A study on authorship attribution of Chinese texts based on discourse information analysis

Shaomin Zhang

School of Foreign Languages
Guangdong Polytechnic Normal University
No. 293, Zhongshan Highway
Tianhe District
Guangzhou, 510665
China

Awarding Institution: Guangdong University of Foreign Studies, China, Guangdong, Guangzhou, Baiyun, 白云大道北2号 邮政编码: 510420

Date of Award: 17 December 2014

KEYWORDS FORENSIC LINGUISTICS, AUTHORSHIP ATTRIBUTION, DISCOURSE INFORMATION ANALYSIS, DISCRIMINANT ANALYSIS
In the era of computers and the Internet, authorship attribution in forensic linguistics ushers in a new stage in which e-texts and typescripts, instead of the traditional handwriting specimen, may become the mainstream data for attribution. Therefore, analyses based on traditional handwriting in the field of justice in China will be less and less usual with the decrease in standard handwritten documents. Under this condition, the present research aims to probe into the method of authorship attribution of Chinese texts, including e-texts and typescripts, on the basis of discourse information analysis (DIA). It is hoped that some reliable and valid features of authorship attribution can be unveiled and fully explained in terms of both texts and text authors (Grant and Baker 2001). Such a research objective is formulated into three specific research questions, which were designed at descriptive, interpretative and explanatory levels respectively, as follows:

1. What are the discourse information (DI) features that have robust discriminating power in authorship attribution of Chinese texts?
2. Which features or feature sets have best discriminating power for the texts of a specific group of authors or of a specific genre?
3. What is the relation between identified text features and the authors’ idiolects?

The present research was conducted in the framework of DIA, idiolect theory, psycholinguistic and sociolinguistic theories.

The texts used in the present research are composed of emails, QQ diaries, microblogs, column articles and some request-writings according to seven designated topics. Two types of experiments were designed: Type I Experiments were carried out for examining the first four types of existing data and Type II Experiments are for the request-writings.

Methodologically, both qualitative and quantitative analyses were adopted in order to discover some reliable and valid DI features. Qualitatively, discourse information analysis was conducted for the purpose of DI feature extraction. During this process, some consistent features among specific authors were extracted (Grant 2010; Grant 2013). Quantitatively, the extracted features were tested mainly through discriminant analysis (DA) for their combined distinctive power of distinguishing texts of different authors.

Through various sub-experiments in Type I and Type II Experiments, a series of DI features proved to be satisfactory in discriminating Chinese texts of different authors in most of the cases. Furthermore, based on the results of the experiments, the main outcomes can be summarised as follows:

1. It was found that ten DI features demonstrated robust discriminating power in authorship attribution for Chinese texts: information units, subjective information units, objective information units, objective information units, information elements, information
elements per information unit, conditions, static processes, dynamic processes, information clusters and sharpening in focus. The different combinations of the DI features could work together in a variety of contexts in order to exhibit their distinguishing power.

2. A series of DI features were found to be closely related to gender, age and text genre in authorship attribution. In particular, in terms of gender, different DI feature sets were tested and proved to be essential for distinguishing the authors of different genders, same gender, different age groups and different genres of texts. In more detail, there were six types of DI features that proved to be significant in distinguishing between male and female authors, three kinds of DI features proved to be significant in distinguishing male authors, and five different types of DI features were significant in distinguishing female authors. In terms of age, eight kinds of DI features were found to be essential in distinguishing authors of the 1960s, the 1970s and the 1980s. In terms of text genres, another eight types of DI features were found to be significant in distinguishing four different text genres: QQ diaries, emails, microblogs and column articles.

3. From perspectives of gender, age and genre, DI features could be used to identify the authors’ idiolects mainly in four dimensions: informativeness, powerfulness, ornament and elaboration.

In addition, there are some other related findings about author set size and text length, which are illustrated as follows.

The author set size exhibited an essential effect on the attribution results of short Chinese texts of less than 500 Chinese characters and five authors proved to be the threshold of the author set size. In a similar vein, text length also exerted a profound influence on the attribution results. To be specific, long Chinese texts of more than 500 characters showed robustness despite the increase in candidate authors. The maximum number of the candidate authors for long Chinese texts in the present study was seven. However, the performance of short Chinese texts was very vulnerable to the author set size in the process of authorship attribution.

Thus, based on the two findings, the interplay between author set size and text length is easily spotted. Long Chinese texts of more than 500 characters were very robust and not vulnerable to number of authors, while short Chinese texts were very vulnerable to the number of authors in authorship attribution.

The present study makes both theoretical and practical contributions to the study of authorship attribution in China in that a new type of theory and related authorship features, i.e. discourse information theory and DI features, were introduced and proved to be effective. On the one hand, the method of the present research may be applied to the investigation of authorship attribution cases in conjunction with other methods such as handwriting analysis and polygraph tests. Thus, it can provide supportive evidence in police investigations and legal
practice in China. On the other hand, the method may also be applied in authorship attribution cases in which handwritten texts and other evidence are not available.

Acknowledgments

I am very grateful to Professor Du Jinbang, Dr Kirchhuebel, Professors Yuan Chuanyou and Xu Zhanghong for their helpful suggestions and comments. This work was also supported by the National Social Science Project of China (Project No.: 13BYY066) led by Professor Du Jinbang

References

