Letter Based Text Scoring Method for Language Identification

Hidayet Takcı and İbrahim Soğukpınar

Gebze Institute of Technology, 41400 Gebze /Kocaeli
{htakci, ispinar}@bilmuh.gyte.edu.tr

Abstract. In recent years, an unexpected amount of growth has been observed in the volume of text documents on the internet, intranet, digital libraries and news groups. It is an important issue to obtain useful information and meaningful patterns from these documents. Identification of Languages of these text documents is an important problem which is studied by many researchers. In these researches generally words (terms) have been used for language identification. Researchers have studied on different approaches like linguistic and statistical based. In this work, Letter Based Text Scoring Method has been proposed for language identification. This method is based on letter distributions of texts. Text scoring has been performed to identify the language of each text document. Text scores are calculated by using letter distributions of new text document. Besides its acceptable accuracy proposed method is easier and faster than short terms and n-gram methods.

1 Introduction

Language identification is the first step of understanding text documents which is written in. It is one of the text mining applications [8]. It can be seen as a specific instance of the more general problem of an item classification through its attributes. Text documents are classified by language identification method based on language categories. Therefore we can solve language identification problem by using of text classification algorithms. There are many algorithms for text classification [1]. Recently, due to its linear time complexity centroid-based (CB) text classification algorithm has been used for text classification [6].

In language identification, generally, short words or common words [9], n-grams [3, 11], unique letter combinations [10] etc. have been used. In addition to these methods, letter distributions of texts can be used for language identification of text documents. It has already been mentioned that, letters could be used for characterization of text documents [4]. Letter distributions can be used as an additional solution and this solution can be used to reduce the size of feature set [12].

In this study, letter based text scoring method has been proposed for language identification of text documents. In our method, text scores have been calculated by using letter distributions of texts. Text scoring has increased the speed of language identification. In the experiments successful results has been obtained by using this method.

T. Yakhno (Ed.): ADVIS 2004, LNCS 3261, pp. 283–290, 2004.

In the proposed method, firstly letter distributions of new text document (observed values) are calculated. Then these letter distributions are multiplied by average letter distributions of languages (estimated values). Average letter distributions have been obtained from training documents. Letter distributions for all languages are used from the ref. [7]. There is an average letter distribution for each of languages. These values are named as centroid values for CB algorithm. k numbers (k is equal to the number of different languages) of text scores are calculated for each new text document. The unit score of each letter is different from the others. Where, unit score is frequency of each letter in centroid values. If one of text scores is maximum then text is mapped into related class, else if there are two or more maximum text scores then text cannot be mapped into anyone class. Maximum text score is used to identify the language of new text document.

The rest of the paper is organized as follows: in the second section of this paper, using letters in language identification has been described. In the third chapter, text scoring system has been explained. In the fourth section, experimental results and analysis have given. The last section contains conclusions.

2 Using Letters in Language Identification

Short terms, n-grams, or unique letter combinations are used generally in the language identification. However, the sizes of feature set of these methods are large and preprocessing costs are very much. Therefore, reducing of their dimension and preprocessing costs is necessary.

Dimension and preprocessing problems can be indeed solved by using letters in the language identification. For example, the feature set sizes of n-grams are estimated as 2550-3560 and common words are 980-2750 [2]. The size of our method is only 30 – 40. Our method uses alphabet and the other methods use dictionary. Document-term frequencies are used by other methods whereas document-letter frequencies are used by proposed method in this work. In letter based language identification, documents are represented by letter distributions.

Letter distributions can be used to distinguish of documents. Figure 1 shows the differences of letter distributions among different languages. The reason of closeness of distributions of some letters is from the neighbor of languages. This situation is from the historical basis of languages. Figure 1 obviously depicts that letters provide distinguishing information for text documents. Letter distributions in documents depend not only on language but also on subject and writer as well.

3 Letter Based Text Scoring Method

Proposed method for language identification is a statistical based method. In our method, text scores are calculated for text classification. Text scores are obtained from letter distributions (D_{LF}) of new text document and average letter distributions (L_i) of languages. After calculating language scores of new text document they are

compared with each other. Then max language score gives the class of new text document.

Architecture of the proposed system explained following paragraphs is shown in Figure 2.

