# Bilingual Tweets Authorship Attribution

Anonymous TACL submission

## Abstract

This document attempted to tackle the authorship attribution (AA) problem across different languages. Focusing on Chinese and English, I have examined the possibilities of three attribution models, especially an aligned word embedding model. The final result showed that even though it didn't surpass the other two models, there is possibility for the aligned word embedding model to solve the cross-language authorship attribution (CLAA) problem.

## 1 Introduction

In recent years, political propaganda has moved a significant amount of resources onto the social media in additional to the tranditional mass media. This has created both positive and negative influences to the crowd. Since then, there has been many attempts on applying NLP techniques to counter-attack the negative affect caused by manufactured information published on the social media. Most of them analyzed stylometric features on mono-language data and showed promising potential in this field (Rocha et al., 2016).

In June 2019, the proposed anti-extradition law in Hong Kong had attracted great controversy on the social media. Political propoganda was so severe that Twitter had to suspend about 1000 twitter accounts violating their platform manipulation policies[1]. A brief study by Wood et al. (2019) had revealed that the languages used in these tweets contents spaned across several languages (but mostly in Chinese and English). As propoganda and manufactured information are now reaching out more people around the world by using multiple language, there is the need to develop AA techniques for cross-language social media texts. In this project, I have decided to try the possibilities for bilingual authorship attribution — chosing English and Chinese as my experiment languages, by applying both machine translation and aligned word embedding models.

The rest of the paper is structured as follows: Section 2 descirbes related works in general AA and CLAA problems. Section 3 explains how the dataset is built as well as the three models I have designed, while Section 4 showed and evaluated their corrsponding results. Finally in Section 5 I summarized the whole project and looked at possible future works for CLAA.

## 2 Related Works

CLAA has less attention compared to AA on monolingual languages. Bogdanova and Lazaridou (2014) explored the possbility of applying machine translation to connect two languages, in combination of tranditional stylometric features such as word-level and char-level n-grams. There are also researches attempted to tackle the problem without bridging different languages by their semantics at all, such as Llorens and Delany (2016) and Sarwar et al. (2018) who both analyzed low level language independent features.

However, in all of these previous works, the languages used in their datasets are all Indo-European Lanuages. In my case, the relationship between Chinese and English are much further than language pairs like Spanish and English. There are research done for distant langauge pairs in cross-language plagiarm detection (Barrón-Cedeno et al., 2010), but to my best knowledge no similar work for CLAA has been released.

Besides datasets, most AA tasks employed machine learning techniques as the classification al-

[1] https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.html

| Chinese | English |
|---|---|
| 是 | am, is, are |
| 的 | of |
| 有 | have, has |
| 在 | at, in, on |

Table 1: Function Words Used in the Query String

| # of tweets in total | 69710 |
|---|---|
| ZH tweets | 27476 |
| EN tweets | 37604 |
| Other lang. tweets | 4630 |

Table 2: Overlook at the Dataset

gorithm, while there are few examples successfully applied Neural Networks to the same problem (Shrestha et al., 2017). This is reasonable as monolingual AA usually catches the quirkiness of spelling, spacing or word richness of a specific writer. While in CLAA, especially for distant language pairs, there might be no directly comparable features betwenn languages. For example in Chinese and English, it is impossible to compare the pattern of misspelling and spacing because they are not existed in Chinese. Therefore we might need to go back to the semantic level to look for similarities between texts to find out if they came from the same author.

## 3 Methodology

### 3.1 Dataset

Since there is currently no publicly available corpus focused in social media texts in both English and Chinese, I have built my own using Twitter Public API.

The first step is to build a query string to collect as many tweets in either English or Chinese as possible. I selected several most frequent function words in Chinese and their English counter parts, which are listed in Table 1. Suppose $C = \{c_1, c_2, c_3, \cdots, c_i\}$ is a group of Chinese function words and $E = \{e_1, e_2, e_3, \cdots, e_j\}$ is the corrsponding English function words, the query string $Q$ would be union set over the Cartesian product $C \times E$.

$$Q = \bigcup_{k=1} s_k, s_k \in C \times E \tag{1}$$

So if we select the first row in Table 1, then our Chinese cadidate function word candidate would be "是" while the English candidates are "am", "is" and "are". The final query string would be

```
(是 am) OR (是 is) OR (是 are)
```

