IDENTIFYING THE POETS OF THE ANONYMOUS POEMS

M.Fatih Amasyalı, Yıldız Technical University, Department of Computer Engineering, Istanbul, Turkey, mfatih@ce.yildiz.edu.tr
Banu Diri, Yıldız Technical University, Department of Computer Engineering, Istanbul, Turkey, banu@ce.yildiz.edu.tr

Abstract

To classify a poem or to recognize its poet there are two ways. To use the content of the poem or the style. In this study 22 of style markers figured out for each poet. By the developed method the poet of a poem can be determined using the style markers formed from a group of poets. The poet group consists of 7 different poets and the success rate has been obtained as %69 in average.

Keywords

Natural Language Processing, Author Recognition, Stylometry, Statistical Learning Algorithms, Statistical Data Analysis

1. INTRODUCTION

Poets and poems are generally very important for the history of the countries. Usually the poets of the famous poems are well known. But also there exist some poems called anonymous, which their poets are unknown. Researchers are trying to find out the owner of these poems in the history. But may it be a possibility to detect these poets by analysing their poems itself?

Poems are different than written and spoken language. Lines of poetry have a harmony and esthetic, usually there are rhymes. Meaning of the words may be different. There are several types of the poems such as lyric, epic, dramatic or pastoral.

In this study, we have tried to identify the poets of the anonymous poems by using text processing techniques. To classify a text there are two different properties. One is the content of text, the other one is the style [6]. 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)[2] figured out the style marker of the text. Biber (1995)[1] added the syntactic and lexical style markers. In the recent improvements on authorship attribution we can see Kessler (1997)[4] he developed a simple and confident method. In 1998 Twedie and Baayen[2] 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 a poet to distinguish from the others? When we read a poem of a poet we can recognize his words, his style and the structure of the sentences if we already read another poet from the same poet. Most of the time a reader can distinguish the poet of his poems. 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 poets 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.

2. THE FORMATION OF THE POET BASED CORPUS

For the poetship attribution a corpus has been developed in which it contains poems downloaded from URL www.siirdefteri.com.tr. These poems have different theme such as passion, life. Poet based corpus contain two sets, test and training. Training set contains 15 different poems for each of 7 different poets. In the table 1 there is the code of each poet and the average word count used in the poem. For the test set there are 5 different poems for each of 7 different poets.

Table 1 The poet-based corpus

Code	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**
# of poem	15	15	15	15	15	15	15
Words(average)	45	71	140	72	98	222	99

3. POETSHIP ATTRIBUTION SYSTEM

First of all the system needed a Turkish dictionary [5] and the rules for the Turkish language. Then a module has been developed for extracting the properties of the poem. The developed module has been implemented on the training set and attribution of the poets has been figured out. In the last section a poem with an unknown poet has been processed. And the system finds out the poet of poem. If the poet is not in the training set the system gives information about the mismatch. The block diagram of poetship attribution system is given in the figure 1.

3.1 Style Markers

15 different poems for each of 7 poets has been taken to form the training set to be able to recognize the poet. In the table 2 there are 22 of style markers. These 22 style marker has been processed for every poem of the poets and by having the average of these 15 poems we could collect 22 style markers per poets.

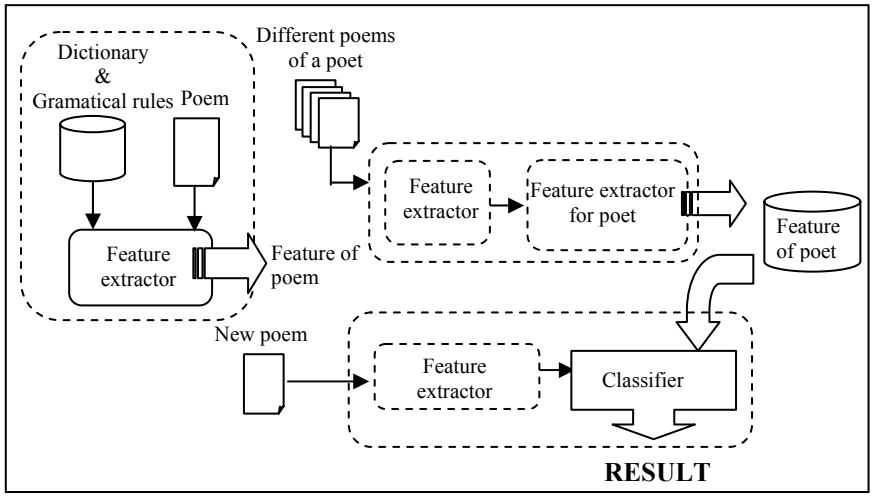

Fig.1. Block sketch of recognize poet system

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

Table 2 Style markers for poem

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

3.2 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.

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 column there are the grammatical types of the word.

