Automatic Author Detection for Turkish Texts

Banu Diri [1], M.Fatih Amasyali [1]

[1] Yildiz Technical University, Department of Computer Engineering
80750 Istanbul, Turkey
{banu,mfatih}@ce.yildiz.edu.tr

Abstract- To classify a text or to recognize its author there are two ways. To use the content of the text or the style. In this study 22 of style markers figured out for each author. By the developed method the author of a text can be determined using the style markers formed from a group of authors. The author group consists of 18 different authors and the success rate has been obtained as %84 in average.

Index Terms- Natural Language Processing, Author Recognition, Stylometry, Statistical Data Analysis

I. INTRODUCTION

Increasing Internet applications and their growing rich content lead the accumulation of huge amount of electronic documents on all over the net servers. Each day gigabytes of data get produced on the Internet. There exists very different type of documents. Some documents are for images, some are for sounds and some are for texts. The electronic form of the published texts gives the ability of processing them by using some special software. The motivation behind these softwares is the need for rapid retrieval of the required data, search for specific information and some language specific techniques such as Natural Language Processing (NLP). In this study a technique has been developed for figuring out the author of a given text by having the knowledge of the author group.

To classify a text there are two different properties. One is the content of text, the other one is the style [1]. There are hundreds of researches about this subject in the last 35 years. The pioneers of authorship attribution are Brinegar (1963)[2] he focused on word lengths, Morton (1965)[2] he focused on sentence lengths, and Brainerd (1974)[2] he focused on syllables per word. Holmes (1992)[2] developed a function to relate the frequency of used words and the text length. Karlgren-Cutting (1994)[3] figured out the style marker of the text. Biber (1995)[4] added the syntactic and lexical style markers. In the recent improvements on authorship attribution we can see Kessler (1997)[5] he developed a simple and confident method. In 1998 Twedie and Baayen[6] showed that the proportion of the different word count to the total word count could be a fair measurement and the results for the texts which are shorter than 1000 word in length could be inconsistent. In the year 2000 Stamatatos-Fakotakis-Kokkinakis[2] have measured a success rate of %65 and %72 in their study for authorship recognition, which is an implementation of Multiple Regression and Discriminant Analysis. They have measured these results on ten authors and also showed that this method can also be used in texts, which are shorter than 1000 words in length.

What are the properties of an author to distinguish from the others? When we read an article of an author we can recognize his words, his style and the structure of the sentences if we already read another articles from the same author. Most of the time a reader can distinguish the author of his newspaper. Is it possible to automate this process?

In this study a new method has been proposed and the presentation of the most deterministic properties of the authors has been given. In the second section of the study exists the way we establish the corpus. In the third section there are the modules of the proposed system and the relation between themselves. In the fourth section there are the results of the tests and the comparisons of these results with the implementation of the Neural Network.

II. THE FORMATION OF THE AUTHOR BASED CORPUS

For the authorship attribution a corpus has been developed in which it contains texts downloaded from URL www.hurriyet.com.tr. These texts have different subjects such as magazine, medical and politics. Author based corpus contain two sets, test and training. Training set contains 15 and test set contains 5 different texts for each of 18 authors. Selected texts comprise essays on politic, magazine and medical. The average length of texts is 456 words.

III. AUTHORSHIP ATTRIBUTION SYSTEM

First of all the system needed a Turkish dictionary [7] and the rules for the Turkish language. Then a module has been developed for extracting the properties of the text. The developed module has been implemented on the training set and attribution of the authors has been figured out. In the last section a text with an unknown author has been processed. And the system finds out the author of text. If the author is not in the training set the system gives information about the mismatch. The block diagram of authorship attribution system is given in the figure 1.

