International Journal of Engineering

Journal Homepage: www.ije.ir

A Signal Processing Method for Text Language Identification

H. Hassanpour[a], M. M. AlyanNezhadi*[b], M. Mohammadi[c]

[a] *Image Processing & Data Mining Lab, Shahrood University of Technology, Shahrood, Iran*
[b] *Department of Mathematics, University of Science and Technology of Mazandaran, Behshahr, Iran*
[c] *Department of Information Technology, College of Engineering and Computer Science, Lebanese French University, KR-Iraq*

PAPER INFO

Paper history:
Received 17 March 2021
Received in revised form 03 April 2021
Accepted 04 April 2021

Keywords:
Language Identification
Signal Processing
Wavelet Packet Transform
Artificial Neural Network

ABSTRACT

Language identification is a critical step prior to any natural language processing. In this paper, a signal processing method for Language Identification is proposed. Sequence of characters in a word and the order of words in stream identify the language. The sequence of characters in a stream provides a signature to recognize the language without understanding its meaning. The signature can be extracted using signal processing techniques via converting texts into time series. Although several research and commercial software have been developed to identify text language, they need a standard dictionary for each language. We proposed a dictionary independent method consisting of three main steps, I) pre-processing, II) clustering and finally III) classification .First, the texts are converted to time series using UTF-8 codes. Second, to group similar languages, the obtained series are clustered. Third, each cluster is decomposed into 32 sub-bands using a Wavelet packet, and 32 features are extracted from each sub-band. Also, a multilayer perceptron neural network is used to classify the extracted features. The proposed method was tested on our dataset with 31000 texts from 31 different languages. The proposed method achieved 72.20% accuracy for language identification.

***doi**: 10.5829/ije.2021.34.06c.04*

1. INTRODUCTION

Natural language processing (NLP) techniques play an important role in the classification and processing of huge digital documents on the Web [1, 2]. Determination of the language of a text's content is called Language Identification (LID classification). This is the initial step in many NLP pipelines such as tagging data stream from Twitter with relevant language, improving search results by searching in the relevant language, and automatically using machine translation [3]. Since most of the later steps are language-dependent, any errors in the first step is compounded by later steps. Although the determination of disjoint languages is not a tough task, distinguishing the languages originated from the same root (e.g., Persian and Arabic or Italian and English) is a difficult task.

The ability to identify the language of a document increases the accessibility of data. It has a vast range of applications, i.e., presenting information in a user's native language is critical in attracting website visitors [4]. Most of the text processing techniques presuppose that the document's language is known. However, in real-world data, automatic LID is required to identify the language of the document .

The rhythm of expression is different in languages. It is created by the sequence of letters. Therefore, in this study, we will use the sequence of letters to identify the language. Of course, due to cultural and political issues, words from languages such as English, Arabic and French have infiltrated other languages. This makes language identification a bit difficult.

In the proposed method's training phase, the text is converted into a time series using UTF-8 coding. The time series is clustered into different clusters then analyzed using the Wavelet packet. The statistical features are extracted from each sub-band and used as the inputs of a multilayer perceptron neural network.

The proposed method is examined with our collected dataset. The provided dataset covers similar languages.

*Corresponding Author Email: alyan.nezhadi@mazust.ac.ir (M. M. AlyanNezhadi)

Please cite this article as: H. Hassanpour, M.M. AlyanNezhadi, M. Mohammadi, A Signal Processing Method for Text Language Identification, International Journal of Engineering, Transactions C: Aspects Vol. 34, No. 6, (2021) 1413-1418

Designing a system to distinguish between similar languages such as Serbian and Croatian [5], language varieties like European Portuguese and Brazilian [6], or a set of Arabic dialects [7] is more challenging than designing systems to discriminate between, for example, Finnish and Japanese [8, 9]. The experimental results show the ability of the proposed method for LID with a similar languages dataset.

The rest of the paper is organized as follows. In the next section, we review several litterateurs dealing with LID, then the proposed method is described. In Section 3, our dataset is introduced, and applying the proposed method to the dataset is given. Finally, the study is concluded in section 4.

2. LITERATURE REVIEW

Some approaches have been proposed in literature for LID based on frequent word counting, unique tokens and n-gram [3] in which features such as the presence of particular characters, words or n-grams [10] are used as discriminators.

In the case of frequent words counting [11], the language is identified based on the frequencies of the words in the predefined dictionary constructed per each language. Another approach is based on n-gram. The n-gram is a contiguous sequence of items from a given text. There are some words with higher frequency for each human language than others, which can be used as discriminator feature. Ng and Selamat [12] studied three n-grams based identification method, i.e. , distance measurement, Boolean technique and optimum profile technique. In the first method, the profiles are produced and sorted based on n-gram frequencies. The minimum distance between testing and training profiles is selected as the winner. In the Boolean technique, the matching rate between testing and training profiles is computed. The language of the text is identified based on the highest matching rate. The first approach suffers from dimensionality problem and the latter fails in the case of the same n-gram frequency for multiple languages. The last approach applies both frequency and position features. The language with minimum converged point is known as the text language [12].