The next step is to fine grain all of the possible twitter users from the previous step. Here I define two bilingual ratio values — let $t_1$ and $t_2$ denote two sets of the tweets in any of the two language written by a user, $T$ is the set of all of his/her tweets. Then we have $R_i$ as the ratio between these two tweets languages.

$$R_i = \frac{\min(|t_1|, |t_2|)}{\max(|t_1|, |t_2|)} \tag{2}$$

And $R_T$ as the ratio between the bilingual content and the total tweets.

$$R_T = \frac{|t_1| + |t_2|}{|\mathrm{T}|} \tag{3}$$

For each user, I crawled his or her first 200 tweets from his or her timeline (exclude retweets). The threshold were set at $R_i \geq 0.5$ and $R_T \geq 0.8$. In this way I could filter out two kinds of users — someone who occasionally tweets in another language than his/her main language, and someone who tweets in all kinds of languages.

After searching and fine graining I have selected 52 valid bilingual twitter accounts. Two of them are non-personal accounts so I have removed them from the list, which shortened the length of the final list of bilingual twitter users to 50. From them I crawled all of their tweets and applied cleaning to these data, including removing URLs, hashtags, mentions, reserved words, emojis and smileys. The CNN classifier cannot take tweets that are shorter then 5 words so I have also segmented both English and Chinese tweets and removed those that are shorter than 5 segments. In the end, I have obtained 69710 tweets that are detailed in Table 2 and 3. Also, each model is evaluated by 10-fold cross validation, except for the aligned word embedding model which was trained and validated in a fixed training and validataion dataset derived from the full dataset by the ratio of 0.6, 0.1 and 0.3.

### 3.2 Vanilla Model

As Rocha et al. (2016) had showned in their work, the best stylometric features for AA on social media text are word-level and char-level n-grams. I have designed a vanilla attribution model with

| | # of tweets/user | Length of raw tweets | Length of ZH tweets | length of EN tweets |
|---|---|---|---|---|
| mean | 1383.4 | 100.309 | 37.664 | 72.59 |
| std | 1211.126 | 38.24 | 29.142 | 31.04 |
| min | 102 | 5 | 5 | 6 |
| 25% | 334.5 | 68 | 16 | 47 |
| 50% | 856 | 109 | 27 | 74 |
| 74% | 2306.75 | 140 | 51 | 100 |
| max | 4195 | 159 | 140 | 146 |

Table 3: Distribution of the Dataset

word-level 1, 2, 3-grams and char-level 1, 2, 3-grams, which were then feeded into a logistic regression classifier and a LIBLINEAR SVM classifier (Fan et al., 2008) implemented by Scikit-learn (Pedregosa et al., 2011). The word-level and char-level features were counted from all available tweets without distinguishing the language by their appearance and tranformed into TF-IDF values.

We have seen from Table 3 that the number of tweets from each user is highly unbalanced. The most frequent user tweets 40 times more than the most quiet user. So I have set both of the logistic regression and SVM classifier to balance out the dataset automatically by assigning more weight to minor classes (users).

### 3.3 Machine Translation Model

I have also adapted machine translation followed by Bogdanova and Lazaridou (2014) to tackle the CLAA problem. I have used the Translator Text API from Microsoft Azure Cognitive Services, manully specifying the source language and the target langauge. One thing to note is that even the state-of-art machine translation service is far from perfect and underperforms on social media texts than formal writings. In other words it will inevitably introduce "distortion" to the raw tweets and worsen the result in theory, though machine translation is one of the most intuitive solution to this multi-language task.

In this second model, I have extracted and divided the raw tweets into the English group and the Chinese group. For tweets that mix both languages I seperated them into these two monolingual groups. Then tweets in each language group will be machine translated into the other language before being feeded into the aforementioned logistic regression and SVM classifier, together to be classified among other translated tweets within the same language group.

### 3.4 Aligned Word Embedding Model

The last model I have applied is a Convolutional Neural Network classifier inspired by Shrestha et al. (2017). They have proposed an architecture using char n-grams models as the single embedding layer to deal with tweets classification. But since my task is to explore the classification problem between two languages, I have switched to the aligned fastText word embeddings (Bojanowski et al., 2017) (Joulin et al., 2018) as my embedding layer.