(D:Dictionary, DL:length of dictionary)

$D(DL_i,1)$ → the word itself $T = \{0,1,2,3,4,5,6,7,8\}$
$(DL_i,2), (DL_i,3), (DL_i,4) \in T$ $i=1,\ldots,35000$

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

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 ($n_i t_j \in T$ $j \in \{1,2,3\}$)
The type of the word as $n_{it} \in T-\{0\}$
The number of types of the word as ($n_{ic} \in \{1,2,3\}$).

For determining the number of types	The word has a type and this type belongs to
$n_{ic}=1 \quad \begin{cases} n_i t_1 \neq 0 & n_i t_2 = n_i t_3 = 0 \\ n_i t_1 = n_i t_3 = 0 & n_i t_2 \neq 0 \\ n_i t_1 = n_i t_2 = 0 & n_i t_3 \neq 0 \end{cases}$	$n_{it}=Z \quad \begin{cases} Z \in n_i t_k \quad k=1,2,3 \\ Z \neq 0 \\ n_{ic}=1 \end{cases}$

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

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 lines of poetry. 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 lines of poetry 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 lines of poetry, this word is verb.
7. Run the below code.

```
f← 0;
for(i=1;i<=DL;i++)
{
      for (j=1;3>=j;j++)
        {  if (n_i t_j) <> 0 f[n_i t_j]++ ;  }
}
n_it=max(f)  // the type which comes from the index of the maximum element of the array f.
```

3.3 Determining Poetship Features

To detect the poetship features a training set has been formed from the 15 different poems of 7 poets. By using the feature extraction module, 22 of style marker has been figured out from all of these poems. And in the last step by taking the average of each poet we have collected a feature vector for each of 7 poets.

3.4 Suggested Poetship Recognition Method

To detect the poet of given a poem we first figure out the mentioned style markers (we take the X[22] vector, has these style markers). And by the help of the poetship attribution module it is possible to classify and detect the poet. The matrix M is calculated from the difference of 7 poet features vector and X^T vector, which is the test data. $M_{ij} = A_{ij} - X_j^T \quad i=1..7 , \ j=1..22$

For each feature of the matrix M there is a score for each poet. Each row of the matrix M belongs to a poet 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 for 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 poet, which has the maximum score, is specified as the owner of the test data.

```
for (i=1;i<=22;i++)
{  for (j=1;j<=7;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++;
       }
   }
}
poet=Max(a_score);
```

4. EXPERIMENTAL RESULTS

To maximize the success of the poetship attribution system we have tested 2 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 35 texts (5 texts per poet).
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 %57. To be able to compare these results we have considered the artificial neural network implementation. After 500 training by using Back Propagation Function (BP) (22-7) we have got a success rate of %66, by using Radial Base Function (RBF) (neuron number=60, spread=0.73) we have seen a recognition rate of %60. These 3 different results can be shown in the table 3.

Table 3 Recognized poem of ratio

	PO1	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**
BP	3/5	3/5	0	3/5	4/5	5/5	5/5
RBF	3/5	3/5	2/5	3/5	3/5	2/5	5/5
Our Method #1	2/5	1/5	1/5	5/5	4/5	4/5	3/5

In the second phase, we have focused to the effectiveness of these 22 different style markers. Not all of them have the same influence on poetship attribution. For example the style marker SM2 has more deterministic effect. So we have tired to give that different weight by multiplying the style marker SM2 by 2. After this modification the success rate has improved approximately to %69. In this last phase we can see how many of poet recognized in the table 4.

Table 4 Recognized poem of ratio

	PO1	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**
OurMethod #2	2/5	2/5	2/5	5/5	4/5	4/5	5/5

5. 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 %57 has been measured. Results with the artificial neural networks have %66 of success rate using Back Propagation Function and %60 of success rate using Radial Base Function. In the second phase, the style markers SM2 has been taken with different weight and we have measured a success rate of %69. This study shows that it is possible to identify the poet of a poem independent of the content and the word count from 7 different poets.

References

[1] D.Biber: Dimensions of Register Variation: A Cross-Linguistic Comparison, Cambridge University Press (1995)
[2] E.Stamatatos, N.Fakotakis and G.Kokkinakis : Automatic Text Categorization in Terms of Genre and Author, Computational Linguistics, pp.471-495 (2000)
[3] D.I.Holmes : A Stylometric Analysis of Mormon Scripture and Related Texts. Journal of the Royal Statistical Society, Series A, 155(1):91-120 (1992)
[4] B.Kessler, G.Nunberg and H.Schutze : Automatic Detection of Text Genre. Proc. of 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL'97), 32-38 (1997)
[5] Türk Dil Kurumu : Türkçe Sözlük, Milliyet Tesisleri (1992)
[6] Y.Yang: An Evaluation of Statistical Approaches to Text Categorization. Kluwer Academic Publishers. Information Retrieval 1, 69-90, (1999)