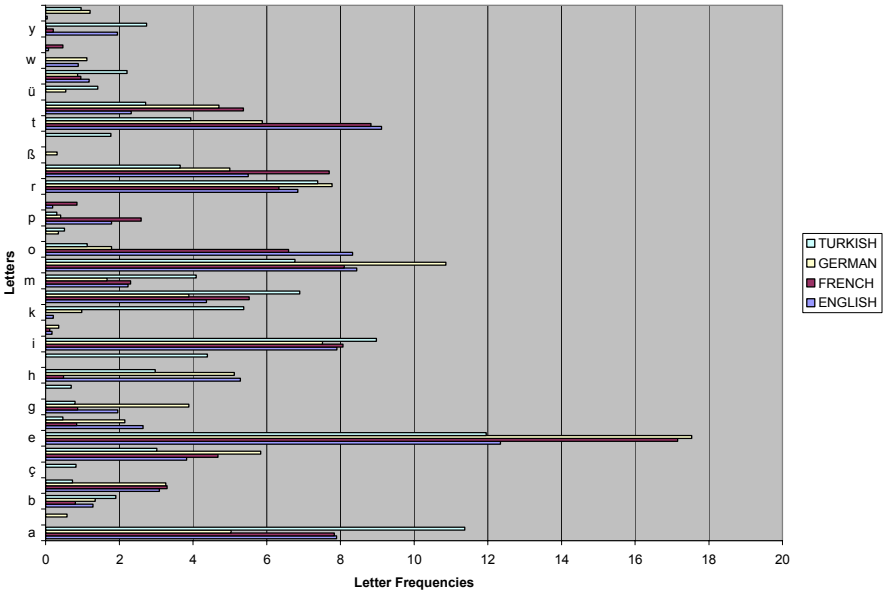

Fig. 1. Letter frequencies of each language

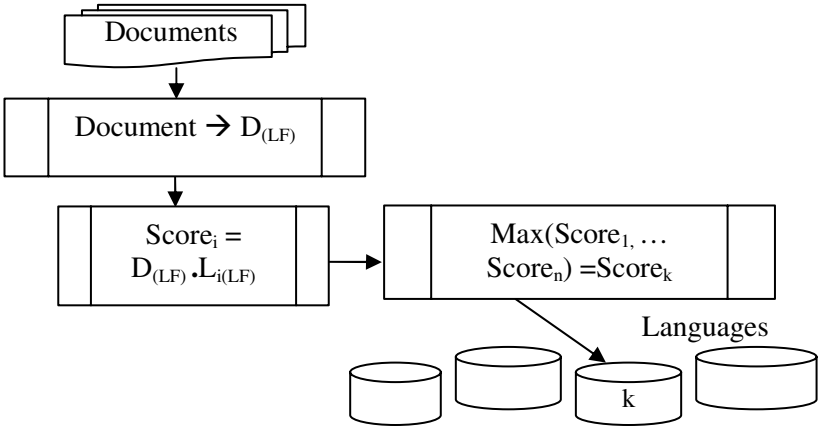

Fig. 2. Proposed language identification system

Firstly, raw text documents are transformed to letter distributions ($D_{(LF)}$). These distributions are called as document profiles. Document profiles are multiplied with centroid values of languages ($L_{i(LF)}$). If k^{th} text score is max score, new text document is assigned into k^{th} class. Using letter based text scoring is the most important difference of our solution from other language identification methods.

Vector space model has been used for representing text documents [5]. In this model, each document ***d*** is represented as a vector in a letter space. In the simplest form each document can be represented by a letter frequency (or distribution) vector shown in equation (1),

$$\vec{d}_{lf} = (lf_1, lf_2, ..., lf_n) \tag{1}$$

Where, lf_i is the frequency of i^{th} letter in the document. As weighting model for the LF vector, the frequency-weighting model has been chosen. Therefore, for each document LF vector will also be its weighting vector.

Finally, normalization is achieved by transforming each document vector into unit vector. Here, relative frequency is used for normalization.

