Character Level Authorship Attribution for Turkish Text Documents

[1]Hidayet Takçı, [2]Ekin Ekinci

[1]Mühendislik Fakültesi, Bilgisayar Mühendisliği, Cumhuriyet Üniversitesi, Sivas-Turkey
[2]Mühendislik Fakültesi, Bilgisayar Mühendisliği, Kocaeli Üniversitesi, Kocaeli-Turkey

htakci@gmail.com, ekin.ekinci@kocaeli.edu.tr

Abstract: Individuals have their own style of speaking and writing. Style of a text can be used as a distinctive feature to recognize its author. In recent years, practical applications for authorship attribution have grown in areas such as criminal law, civil law and computer security. Recent research has used techniques from machine learning, information retrieval and natural language processing in authorship attribution. In this paper, Statistical Language Modeling is utilized in Authorship Attribution. Each author is represented with feature statistics. Letters, punctuations and special characters which build up the feature set are utilized to calculate the profiles of the authors.

Key words: Authorship Attribution, Character Level Method, Centroid Values, Centroid Vector, Document Vector

Introduction

The topic of the article is authorship attribution and this study aims to recognize authors of Turkish texts automatically. In addition, it can be utilized in different areas such as spam filtering, determining plagiarism cases, identifying author of program code and in forensic analysis. The output of this study will be classification of texts based on the authors, determining the authors with similar styles in writing and classification of authors depending on their styles. Similarity in the authors' styles is related to their cultural or geographical backgrounds. This situation makes us able to reach interesting information about the authors.

In authorship attribution studies researches have experienced different features such as function words, content words, character n-gram, word and punctuation marks profile etc. Their performances are changeable. While some of these methods give best results, some of them don't give because of preferred dataset. Stylistic and statistical methods can be utilized for authorship attribution. This study deals with the statistical methods in authorship attribution. Recognizing the author by statistical methods necessitates accurate expressions of numerical data.

In the proposed method; letters, punctuation marks and some special characters are added to the feature set individually. Feature set of this method is considerably small compared with the feature sets of other methods. In the character level authorship attribution, including the punctuation marks and special characters such as "space" and "enter", all characters that form the text are all members of the feature set. For this reason, no preprocessing step is required and we are able to work with raw data. And we can easily acquire this type of information. With the proposed character level method, we profit from the costly preprocessing steps. Proposed method uses centroid based classification algorithm which is a very successful algorithm in text classification as well as Bayesian text classification (Han & Karypis, 2000).

The rest of the paper is organized as follows. Next, in Section 2, we give briefly related works. Section 3 describes the character level authorship attribution. After that, Section 4 gives information about experiments and results, and finally Section 5 includes our conclusion that we have been able to achieve so far.

Related w-Works

The authorship attribution system is an application which aims to recognize the author of a text and it is in relation with different areas such as speech recognition, spam filtering and copyright. The studies on authorship attribution have been continuing since 19th century. In 1887 Mendenhall made first known study about authorship attribution and he used words as feature (Mendenhall, 1887). Zipf and Yule used statistical method for authorship attribution respectively 1932 and 1938 (Statamatos, 2008). In 1964 Mosteller and Wallace used Bayesian Analysis of 90 functional words to find authorship of "The Federalist Papers" (Mosteller &

Wallace, 1964). This study is accepted milestone of authorship attribution. After Mosteller and Wallace study, functional words were started to use in many studies (Koppel, Scheler & Argamon, 2008).

In 1990s, researchers started to use linguistic style for authorship attribution (Statamatos, 2008). Using linguistic style for text is called stylometry. Stylometry originates with the suggestion of Augustus de Morgan in 1851 that "it might be possible to identify authors because one might deal in longer words" (Morgan & Elizabeth, 1882).

In 2001 Grant and Baker described an approach known as Principal Component Analysis. This approach identifies which marker or combinations of markers are effective in discriminating the author of a text (Grant & Baker, 2001). In Baayen and his colleagues' study they proved that authors have 'textual fingerprints'. Statistical methods were used in their study and according the results discriminant analysis is a more powerful technique than principal component analysis. Using punctuation marks with function words and content words increase the classification accuracy (Baayen, Halteren, Neijt & Tweedie, 2002).