Common words such as conjunctions, determiners, and prepositions can be used to extract LID features. Dunning [13] used byte level n-grams of the entire string instead of the word's character level n-grams. Although n-gram based methods provide high accuracy in LID, but these methods suffer from high order of time complexity [14]

N-gram based methods are the most common LID methods in the literature. Several methods in combination with this approach have been developed, like SVM [15], Naive Bayes [16], prediction partial matching (PPM) [17], deep learning [18] and a combination of multiple classifier [19]. There are also benchmarking solutions to the LID. Google compact language detector (CLD) and TextCat employ n-gram based method[1], LogR [20] uses a discriminative strategy with regularized logistic regression [16]. Cavnar and Trenkle [14] provided outstanding results compared to the other state-of-the-art methods. They used rank order statistic as distance measure. The weakness of this method is that it relies to the tokenization while many languages have no boundaries.

The languages with the same origin are very similar in appearance and n-grams. For example, Arabic and Persian languages are of the same origin. As shown in Figure 1, they are very similar to each other. As can be seen in Equation (), 50% of 2-grams (bigrams) from the two example texts are joint. In addition, 59% and 76% of the two sample texts, the Persian and Arabic are existed in intersection of 2-grams sample texts (Equations () and ()). Therefore, the n-gram based approaches are unable to distinct between these languages.

$$\frac{bigram_{Persian} \cap bigram_{Arabic}}{bigram_{Persian} \cup bigram_{Arabic}} = \frac{29}{58} = 0.5 \quad (1)$$

$$\frac{bigram_{Persian} \cap bigram_{Arabic}}{bigram_{Persian}} = \frac{29}{49} = 0.59 \quad (2)$$

$$\frac{bigram_{Persian} \cap bigram_{Arabic}}{bigram_{Arabic}} = \frac{29}{38} = 0.76 \quad (3)$$

Various LID systems exist in literature for identifying text language but with a limited number of languages to identify. Shekhar et al. [21] and Gupta et al. [22] proposed an LID system to recognize Hindi and English languages.

After a rigorous search we found no prior study that employed signal processing techniques to identify the language of a text. In this paper, we address the problem of LID from a signal processing perspective. Different languages have different tones. Someone can recognize the language of a conversation if has previously heard such a conversation, even tough he/she does not understand the concept. Different languages show different frequency characteristics. This fact also can be observed from the text indicating correspondent phrases in a language. For each language, there is a dependency between components of a sentence as well as components that construct the word. These dependencies can be observed using both Fourier transform and Wavelet transform methods. The Fourier transform is not a good choice as it provides more diversity for the spectrum representation of the texts from the same language.

[1] https://github.com/google/cld3. [Last visited. 2021]

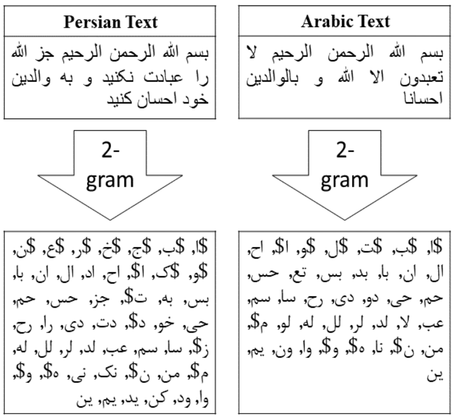
Persian Text
بسم الله الرحمن الرحیم جز الله را عبادت نکنید و به والدین خود احسان کنید
Arabic Text
بسم الله الرحمن الرحيم لا تعبدون الا الله و بالوالدين احسانا
2-gram
2-gram
$ا, $ب, $ج, $خ, $ر, $ع, $ن, $و, $ک, ا$, اح, اد, ال, ان, با, بس, به, ت$, جز, حس, حم, حی, خو, د$, دت, دی, را, رح, ز$, سا, سم, عب, لد, لر, لل, له, م$, من, ن$, نک, نی, ه$, و$, وا, ود, کن, ید, یم, ین
$ا, $ب, $ت, $ل, $و, ا$, اح, ال, ان, با, بد, بس, تع, حس, حم, حی, دو, دی, رح, سا, سم, عب, لا, لد, لر, لل, له, لو, م$, من, ن$, نا, ه$, و$, وا, ون, یم, ین

Figure 1. Samples of Persian and Arabic texts. As can be seen these languages are very similar to each other both in appearance and 2-grams

3. THE PROPOSED METHOD

Difference between languages is mainly due to their tunes. A language's tune depends on the succession of letters in a word and the pronunciation of successive words in a statement. Once we listen to a speech, we initially recognize the language and subsequently understand its meaning depending on the language's grammar. We may not fully understand a speech with a low volume level or low Signal to Noise (SNR) ratio, but we can still recognize the language spoken in if we know the language.

