# Bilingual Tweets Authorship Attribution

Anonymous TACL submission

## Abstract

This document contains the formatting requirements for TACL submissions. These formatting rules take effect for all submissions received from September 2, 2018 onwards.

## 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, which has created both positive and negative effects in the crowd. Since then, there has been many researches on applying NLP techniques to counter-attack the manufactured information published on the social media. Most of them analyzed stylometric features on mono-language data and showed promising potential in this field such as 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 also had played a big role in this movement. In fact, it 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), which proposed challenges for cross-langauge authorship attribution on short social media texts.

Unlike the authorship attribution in English social network texts, cross-language authorship attribution has not been discovered extensively yet. However as propoganda now reaching out more people around the world by using multiple language, there is the need to develop authorship attribution techniques for cross-language social media texts. In this project, I have decided to try the possibilities for bilingual authorship attribution — focuing on English and Chinese, by applying both machine translation and aligned word embeddings.

[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

## 2 Methodology

### 2.1 Dataset

To my best knowledge there is currently no publicly available corpus focused in social media texts in both English and Chinese, hence I have decided to build 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 frequent function words in Chinese and their English counter parts, as shown in Table 1. Let's say $C = \{c_1, c_2, c_3, \cdots, c_i\}$ is a group of Chinese function words and $E = \{e_1, e_2, e_3, \cdots, e_j\}$ is their English siblings, the query string $Q$ would be union set over the Cartesian product $C \times E$.

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

For example if we select the first row in the table, then our Chinese cadidate function word is "是" while the English candidates are "am",

"is" and "are". Our 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 $L_1$ and $L_2$ denote two sets of the tweets in a certain language by a user, then we have $R_{inner}$ as the ratio between these two tweets languages.

$$R_{inner} = \frac{min(|L_1|, |L_2|)}{max(|L_1|, |L_2|)} \quad (2)$$

And $R_{overall}$ as the ratio between the bilingual content and the total tweets.

$$R_{overall} = \frac{|L_1| + |L_2|}{|\text{All tweets}|} \quad (3)$$

For each user, I crawled his or her first 200 tweets from his timeline (exclude retweets) and set the threshold at $R_{inner} \geq 0.5$ and $R_{overall} \geq 0.8$. 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 these two steps I have selected 52 valid bilingual twitter accounts. Two of them are obviously non-personnal users so I have removed them from the list, which made 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 not longer than 5 segments. In the end, I have obtained 69710 tweets that are detailed as below in Table 2. Also, each model is trained and validated in fixed training and validataion dataset, derived from the full dataset. The ratio for dividing the training, validation and test subset are 0.7, 0.1 and 0.2, respectively.

### 2.2 Vanilla Model

As Rocha et al. (2016) had showned in their work, the best stylometric features for authorship attribution on social media text are word-level and char-level n-grams. I have designed a vanilla attribution model with word-level 1, 2, 3-grams and char-level 1, 2, 3-grams, which is then feeded into a logistic regression classifier and a LIBLINEAR(Fan et al., 2008) SVM classifier implemented by Scikit-learn (Pedregosa et al., 2011). The word-level and char-level features are calculated from all available tweets without distinguishing the language. I have also transformed all of the appearance counts to TF-IDF values to deminish the effect of trending words.

We have seen from Table 2 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 LIBLINEAR SVM to balance out the dataset automatically by assigning more weight to minor classes (users).

### 2.3 Machine Translation Model

Inspired by Bogdanova and Lazaridou (2014) I have adapted to tackle the cross-langauge authorship attribution problem is by using machine translation. I have used the Translator Text API from Microsoft Azure Cognitive Services. The state-of-art machine translation service is far from perfect and underperforms on social media text than formal writings. In order words it will inevitably introduce "distortion" to the raw tweet and worsen the result in theory, I still argue that machine translation is one of the cheapest and the most intuitive solution to multi-language tasks in NLP.

In this second authorship attribution 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 with other translated tweets within the same group.

### 2.4 CNN 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-gram 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 tweets | length of raw tweets | length of CHN tweets | length of ENG 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 2: Distribution of the dataset

| Hyperparameters | Value |
|---|---|
| # of embedding layers | 1 |
| dimension | 600 |
| # of convolutional layers | 3 |
| 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 3: Hyperparameter settings in the CNN model

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 adapted Adam optimizer (Kingma and Ba, 2014) and all of the hyperparams are shown in Table 3.

## References


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.

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*.

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

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.

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.

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.

Daniel Wood, Sean McMinn, and Emily Feng.

2019. China used twitter to disrupt hong kong protests, but efforts began years earlier.