In the aligned fastText word embeddings, each word is represented by a 300 dimension vector. I concatenated two embeddings to form a bunch of 600 dimension word embeddings for the possible bilingual vocabularies, padded them and send them to the next layer. Each OOV words are marked as `<UNK>` and are giving an embedding of zeros.

The CNN network also has four convolutional layers, each of them has 100 kernels sizing from 2, 3, 4 or 5. They are designed to catch the information hidden inside the word bigram. trigram and quadgrams before being max-pooled. I have also used Adam optimizer (Kingma and Ba, 2014) and all of the hyperparameters are listed in Table 4.

## 4 Results

### 4.1 Best Features

For the first two models I have applied grid search to find the best stylometric feature for bilingual AA. As we see in Table 5, the best features are usually the shorter word unigrams or bigrams for each classifier. Only in the translated Chinese group the best results appeared in the char-level bigrams and trigrams. But this can be explained by the fact that in Chinese the average word length

| Hyperparameters | Value |
|---|---|
| # of embedding layers | 1 |
| dimension | 600 |
| # of convolutional layers | 4 |
| kernel size | [2, 3, 4, 5] |
| # of kernels | 100/layer |
| pooling | max |
| Dropout | 0.5 |
| Learning rate | 0.001 |
| Max epochs | 10 |
| Batch size | 32 |

Table 4: Hyperparameter settings in the Aligned Word Embedding model

is about one to two characters while in English it is about four to five letters. (Chen et al., 2015) (Bochkarev et al., 2015). In other words, Chinese characters them alone can carry as much information as English words. The result that word-level n-grams are more effective than char-level is aligned with results from many previous works in many other AA tasks (Kestemont et al., 2018) (Rangel et al., 2019).

Also the SVM classifier outperformed the logistic regression classifier in nearly all comparisions, except in the char-level unigram one. This can be seen in Rocha et al. (2016)'s work as well. They attributed it to the application of Maxiumn Margin Princple in SVM classifiers. However for AA task on long articles, logistic regression could be better than SVM (Bogdanova and Lazaridou, 2014). Thus I think SVM is a better choice for short social media texts than logistic regression.

### 4.2 Distortion from Machine Translation

As mentioned previously, machine translation will inevitably introduction noise into the text and will bring down the accuracy. In Table 5 I have also calculated the impact of machine translation compared to the untranslated original text. Word-level bigrams have topped the chart in almost every group, followed closely by word-level trigrams and unigrams. The result once again showed that word-level n-grams are better features than char-level n-grams in our bilingual tweet dataset. Further more, LIBLINEAR SVM classifier still achieved higher accuracy than the logistic regression classifier after machine translation, showed that it is more suitable for short social media text no matter what language the text is in.

Move on to the performance difference between translated Chinese and English texts, we can see that Chinese suffered more than English if it been translated. Especially in the case of char 1-gram, when all Chinese text are translated into English the performance dropped more than 50%. In contrast while we turned all English content into Chinese, we have manageed to improve the performance by nearly 30%. Since many Chinese characters can serve as a word alone, short char-level n-grams in Chinese can be viewed as word-level n-grams in English. Thus explains why we had a large gain on performance when we translate everything into Chinese.

### 4.3 Aligned Word Embeddings

Finally the results for the aligned word embedding model are shown in Table 6. Because fastText only provides word embeddings for Tranditional Chinese, I have used OpenCC[2] to convert all the Simplified Chinese content to Tranditional Chinese. Furthermore I have also added a reference group using the unaligned common fastText embeddings. Both the aligned and the unaligned word embeddings for Chinese and English were pre-trained on Wikipedia text (Bojanowski et al., 2017).

The aligned word embedding model didn't surpass my vanilla model, however it is much closer to it than the machine translation model. Also the performance gain by using bilingual embeddings was subtle compared to monolingual embeddings, only around 2%. The unaligned model performed worst among these three kinds of embeddings, which is expected.