Centroid values (average letter distributions) are language profiles. If there are k classes in training set, then, k centroid values are calculated. $\{\vec{C}_1, \vec{C}_2, ..., \vec{C}_k\}$, Each $\vec{C}_i$, is centroid value of the i^{th} class. Centroid value mentioned the mean of elements in the class. Mean value for a class is assumed to characterize the whole class. If $\vec{C}$ is to be defined as a centroid value for a document set (category or class) formed by S documents, the value of this vector is obtained as follows:

$$\vec{C} = \frac{1}{|S|} \sum_{d \in S} \vec{d} \tag{2}$$

This centroid value has been used to calculate text score value. This score value computed with equation (3) is used for identification of language of texts.

$$\text{Score}_i = D_{(LF)} \,.\, L_{i(LF)} \text{ (dot product multiplication)} \tag{3}$$

3.1 Using Letters and Text Scoring

Text scoring is a score computing task. Language scores of each text are computed as letter based. Average frequencies of letters are unit letter scores for languages. Texts are classified into language classes based on language scores. The text scoring algorithm which is shown in Fig. 2 is represented in the following pseudo-codes. Firstly some terms used in algorithm are presented

D	*: Document*
ch	*: Letter*
$\boldsymbol{lf_i}$	*: letter frequency of ith letter*
$\boldsymbol{letter_i}$	*: ith letter*
n	*: the size of letter features*
k	*: the size of language categories*
$\boldsymbol{weight_{ij}}$	*: the weight of ith letter according to jth language*
$\boldsymbol{Score_j}$	*: language scores of document*
Max	*: Maximum language score*

```
Open a text document // D
…
Read ch;
    Case ch of
   Letter_i : lf_i:= lf_i+1;
   …
          end;
for j:=1 to k do
 for i:=1 to n do
   score_j:=score_j+ lf_i * weight_ij;
Max:=0;
For j:=1 to k do Begin
   If dj>Max then Begin Max:=score_j; Class_No=j; End;
    End;
If Max>0 Then Write('Document is assigned to class
number ',Class_No)
```

4 Experimental Results and Analysis

In order to obtain successful results from language identification, training set must contain as much text documents as possible. Today it is obvious that the biggest data and documents storage is the Internet. Therefore, it is the most practical solution to get training documents from the Internet. To accomplish language identification 36285 letters have been selected from documents, distributions of letters are in Table 1.

Table 1. The sample sizes of letters according to languages

Language	Sample Letter Count
English	8502
French	9415
German	9831
Turkish	8537

In the test, English, French, German, and Turkish were used as document languages. Firstly, language scores are computed for each of texts. To compute the language scores of texts we have a score table. In this table there is a unit score of each letter according to languages. These unit scores are shown in Table 2. In fact these values are centroid values.

Table 2. The unit scores of letters

Centroid	**A**	**Ä**	**B**	**C**	**Ç**	**D**	**E**	**F**	**G**	**Ğ**	**H**	**...**	**W**	**X**	**Y**	**Z**
English	8	0	1	3	0	4	12	3	2	0	5	...	1	0	2	0
French	8	0	1	3	0	5	17	1	1	0	0	...	0	0	0	0
German	5	1	1	3	0	6	18	2	4	0	5	...	1	0	0	1
Turkish	11	0	2	1	1	3	12	0	1	1	3	...	0	0	3	1

Language scores of text are computed from the values of score table and letter distributions. Letter distributions for a text document are shown in Table 3.

Table 3. Letter frequencies of text documents

Letter Frequencies	**A**	**Ä**	**B**	**C**	**Ç**	**D**	**E**	**F**	**G**	**Ğ**	**H**	**...**	**W**	**X**	**Y**	**Z**
$\mathbf{Dlf_1}$	8	0	1	3	0	4	12	3	2	0	5	...	1	0	2	0
$\mathbf{Dlf_{2 \ldots} Dlf_{n-1}}$	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
$\mathbf{Dlf_n}$	11	0	2	1	1	3	12	0	1	1	3	...	0	0	3	1

Maximum language score are used for language identification. If all of the language scores are equal to zero or some of language scores are equal to each other then classification cannot be performed.

After representing documents in the form of LF as in Table 3, centroid values are found for each class. Centroid values are vectors formed by mean values belonging to classes. Centroid values for classes are shown in Table 2. In this system centroid values are unit scores.

This study is the modified method of "Centroid-Based Language Identification Using Letter Feature Set" [12]. At the paper letter distributions have been used for language identification. [12] In this study, we joined text scoring method into letter based language identification system. Classification accuracy of language identification has been increased by this new method. Some of the experimental results have been presented in Table 4. As it will be seen from the table 4 detection rates for letter based language identification are acceptable.

In addition to, we have also tested two methods for Turkish documents. In comparison of the results of proposed new method with results of the Letter based method in ref [12] the more accurate result has been obtained. Its classification accuracy is presented in Table 5.

