fed into RNN in one step. So the number of steps of RNN is the number of words in each article. Since there are typically a large number of words in each article, there will be too many steps, which has caused the gradient vanishing problem. The result we have obtained is poor, with the highest F_1 score around **0.2**. Besides, since each article has a different length, different models for each article have to be built separately. As a result, the training is undesirably slow. Paddings have been added to the articles to make them have the same length. But unfortunately, since the length of articles covers a wide range - from 200 to 1500, adding paddings does not enhance the performance significantly.
- Sentence level approach. The mean vector of each sentence is calculated and fed into the RNN model at each step. This approach also captures the word / sentence sequence information since the RNN process is recurrent. Because that the average number of sentences of all articles is around 15-20, this model does not suffer from the gradient valishing problem, so it significantly outperforms the word level approach. The highest F_1 score we have obtained is around **0.6**. In addition, the training is relatively fast compared with the word level RNN approach.

Confusion matrix visualizations of sentence level RNN with different GloVe word vector dimensions are shown below.

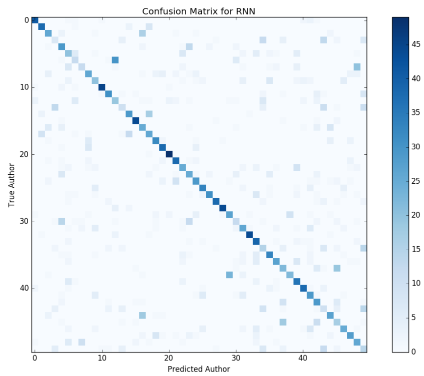

Figure 5: Confusion Matrix for RNN (D = 50)

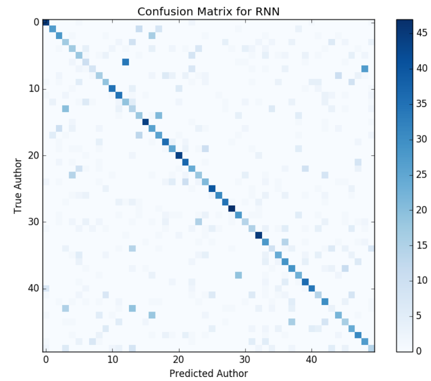

Figure 6: Confusion Matrix for RNN (D = 100)

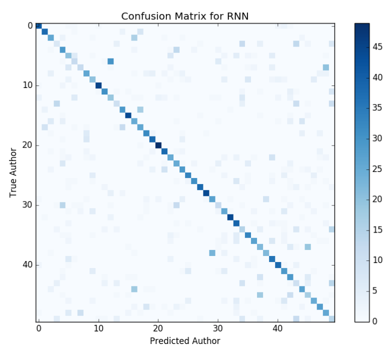

Figure 7: Confusion Matrix for RNN (D = 200)

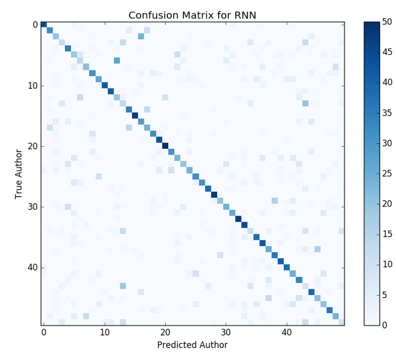

Figure 8: Confusion Matrix for RNN (D = 300)

We can see from the figures that the sentence level RNN Classifier with GloVe word vectors of dimension 300 has the best performance, with the F_1 score around **0.6**.

The training loss and validation loss values of sentence level RNN are shown below

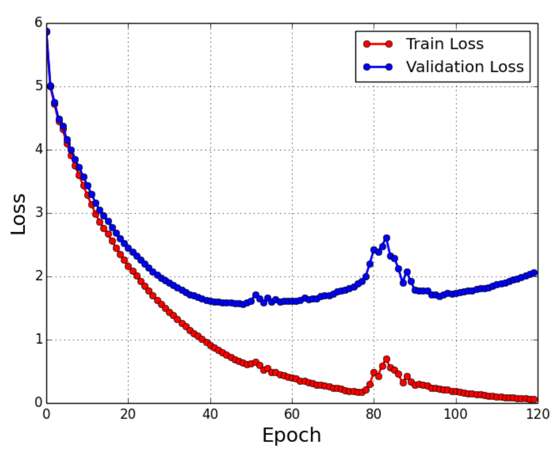

Figure 9: Loss History for RNN

It can be observed from the figure that despite some small spikes in the curve, the training loss value keeps decreasing. The validation loss value drops to a minimum value and then starts to increase.