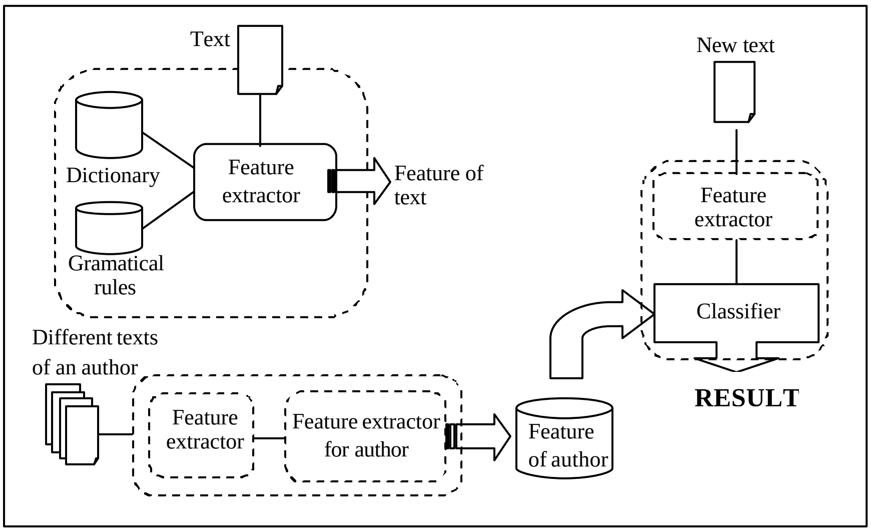

Fig.1. Block sketch of recognize author system

A. Style Markers

15 different articles for each of 18 authors has been taken to form the training set to be able to recognize the author.

Table 2 Style markers for text (SM: Style Marker)

Code	Style Markers
SM1	# of sentences
SM2	# of words
SM3	Avg. # of words in a sentence
SM4	Avg. word length
SM5	# of different words
SM6	Word richness
SM7	Avg. # of nouns in a sentence
SM8	Avg. # of verbs in a sentence
SM9	Avg. # of adj. in a sentence
SM10	Avg. # of adverb in a sentence
SM11	Avg. # of particle in a sentence
SM12	Avg. # of pronoun in a sentence
SM13	Avg. # of conjunctions in a sentence
SM14	Avg. # of exclamations in a sentence
SM15	# of point
SM16	# of comas
SM17	# of colons
SM18	# of semicolons marks
SM19	# of question marks
SM20	# of exclamation marks
SM21	# of inverted / # of all sentences
SM22	# of incomplete / # of all sentences

In the table 2 there are 22 of style markers. These 22 style marker has been processed for every text of the authors and by having the average of these 15 articles we could collect 22 style markers per author.

Style markers are determining features about number of word and sentences between SM1 and SM6, word type between SM7 and SM14, number of punctuation marks between SM15-SM20 and type of sentences in SM21 and SM22.

B. Components of Feature Extractor

As given in the figure 1 the feature extractor module consist of two parts. The first one is word database module and the other one is grammatical rule module.

1) The Word Database Module In this study the developed word database has been based on the dictionary of Turkish Language Society which consists of 35,000 words. As in the Turkish language a word could be an adjective, a noun, an adverb or a different grammatical type we needed to include the grammatical type of the word too. Maximum of 3 grammatical types which are the most used ones has been included in the system.

The word database module is in the matrix form and in first column there is the word itself, in the second, third and fourth columns there are the grammatical types of the word.

Determining word's type:

As shown in the figure 2 each word has at least one type and at most three types.

The word *i* is given as (n_i) and types are (t_1, t_2, t_3)

The types of the word as ($n_i t_j$? T j ? {1,2,3})

The type of the word as n_{it} ? T-{0}

The number of types of the word as (n_{ic} ? {1,2,3}).