Vocabulary richness and repetition; word type frequencies and distributions; word, sentence clause and paragraph lengths and distributions; syntactic analysis, co-occurrence and collocations; and content analysis are the other valid criterions in authorship attribution. Diri and Amasyalı used these criterions to identify authors of Turkish texts and developed a new classification technique. In their study 22 of style markers figured out for each 18 authors and %84 success rate has been reached in average (Diri & Amasyalı, 2003). In 2007 Taş and Görür developed a new classification technique to identify author for Turkish texts. For identifying the authors, 35 of style markers have been figured out. Their experimental group consists of 20 authors and with the developed method they obtained a success rate of %80 in average (Taş & Görür, 2007). Grieve used thirty nine different types of textual measurements in attribution studies in 2007, with word and punctuation mark profile they reached best results, also 2-gram and 3-gram profiles give best results (10-author limit) (Grieve, 2007).

Character Level Authorship Attribution

Character level authorship attribution is an author recognition method which deals with individual characters that compose the text. Characters can also be utilized by the other author recognition methods. But in those methods characters are generally taken into consideration as combinations of characters not individually. In the proposed system, each character individually is a member of the feature set. Besides characters such as "enter" and "space" which can provide vital information about the author's style are also added to the feature set. Character level method is very effective technique for author attribution. Characters were also used in identification of languages and best results were acquired. Language identification studies with characters using centroid based model gave best results. Therefore, in this study character level features and centroid based model were used.

Feature selection which determines the feature set is a very important process. Dimension of the feature set is another important aspect for studies in author recognition. In some methods, such as n-grams and functional words, large feature set decrease the effectiveness of the authorship attribution system. For example, in order to recognize the author of a Turkish text by using the functional words method, all the frequently used words (adjectives, pronouns, adverbs, conjunctions…) are required to be added to the feature set. On the other hand, the feature set of character level model is quite smaller than other methods. Despite the small size of the feature set, features are very successful at representing the text. The other approach in authorship attribution is word level analysis has also some problems. While using word level analysis, morphological features is not important and when studying with some Asian languages which have no explicit boundaries researchers face with problems (Keselj, Peng, Cercone & Thomas, 2003), character level method avoids such problems.

By using individual characters instead of n-grams or functional words, the feature set would be limited with the letters and punctuation marks that are included in the alphabet. For the authorship attribution in Turkish text documents, it is possible to make a feature set consisting of 29 letters of Turkish alphabet and the punctuation marks that are most frequently used in the language. So, individual characters can be used in authorship attribution for real time applications where effectiveness has a vital importance.

When we examine texts from different authors, we find out that different texts of an author have similar character frequencies. Therefore, character frequencies can be utilized to find out the author of a text. Texts written by the same author as well as texts written by different authors have distinct character frequencies. But, while character frequencies of texts written by the same author are very similar to one another, the frequency of texts written by different authors has quite different character frequencies. This case constitutes the basis of the character level model. Hypothesis of our study is "Characters are discriminative markers for authors and texts can be classified due to the frequencies of characters that it includes. Owing to this, each text can be designated to the related group of its author."

Author's style can be used to identify it. This is the second basis of authorship attribution. For example; while some authors hardly ever use exclamation mark, some authors use the exclamation mark quite often, some authors use comma frequently because they like long sentences while some authors use dot more frequently by using short sentences in their writings. These kinds of details in the text have vital importance in authorship attribution.