Main advantage of our method is in detection speeds. We may take the size of the features of these methods for speeds of the methods. Thus, it can be claimed that when centroid based document classification is supported with letter feature set, its existing better performance is further increased and operation time is further lowered. The operation costs of three methods are presented in Table 6.

Table 4. The comparison of letter based language identification and the other techniques

	Number of words in sentence						
	1-2	3-5	6-10	11-15	16-20	21-30	31 >
English							
3gram	78.9	97.2	99.5	99.9	99.9	100.0	99.9
Short	52.6	87.7	97.3	99.8	99.9	100.0	99.9
Letter Based	54.0	78.2	96.0	99.0	99.0	99.0	100.0
Text Scoring	60.0	80	93.0	96.0	98.0	99.0	100.0
French							
3gram	69.2	93.0	94.5	93.6	99.8	100.0	99.9
Short	30.8	81.8	96.0	97.2	99.8	100.0	100.0
Letter Based	66.0	92.0	96.0	95.0	97.0	100.0	100.0
Text Scoring	85.0	98.0	100.0	100.0	100.0	100.0	100.0
German							
3gram	90.3	97.2	99.3	99.8	99.9	100.0	100.0
Short	30.8	81.8	96.0	97.2	99.8	100.0	100.0
Letter Based	65.0	90.0	92.0	95.0	96.0	100.0	100.0
Text Scoring	75.0	83.0	99.0	98.0	100.0	100.0	100.0

Table 5. The comparison of letter based language identification and text scoring

	Number of words in sentence						
	1-2	3-5	6-10	11-15	16-20	21-30	31 >
Turkish							
Letter Based	76.0	93.0	96.0	100.0	100.0	100.0	100.0
Text Scoring	93.0	97.0	99.0	100.0	100.0	100.0	100.0

Table 6. Operation costs of Language identification systems

Operation cost	
3gram	2550 ops.
Short	980 ops.
Text Scoring	37 ops.

It is assumed that operation costs depend on the size of feature sets of language identification systems.

5 Conclusions

Data mining is a partially new technique of finding meaningful information and useful patterns from large amount of data. It has been applied generally to structural data stored at database. Therefore, data mining is considered as one of knowledge discovery steps from databases. After realizing that concepts of data mining can be applied also to data, which are not structural, text mining is born as a new field. Text

categorization is being at an important position at text mining. Therefore, text documents can be automatically assigned to previously defined classes by this technique. In text categorization operation, documents are represented according to frequency information of words that are concerned in these documents and in that way documents enter into the classification process.

In this study, we have proposed that documents could be represented by letter frequencies instead of word frequencies. Text scoring method has been used for language identification. Experiments have been conducted for four languages. Pleased results are achieved in the end of this experiment. Thus, it has been revealed that letter feature set can be used for language based recognition types. As a result, it is shown that the usage of letter feature sets can be used for language identification of text documents. For future study, experiments can be extended for more different languages.

References

1. Dumas, S., Plat, J., Heckerman, D., and Sahami, M.: Inductive learning algorithms and representation for text categorization. In Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management, pages 148-155, 1998.
2. Grefenstette, G.,: Comparing two language identification schemes, in Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT?95), Rome, Italy, December 1995.
3. Cavnar, W *and* Trenkle, J. *:* "N-gram-based text categorization," in Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp.161--175, 1994.
4. Benedetto, D., Caglioti, E. and Loreto, V.: Language trees and zipping. Physical Review Letters, 88:4(2002).
5. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989
6. Han, E.-H. and Karypis, G.: Centroid-based document classification: Analysis and experimental results. In Principles of Data Mining and Knowledge Discovery, pages 424--431, 2000.
7. Pawlowski, B.,: Letter Frequency Statistics, URL : //www.ultrasw.com/pawlowski/brendan/Frequencies.html
8. Visa,A.: Technology of Text Mining, (MLDM 2001), Perner, P. (Ed.), LNAI 2123, pp. 1–11, 2001.
9. Johnson, S.,: Solving the problem of language recognition Technical report, School of Computer Studies, University of Leeds, 1993
10. Churcher, G.,: Distinctive character sequences, Personal communication, 1994
11. Hayes, J.,: Language Recognition using two and three letter clusters. Technical report, School of Computer Studies, University of Leeds, 1993
12. Takçı, H., Soğukpınar, İ.,: Centroid-Based Language Identification Using Letter Feature Set, Lecture Notes in Computer Science, (CICLING 2004) Springer-Verlag, Vol. 2945/2004, pages 635-645, February 2004