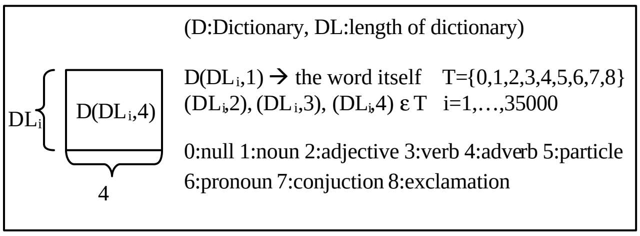
(D:Dictionary, DL:length of dictionary)
DLi
D(DLi,4)
4
D(DLi,1) → the word itself T={0,1,2,3,4,5,6,7,8}
(DLi,2), (DLi,3), (DLi,4) ε T i=1,…,35000
0:null 1:noun 2:adjective 3:verb 4:adverb 5:particle
6:pronoun 7:conjuction 8:exclamation

Fig. 2. The structure of word database module

For determining the number of types

$$n_{ic} = 1 \quad \begin{cases} n_i t_1 \neq 0 & n_i t_2 = n_i t_3 = 0 \\ n_i t_1 = n_i t_3 = 0 & n_i t_2 \neq 0 \\ n_i t_1 = n_i t_2 = 0 & n_i t_3 \neq 0 \end{cases}$$

The word has a type and this type belongs to

$$n_{it} = Z \quad \begin{cases} Z \in n_i t_k \quad k = 1,2,3 \\ Z \neq 0 \\ n_{ic} = 1 \end{cases}$$

if a word (n_i) has more than one type (n_{ic} ? 1), the type is determining according to grammatical rules.

2) Grammatical Rules As expressed in the word database module each word may have maximum of 3 grammatical types. The system automatically detects the type by implementing the grammatical rules on the sentence for this process the below rules get implemented in the below order.

1. If a word’s possible types include adjective and a word has no affix and the next word is noun or pronoun this word is adjective.
2. If a word’s possible types include adjective and a word has affix, adjective is removed from word’s possible types and number of type is decreased. If number of type is fall down to one this type is word’s type.
3. If a word’s possible types include adjective and a word has affix, go to rule 7.
4. If a word’s possible types don’t include adjective but adverb and the next word is verb or the word is at the end of sentences this word is adverb.
5. If a word’s possible types don’t include adjective but adverb and the next word is noun, adverb is removed from word’s possible types and the number of type is decreased. If number of type is fall down to one this type is word’s type.
6. If a word’s possible types don’t include adjective but adverb and the word is at the end of sentence, this word is verb.
7. The word’s type is the most used type in the text.

C. Determining Authorship Features

To detect the authorship features a training set has been formed from the 15 different articles of 18 authors. By using the feature extraction module 22 of style marker has been figured out from all of these articles. And in the last step by taking the average of each author we have collected a feature vector for each of 18 authors.

D. Suggested Authorship Recognition Method

To detect the author of given a text we first figure out the mentioned style markers (we take the X[22] vector, has these style markers). And by the help of the authorship attribution module it is possible to classify and detect the author. The matrix M is calculated from the difference of 18 author features vector and X^T vector which is the test data.

$M_{ij} = A_{ij} - X_j^T$ i=1..18 , j=1..22

For each feature of the matrix M there is a score for each author. Each row of the matrix M belongs to an author and there is score between 1-10 for the 22 feature. Concerning the method for each column of the matrix M we calculate minimum, maximum and standard deviation per feature.

min(M_j) : The minimum value of j th column of M
sd(M_j) : The standard deviation value of j th column of M
max(M_j) : The maximum value of j th column of M

Each column of the matrix M gets divided by intervals by adding the minimum value of M_j to the value of sd(M_j). Each interval represents a score. Scoring starts at the value 10 and decrease one by one by adding the standard deviation value. The author, which has the maximum score, is specified as the owner of the test data. This algorithm can be seen in below code.

```
for (i=1;i<=22;i++)
{  for (j=1;j<=18;j++)
    {k=0;
      while (k<=10)
      {   if(min(Mi)+(k*sd(Mi))<Mi)&&(Mi   <
min(Mi)+((k+1)*sd(Mi))
            {score=10-k;
k=11;a_score[j]=a_score[j]+score;}
        k++;
      }
    }
}
author=Max(a_score);
```

IV. EXPERIMENTAL RESULTS

To maximize the success of the authorship attribution system we have tested 3 different methods. The most successful one was the last one. The success rate of the developed system depends on a test set which consists of 90 texts (5 texts per author).

