Final Report - Author Identification in English and Chinese Text

Introduction

Author identification is the task to determine the author of a piece of text. As we are now in the age of internet, there is a demand for identifying the author of anonymous post on internet for forensic, legal or journalistic reasons. Thus I find author identification to be a topic worth exploring. My preliminary search suggests that there are not that many studies done on authorship attribution of Chinese text. Therefore, in this project I am interested in applying the same identification methods on both English and Chinese text to see if they are equally effective on both text.

Related work

Author identification has long been a topic of interest. Among the first statistical computation studies was the study on the authorship of the Federalist Papers. For the current project, Digamberrao and Prasad's analysis[1] of author identification was an inspiration.

Approach and Experimental setup

The English dataset is provided on Kaggle[2] under the name "spooky author identification". The dataset contains sentences extracted from the work of three authors, Edgar Allan Poe (EAP), H.P. Lovecraft (HPL), and Mary Shelley (MWS). The dataset contains 19,579 entries in total. The Chinese dataset is built from text files of works of three authors, Lu Xun (LX), Lao She (LS) and Yu Dafu (YD). The dataset contains 17,812

---

[1] Digamberrao, Kale Sunil, and Rajesh S. Prasad. "Author Identification on Literature in Different Languages: A Systematic Survey." *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, 2018. https://doi.org/10.1109/icacct.2018.8529635.

[2] https://www.kaggle.com/c/spooky-author-identification/data

entries in total. For both the datasets, 90% was used for training and 10% was used for testing.

Three features were used for analysis. Tf-idf, word count and word vector. Td-idf and word count feature extraction was implemented using TfidfVectorizer and CountVectorizer from the sklearn library. For each of the feature extraction method, both logistic regression classifier and naive bayes classifier were tested. English word vectors were calculated from the word embedding from GloVe[3]. The Chinese word embedding uses a literature corpus and is available on GitHub[4]. The Chinese text was analyzed on character level with no word segmentation, as study has shown that character-level analysis can perform reasonably well on author identification tasks[5]. For the word vector feature, logistic regression classifier and XGBoost classifier were tested.

Results

English dataset: TF-IDF With Logistic Regression Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
EAP	0.81	0.87	0.84	0.533	83.2%
HPL	0.87	0.80	0.83		
MWS	0.84	0.81	0.83		

English dataset: TF-IDF With Naive Bayes Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
EAP	0.78	0.92	0.84	0.562	84.1%
HPL	0.91	0.76	0.82		

[3] https://nlp.stanford.edu/projects/glove/

[4] https://github.com/Embedding/Chinese-Word-Vectors

[5] Peng, Fuchun, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. "Language Independent Authorship Attribution Using Character Level Language Models." *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - EACL 03*, 2003. https://doi.org/10.3115/1067807.1067843.

MWS	0.89	0.82	0.85		

English dataset: Word Count With Logistic Regression Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
EAP	0.79	0.88	0.83	0.445	82.1%
HPL	0.86	0.77	0.81		
MWS	0.83	0.80	0.81		

English dataset: Word Count With Naive Bayes Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
EAP	0.86	0.85	0.86	0.636	85.5%
HPL	0.90	0.83	0.86		
MWS	0.82	0.89	0.85		

English dataset: Word Vector With Logistic Regression Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
EAP	0.73	0.73	0.73	0.692	72.1%
HPL	0.72	0.72	0.72		
MWS	0.72	0.71	0.71		

English dataset: Word Vector With XGBoost

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
EAP	0.76	0.75	0.76	0.610	74.5%
HPL	0.75	0.76	0.76		
MWS	0.72	0.72	0.72		

Chinese dataset: TF-IDF With Logistic Regression Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
LS	0.95	0.93	0.94	0.362	92.3%
LX	0.92	0.90	0.91		
YD	0.90	0.93	0.92		

Chinese dataset: TF-IDF Naive Bayes Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
LS	0.97	0.94	0.96	0.360	92.3%
LX	0.93	0.88	0.91		
YD	0.87	0.95	0.91		

Chinese dataset: Word Count With Logistic Regression Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
LS	0.94	0.93	0.94	0.225	91.7%
LX	0.91	0.90	0.90		
YD	0.90	0.92	0.91		

Chinese dataset: Word Count With Naive Bayes Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
LS	0.96	0.94	0.95	0.496	92.8%
LX	0.94	0.90	0.92		
YD	0.88	0.94	0.91		

Chinese dataset: Word Vector With Logistic Regression Classifier

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
EAP	0.75	0.78	0.76	0.690	72.3%
HPL	0.71	0.67	0.69		
MWS	0.73	0.74	0.73		

Chinese dataset: Word Vector With XGBoost

	Precision	Recall	f1	Multi-class log loss	Overall Accuracy
EAP	0.79	0.83	0.81	0.547	78.0%
HPL	0.76	0.74	0.75		
MWS	0.78	0.77	0.78		

Discussion and Conclusion

As shown ins the results, classification based on word vector performs significantly worse than those based on tf-idf and word count. Before the analysis of the Chinese dataset, I hypothesised that the poorer performance of word embedding on English dataset is due to a large amount of out of vocabulary words (about 3000 in the training set). However, that seems to not be the case, as the number of out of vocabulary characters in the Chinese dataset is very low (about 200 in the training set). Thus it may be concluded that word vector is not an informative feature for author identification compared to tf-idf and word count. It should also be noted that word vector analysis is much more time consuming.

Surprisingly, with almost no modification, all of the models performs better on Chinese text than on English text. The reason for the better performance is unclear.

From the results we can also see that naive bayes tends to be a better classifier than logistic regression, though the difference is not that significant and is more prominent on English text than on Chinese text.

When adjusting the models, I have also realized that excluding stopwords from analysis actually tends to decrease classification accuracy. This suggests that the pattern of stopwords usages is also informative about the authorship and should not be disregarded.

In conclusion, tf-idf and word count very informative features when it comes to author identification, and are effective for both English and Chineses text. Word vector, on the other hand, is not ideal. Naive bayes is a better choice as a classifier than logistic regression, but both are decent.