Typically we can get the best model for prediction (test accuracy) when the validation loss reaches the minimum value. Usually the training algorithm will terminate at the first small spike of the validation loss curve. But in order to show the full trend of the training and validation loss, more epoches have been run in our attempt.

3.4 Comparing with Other Machine Learning Algorithms

Several other machine learning algorithms that are typically powerful in various of prediction tasks are also implemented. These algorithms include Random Forest Classifier, Extra Tree Classifier, Decision Tree Classifier, AdaBoost Classifier, and Gradient Boosting Classifier. For these methods, GloVe word vectors are used to encode the articles and calculate the mean vector for each article and predictions are based on the calculated mean vector. The comparison results of different machine learning algorithms are shown below.

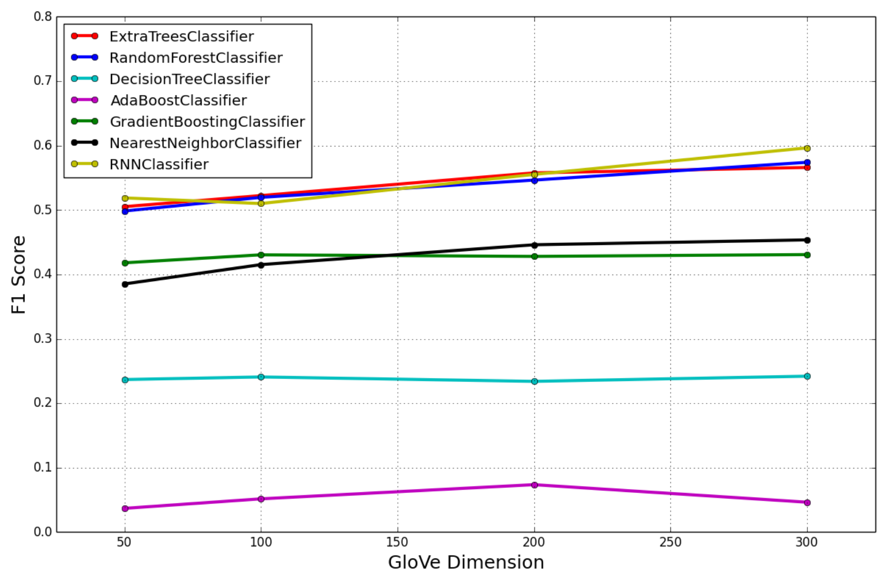

Figure 10: Different ML Methods with GloVe

4 Conclusion

We have presented various methods for authorship identification task with a focus on the Recurrent Neural Network approach. Stylometry features lose information and performs poorly in the authorship identification task. GloVe captures context information for different authors and performs relatively better. RNN, in addition to capturing context information, also captures word / sentence sequence information and has the best performance with a **0.6** F_1 score.

However, the RNN approach does not show significant advantage over other machine learning methods. For future work, we can attempt to improve the performance of RNN to truely unleash its power. Some strategies are as follows.

- Use a larger dataset. This is extremely important since our model is already suffering from overfitting problem and a larger dataset can help us remedy the situation.
- Currently, we are using GloVe pre-trained vectors as our constant word vectors in our model. We can let the GloVe pre-trained vectors be the initial values for the word embedding variables and update them in each RNN step.
- For large enough dataset, we can use neural network to train our own word vectors instead of using the pre-trained GloVe word vectors.

- The more complex RNN models such as stacked RNN, bidirectional RNN, inductive transfer can be applied with a corresponding large dataset.

References

[1] Abbasi A, Chen H. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace[J]. ACM Transactions on Information Systems (TOIS), 2008, 26(2): 7.

[2] Castro A, Lindauer B. Author Identification on Twitter[J]. 2012.

[3] Narayanan A, Paskov H, Gong N Z, et al. On the feasibility of internet-scale author identification[C]. Security and Privacy (SP), 2012 IEEE Symposium on. IEEE, 2012: 300-314.

[4] Nirkhi S, Dharaskar R V. Comparative study of authorship identification techniques for cyber forensics analysis[J]. arXiv preprint arXiv:1401.6118, 2013.

[5] Pennington J, Socher R, Manning C D. Glove: Global Vectors for Word Representation[C]. EMNLP. 2014, 14: 1532-1543.