In the first method we have used all of 22 features which are calculated as equal weights. When we tested the proposed method in the test set we have measured a success rate of %67. To be able to compare these results we have considered the artificial neural network implementation. After 5,000 trainings by using Multilayer Perceptron (MLP) (22-75-18) we have got a success rate of %60, by using Radial Base Function (RBF) (neuron number=60, spread=75) [8] we have seen a recognition rate of %72. These 3 different results can be shown graphically in the figure 3.

In the second phase (method-2) our aim was to improve the success rate more than %70. As all the style markers have not the same influence on the results we have selected only 11 of style markers which are the most effective ones. These style markers (SM3,SM4,SM7,SM8,SM9,SM12,SM13,SM15,SM17, SM21,SM22) have the most deterministic value for authorship attribution. After having implemented these 11 style markers to the same test set we have measured a success rate of %78. But the MLP results were the same, %60. Also the results for Radial Base Function were even worse, %61. These 3 different results can be shown graphically in the figure 4.

Table 3 Recognized text of ratio (method-3)

Author	Success Ratio	Author	Success Ratio
AU01	4/5	AU10	5/5
AU02	5/5	AU11	4/5
AU03	5/5	AU12	4/5
AU04	4/5	AU13	4/5
AU05	4/5	AU14	5/5
AU06	4/5	AU15	4/5
AU07	3/5	AU16	4/5
AU08	2/5	AU17	5/5
AU09	5/5	AU18	5/5

In the third phase (method-3) we have focused to the effectiveness of these 11 different style markers. Not all of them have the same influence on authorship attribution. For example the style markers SM3, SM17, SM21 and SM13 have more deterministic effects. So we have tired to give them different weights by multiplying the style markers SM3, SM17 and SM21 by 4 and the style marker SM13 by 3. After this modification the success rate has improved approximately to %84. In this last phase we can see how many of author recognized in the table 3.

V. CONCLUSION

In this study a new classification technique which is developed by the help of the known methods has been used and it is compared with the known techniques. At the beginning 22 of style markers has been figure out and by considering them as having equal weights a success rate of %67 has been measured. Results with the artificial neural networks have %60 of success rate using MLP and %72 of success rate using Radial Base Function. In the second phase 11 of style markers among the 22 style marker has been selected as equal weights and the success rate improved to %78. But the MLP success was %60 and Radial Base Function success was %61. In the third phase the style markers SM3, SM13, SM17 and SM21 has been taken with different weights and we have measured a success rate of %84. This study shows that it is possible to identify the author of a text independent of the content and the word count from 18 different authors.

REFERENCES

[1] Y.Yang, " An Evaluation of Statistical Approaches to Text Categorization,". *Kluwer Academic Publishers. Information Retrieval 1*, 69-90, (1999)

[2] E.Stamatatos, N.Fakotakis,G.Kokkinakis, " Automatic Text Categorization in Terms of Genre and Author", *Computational Linguistics*, pp.471-495 (2000)

[3] J.Karlgren, D.Cutting, "Recognising Text Genres with Simple Metrics using Discriminant Analysis", *Proc. of the 15th International Conference on Computational Linguistics* (COLING'94), 1071-1075 (1994)

[4] D.Biber, "Dimensions of Register Variation: A Cross-Linguistic Comparison",Cambridge University Press (1995)

[5] B.Kessler,G. Nunberg, H.Schutze, "Automatic Detection of Text Genre. Proc. of 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL'97), 32-38 (1997)

[6] F.Tweedie, H.Baayen, "How Variable may a Constant be Measures of Lexical Richness in Perspective", Computers and the Humanities, 32(5):323-352 (1998)

[7] Türk Dil Kurumu : Türkçe Sözlük, Milliyet Tesisleri (1992)

[8] Haykin, S.: Neural Network, Prentice Hall (1999)