Author Identity Classification in Criminal Investigations

1810225 Shionaga Masanao

In recent years, cybercrime has been on the rise due to the widespread use of the Internet, the increase in digital contents, and the growing number of users of social media. One of cybercrime is "impersonation" on texts in blogs, microblogs, electronic bulletin boards, e-mails, and chat rooms. In order to prevent crimes caused by impersonation, it is necessary to develop a technique that automatically estimates the author of a text based on its contents and detects impersonation.

There is a long history of research on author estimation by quantitative analysis of stylistic features of texts. There are two major tasks in author estimation: author classification and author identity classification. Author classification is a task to classify an author of a given text into several author candidates. On the other hand, author identity classification is a task to determine whether two texts are written by the same author. Although there are many previous studies on author classification, a few studies are carried out on author identity classification for Japanese documents. In particular, no research has applied supervised machine learning, which has been successful in the field of natural language processing in recent years. Therefore, author identity classification is ongoing and still an important research topic.

This study proposes a method for author identity classification of Japanese texts toward automatic detection of impersonation. Texts on blogs are used to develop our proposed system. When a text written by an unknown author (uncertain text) and a set of texts written by a known author (reference text), we aim at automatically determining whether they are written by the same author or not. We train such a model by supervised machine learning. In addition, it is supposed that most of the texts are written by the same author when author identity classification is performed for impersonation detection. Therefore, we propose a model optimized for impersonation detection so that it can accurately detect a small number of text pairs written by different authors.

This research contains two major topics: "construction of the dataset" and "construction of the author identity classification model". The dataset in the first topic is used for training and evaluation of the model.

The dataset used in the proposed method is constructed by the following procedures. First, we collect blog articles from Ameba blog with author IDs. Each retrieved article consists of 300 words or more, and the number of articles per author is 10 to 20. The number of authors is 596, and the number of articles is 10,433. Next, from the retrieved blog articles, we obtain pairs of

reference texts and uncertain texts (hereinafter referred to as "instances"). Two types of instances are made: "same author pair" where the authors of the reference and uncertain texts are the same, and "different author pair" where the authors are different. If the same author pair and different author pair are created by all combinations of authors, the number of different author pairs will be much larger than the same author pairs. Therefore, by suppressing the creation of the different author pairs, we obtain the same number of the same and different author pairs. Finally, we divide the obtained instances into training, development, and test data. More precisely, after dividing the authors in the dataset into three sets, the same and different author pairs are created using the method described above. The training data is used for training the author identity classification model, the development data is used for parameter optimization of the model, and the test data is used for evaluation of the method. The ratio of the training, development, and test data is 8:1:1. In addition, considering the situation of impersonation detection, the number of impersonation instances (different author pairs) is expected to be much smaller than non-impersonation instances (same author pairs). Therefore, in the development and test data, we randomly remove different author pairs so that the ratio of same author pairs to different author pairs becomes 10:1. On the other hand, in the training data, the numbers of two types of instances are kept equal.

The author identity classification model is constructed by the following procedures. First, as a preprocessing of text, we perform morphological analysis on the texts using the morphological analysis engine MeCab. Next, we extract features using results of morphological analysis, i.e. word segmentation and part-of-speech (POS). Four types of features are extracted: unigram of words, bi-gram of particles, tri-gram of POSs, and words before the comma. These features might be effective in author identity classification because they represent the author's writing style and habits, not contents of the text. Next, we create a "document vector" that represents a blog article as a feature vector. Values of the feature vector (feature weights) are set to the relative frequency of the feature in a text, which considers difference of length of the documents. Next, we create an "instance vector", which is a vector of an instance, in other words, pairs of uncertain and reference texts. The document vectors of uncertain texts and reference texts are made by the above procedure. A set of reference texts is represented as a single vector by averaging vectors of individual reference texts. The vectors of uncertain and reference texts are combined to form an instance vector in three ways: difference, sum, and concatenation of two vectors. Next, we learn two models for author identity classification from the training data. The first model is a classifier obtained by Random Forest, a machine learning algorithm that has

been reported to be useful for the author estimation task in previous studies. It is denoted as "Model R" hereafter. The second model is a decision system that chooses the same author pair when the reliability of the classification by Random Forest is less than a pre-defined threshold T, otherwise chooses the class classified by Random Forest as is. It is a bias model preferring the same author pair. The threshold T is optimized using the development data. Hereafter, it is denoted as "Model B". A classifier trained from the totally balanced dataset tends to wrongly classify a same author pair as a different author pair when it is applied to imbalance data where the number of the different author pair is quite a few. Model B is designed to tackle this problem.

In the experiments to evaluate the proposed method, the performance of the author identity classification was evaluated by a 5-fold cross validation using the dataset. We compared six models obtained by combination of classification models (Model R and Model B) and instance vector creation methods (difference, sum, and concatenation of vectors). In addition to the precision, recall, and F-measure of the detection of different author pairs, the specificity, FP rate, and FN rate were used as evaluation criteria. Comparing three methods of creating the instance vectors, the model based on the difference of the vectors had the highest F-measure. The performance of the model based on the sum and concatenation of the vectors were very poor. Comparing Model R and Model B, Model B was higher for most evaluation indices. The bias toward classification into the same author pairs reduced the number of false positives, resulting in the improvement of the F-measure. Finally, the best F-measure, 0.69, was obtained by Model B based on the instance vectors made by the difference of two vectors. The recall and the specificity were 0.76 and 0.95, respectively.

From the above results, we have confirmed the effectiveness of the proposed method in author identity classification toward automatic detection of impersonation.