A model is a simplified prototype of a system. When the character level authorship attribution is considered as a classification problem, the model of the system will consist of training and test phases. Character level model can be stated as follows.

d_i	i^{th} document in the corpus
fr_{ip}	the frequency of p^{th} character in document i
d_{ip}	the relative frequency or n normalized value of fr_{ip}
$\hat{y}_i$	Represents authors of document (training phase)
x_i	Represents authors of document (test phase)
Cj	centroid value for j^{th} author
A_k	represents the average character frequency for k^{th} author
$a_{(T)p}$	represents the total usage frequencies of p^{th} character
c_{jp}	represents the centroid value for p^{th} character in texts of j^{th} author
m	Number of features

Table 1: Parameters

Each document has at least one author. The relation between the documents and their authors (authors are defined with numbers) is stated as follows.

$$D \longrightarrow \{1,2,..,k\}$$

In this study, characters are used as features of the documents, and feature values are the frequencies of these characters in the documents. Values of the determined features are generally presented by vector space model. d_i is represented with a document letter vector as below.

$$\vec{d}_i=(d_{i1}, d_{i2},\ldots, d_{im})$$

Relative frequency ($d_{i\,p}$) is calculated in order to prevent errors caused by the length of the document. The relation between fr_{ip} and d_{ip} is as follows.

$$d_{ip} = \frac{fr_{ip}}{\sum_{p=1}^{m} fr_{ip}}$$

d_{ip} is the p^{th} dimension of vector d_i. Each dimension of document vector represents frequency value of a character. The documents whose authors are unknown are represented by X and the document i is represented by the statement x_i. $\hat{y}_i$ is different from y_i because $\hat{y}_i$ is an estimated value, not an accurate value. It is inevitable to make clear the relation between $\hat{y}_i$ and x_i in order to find the author of a document. The equation :

$$\hat{y}_i = \underset{j=1\ldots k}{\mathrm{argmax}}(\mathrm{Sim}(\vec{x}_{i,}\vec{C}_j))$$

can be used to show this relationship. Cj value is required for authorship attribution. Before centroid values, average character frequencies for each author have to be calculated. This process aims to find the average character frequencies of the samples. For example, after getting the character frequencies of 100 sample 1KB documents, we can obtain average frequency value for each author by calculating the average value of these frequencies. Average character frequency calculation is as below.

$$\vec{A}_j = \frac{1}{n}\sum_{i=1}^{n} \vec{d}_{ji}$$

Character frequencies for each author can be stated as $A=(a_1, a_2,\ldots,a_m)$. For the author with indice j, presentation of average frequency values by means of features is as follows.

$$A_j=(a_{j1}, a_{j2},\ldots, a_{jm})$$

Following equations can be written where $a_{(T)p}$ represents the total usage frequencies of p^{th} character for all the authors and c_{jp} represents the centroid value for p^{th} character in texts of j^{th} author.

$$a_{(T)p} = \sum_{i=1}^{k} a_{jp}$$

$$c_{jp} = \log(a_{jp} * 100 / a_{(T)p})$$

A function called Sim is used for finding similarity. This function is cosine similarity function. Function is stated in 3.9. When the denominator of this equation is removed, we can obtain text scores.

$$Sim(\vec{x}_i, \vec{C}_j) = \frac{\sum_{p=1}^{m} \vec{x}_{ip} . \vec{c}_{jp}}{\sqrt{\sum_{p=1}^{m} (\vec{x}_{ip})^2} . \sqrt{\sum_{p=1}^{m} (\vec{c}_{jp})^2}}$$

Experiments and Results

Data set, that was required for training and test phases of the character level author attribution system, was formed from the articles of a daily newspaper SABAH (www.sabah.com.tr). Articles of the authors who write about different topics such as politics, magazine and medical were preferred. Training set consists of 10 different texts written by 10 different authors and the test set consists of 10 sample texts for each author. The feature set initially consisting of 29 letters of the Turkish alphabet, has been extended to 42 features by adding punctuation marks and some special characters in progress.

Texts have to be presented by numerical data because classification algorithm is used in authorship attribution. For this reason, the frequencies of letters and punctuation marks in the texts are calculated. Characters are counted simply in order to find the frequencies of characters. As the raw data about the character frequencies can't help much, relative frequencies also have to be calculated. After reaching the relative character frequencies for documents, each document is represented by a document vector in document character space.

It is possible to consider the centroids as vectors that present the authors' characteristics. While centroid vectors represent authors, document character vectors represent the documents. Centroid values are obtained from the frequencies of characters which are used by each author. After getting the centroid vectors, similarities between test document and centroid vectors are examined in order to perform authorship attribution process. By applying test scoring method for similarity, it is possible to find the authors' scores of the test document from the dot product of document character vector and centroid vector. After the author scores for the test document are examined, the test document is classified.