The result for the aligned word embedding model might suggest a better combination for bilingual word embeddings rather than concatenation. But there are other facts that could also affect the final result. First, the fastText word embeddings were trained on a domian that is far away from social media text. Wikipedia is more formal, serious and comprehensive place than Twitter, and the topics it includes had little intersection with the topics from Tiwtter. The results could be improved by training a dedicated word embeddings from Twitter corpus. Another things to note is the imbalance between English and Chinese word embedding sizes. The English embeddings has 2519370 items and is 7.5 times larger than the Chi-

[2] https://github.com/BYVoid/OpenCC

| | LR | SVM | LR+MT(EN) | SVM+MT(EN) | LR+MT(ZH) | SVM+MT(ZH) |
|---|---|---|---|---|---|---|
| Word 1-gram | 0.659 | 0.714 | 0.533 (-19.1%) | 0.553 (-22.5%) | 0.614 (-6.8%) | 0.642 (-10.1%) |
| 2-gram | 0.648 | 0.744 | 0.543 (-16.2%) | 0.602 (-19.1%) | 0.612 (-5.6%) | 0.680 (-8.6%) |
| 3-gram | 0.620 | 0.733 | 0.523 (-15.6%) | 0.599 (-18.3%) | 0.595 (-4.0%) | 0.673 (-8.2%) |
| Char 1-gram | 0.452 | 0.450 | 0.216 (-52.2%) | 0.197 (-56.2%) | 0.575 (+27.2%) | 0.583 (+29.6%) |
| 2-gram | 0.592 | 0.655 | 0.415 (-29.9%) | 0.433 (-33.9%) | 0.622 (+5.1%) | 0.680 (+3.8%) |
| 3-gram | 0.637 | 0.723 | 0.500 (-21.5%) | 0.555 (-23.2%) | 0.615 (-3.5%) | 0.682 (-5.7%) |

Table 5: Results for the Vanilla Model and the Machine Translation Model

| Word Embeddings | Accuracy | Loss |
|---|---|---|
| Aligned Bilingual | 70.1% | 1.197 |
| ZH Only | 68.0% | 1.242 |
| EN Only | 69.0% | 1.195 |
| Unaligned Bilingual | 64.8% | 1.816 |

Table 6: Results for the Aligned Word Embedding Model

nese embeddings, just as the English Wikipedia articles are around 5 times more than articles in Chinese [3] [4]. Smaller embeddings size will introduce more OOV words and will lower the overall accuracy.

## 5 Conclusion

In this project I have explored the possibilities of three different model for a bilingual AA question. On a dataset collected from Twitter, the simple SVM classifier with word-level and char-level n-grams achieved the highest accuracy, followed by the aligned word embedding model. I have also discussed the distortion brought by machine translation. It turned out that even though the aligned word embeddings didn't give the best result, it has the potential to be one of the solutions to the CLAA problem as well.

In futher works, the emphisis should be on a more sophiscated assembling of bilingual word embeddings. In addition, world clusters (Täckström et al., 2012) is another promising model to transfer from one languages to another. Either of them avoids the inevitable distortion from machine translation, hence can be the new features for CLAA.

[3] As Dec 2018, https://stats.wikimedia.org/EN/TablesWikipediaEN.htm

[4] As Dec 2018, https://stats.wikimedia.org/EN/TablesWikipediaZH.htm

## References

Alberto Barrón-Cedeno, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 37–45. Association for Computational Linguistics.

Vladimir V Bochkarev, Anna V Shevlyakova, and Valery D Solovyev. 2015. The average word length dynamics as an indicator of cultural changes in society. *Social Evolution & History*, 14(2):153–175.

Dasha Bogdanova and Angeliki Lazaridou. 2014. Cross-language authorship attribution. In *LREC*, pages 2015–2020. Citeseer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Heng Chen, Junying Liang, and Haitao Liu. 2015. How does word length evolve in written chinese? *PloS one*, 10(9):e0138567.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018.

Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.*, pages 1–25.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Marisa Llorens and Sarah Jane Delany. 2016. Deep level lexical features for cross-lingual authorship attribution.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Francisco Rangel, Paolo Rosso, L Cappellato, N Ferro, H Müller, and D Losada. 2019. Overview of the 7th author profiling task at pan 2019: Bots and gender profiling. In *CLEF*.

Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. 2016. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33.

Raheem Sarwar, Qing Li, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. A scalable framework for cross-lingual authorship identification. *Information Sciences*, 465:323–339.

Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 477–487. Association for Computational Linguistics.

Daniel Wood, Sean McMinn, and Emily Feng. 2019. China used twitter to disrupt hong kong protests, but efforts began years earlier.