Each author has its own style of writing and this is the main idea of character level authorship attribution. In this respect, each author expresses his taught and ideas with different words and different styles of sentences. An author's style makes us possible to recognize the author of a text. Character frequencies being able to let us recognize the author of a text will be the proof of character level method. High success ratio for the authorship attribution has been aimed. For this reason, different experiments have been held in order to find an optimum success ratio. These experiments aim to show that which numerical values should be used to represent the data and which similarity methods should be used.

It is possible to use average frequency and centroid values of texts for recognizing the author. But these experiments show that using centroid values more successful than using average frequency. The most important difference between centroid values and average character frequencies is the logarithm transformation. So, owing to the logarithm transformation high success ratio can be obtained with the centroid values.

Functional words method used besides character based method in respect of correct identification and performance. Accuracy rates are shown in the table below.

ID	Author	Functional words	Characters
1	Author1	%60	%90
2	Author2	%40	%90
3	Author3	%50	%60
4	Author4	%70	%100
5	Author5	%50	%90
6	Author6	%20	%60
7	Author7	%40	%80
8	Author8	%10	%100
9	Author9	%100	%90
10	Author10	%90	%100
Average Success Rate		%53	%86

Table 2: Accuracy Rates

It is examined from the results of the experiments made with the same dataset that character based method is more successful than functional words method. In this stduy 67 functional words were used. While the success ratio for the functional words method is 53%, this ratio is 86% for character based method. While character based method not required preprocessing step, functional words method requires this step. Preproceesing step which is a process applied to each article in the dataset neccessitates a lot of time. Besides, small feature set of the character based method makes it superior to the other methods with large feature set.

Conclusion

With this study, an authorship attribution system has been developed with a character level method and it has been compared to the preceding systems with functional words method. Each author in the character level authorship attribution system has been represented by centroid vectors. The author of a test document is identified after examining the similarities between the document character vector of the document and the centroid vectors of the authors. Cosine method has been used to find similarities.

Character based authorship attribution is superior to other methods in respect of performance and effectiveness. Small feature set, studying with raw data makes this method effective. In respect of correct identification and performance, character based method is the most appropriate and successful method for daily articles. For this reason, it is suitable for cases where performance is important. Character based method can be used in spam filtering or plagiarism detection because these processes are also performed by examining the characteristic features of a text.

References

Han, E., & Karypis., G. (2000). Centroid-Based Document Classification Algorithms: Analysis & Experimental Results. *In Principles of Data Mining and Knowledge Discovery*, 424-431.

Mendenhall, T. C. (1887). The Characteristic Curves of Composition. *Science*, IX, 237-246.

Statamatos, E. (2008). A Survey of Modern Authorship Attribution Methods. Journal *of American Society for Information Science and Technology*, 60(3):538-556.

Mosteller, F., & Wallace, D.C. (1964). Inference and Disputed Authorship: The Federalist. *Addison Wesley*.

Koppel, M., Scheler, J., & Argamon, S. (2008). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, 60, 9-26.

Morgan, D., & Elizabeth, S. (1882). Memoir of Augustus De Morgan. *Longmans, Green and Co: London*, 401-415.

Grant, T., & Baker, K. (2001). Identifying Reliable, Valid Markers of Authorship: A Response to Chaski. *Forensic Linguistics*, 8(1):, 66–79.

Baayen, H., Halteren, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. *6es Journ´ees internationales d'Analyse statistique des Donn´ees Textuelles*, 69-75.

Diri, B., & Amasyalı, F. (2003). Automatic Author Detection for Turkish Texts. *ICANN/ICONIP 2003, İstanbul*, 138-141.

Taş, T., & Görür, A.K. (2007). Author Identification for Turkish Texts. *Journal of Arts and Sciences*, 7, 151-161.

Grieve, J. (2007). Quantiative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22, 251-270.

Keselj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-Gram Based Author Profiles for Authorship Attribution. *Pasific Association for Computational Linguistics*, 255-264.