Once a French speaker talks English, the English speech is understandable even with a French accent. Indeed, speech in a language is a succession of letters that construct a tune. We propose a signal processing technique to convert text to time series and extract the tune for language identification. The flowchart of the algorithm is depicted in Figure 2.

In the proposed method, texts are converted to time series. The texts can be converted to time series using suitable coding. All texts in our dataset are coded with UTF-8.

Characters such as @,-,+ and # may exist in different texts. Therefore, they are removed from the time series . The obtained signal is clustered into several clusters using K-means method. The only feature for clustering is the mean of the signals UTF-8 codes. Some languages having unique UTF-8 codes are clustered in groups with only one language per group. However, the languages with similar UTF-8 codes appear in the same cluster. Thus, further processing is required to detect the language of texts in each cluster. To this aim, for each cluster, a model is trained and tested. The steps for the training are: I) feature extraction and II) classification.

Feature Extraction: The clustered signals are analyzed using the Wavelet packet transform, which is an extension of the Wavelet transform provided with more information regarding both high and low-frequency bands of the signal. By using the Wavelet packet transform, each signal is decomposed into 32 sub-bands.

The median of partial energy related to the sub-band coefficients is extracted from each sub-band as a feature.

$$F_x = \log\left(\left|median\left((x^2)\right)\right|\right), \tag{4}$$

where, x is the sub-band coefficients. This feature is used for language identification introduced by AlyanNezhadi et al. [23].

Classification: Many techniques in literature employ neural networks in their classification [24-26]. In this paper, the extracted features are fed into multilayer perceptron neural network with one hidden layer and the network is trained to classify the languages. For testing a new text, first, the cluster of the text is determined. If the cluster contains only one language, that language will be assigned to the text. Otherwise, the correspondent model will be used to detect the language in the cluster.

4. DISCUSSION AND RESULTS

In the following first, we describe the dataset we have prepared, then the obtained results are discussed.

4. 1. Dataset We have selected 31 languages to assess the performance of the proposed method. For each language, 1000 texts were randomly extracted from Wikipedia. The minimum length for each text is 5000 characters. The languages include English (en), France

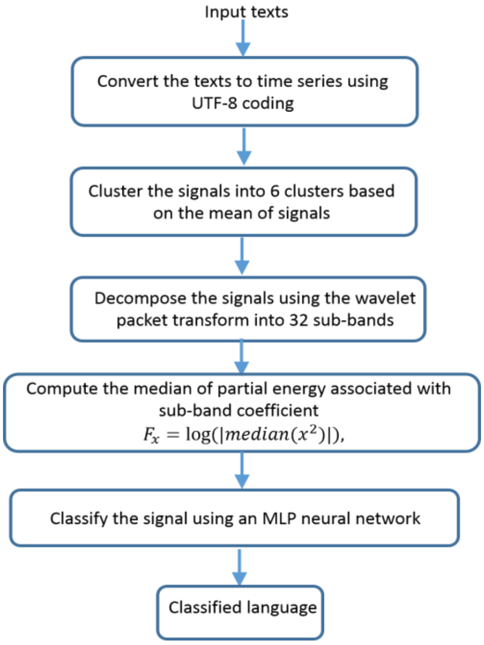
Input texts
Convert the texts to time series using UTF-8 coding
Cluster the signals into 6 clusters based on the mean of signals
Decompose the signals using the wavelet packet transform into 32 sub-bands
Compute the median of partial energy associated with sub-band coefficient
$F_x = \log(|median(x^2)|)$,
Classify the signal using an MLP neural network
Classified language

Figure 2. The flowchart of the proposed method

(fr), Italy (it), Arabic (ar), Russian (ru), arz (Egyptian Arabic), Azerbaijan (az), Byelorussian (be), Bulgarian (bg), Catalan (ca), ckb (Sorani, Central Kurdish), Czech (cs), Deutsch (de), Espernato (eo), Espanish (es), Persian (fa), Finnish (fi), Galician (gl), Hebrew (he), Hindi (hi), Croatian (hr), Indonesian (id), Dutch (nl), Polish (pl), Pashto (ps), Portoguese (pt), Romanian (ro), Tamil (ta) and Turkish (tr). Therefore, there is 310000 texts from 31 different languages.

4. 2. Evaluation To evaluate the performance of the proposed method, we used 31000 texts from our dataset. 80% of the texts are used for clustering, training and validation, and the remaining 20% of texts is used for testing randomly. The texts are converted to time series using UTF-8 coding and then the common characters are removed from them .

In the first stage, the K-means is used to cluster the whole data into six clusters based on only one simple feature (the average of the signals after elimination of common characters in the time domain). Table 1 shows the clustered languages in six groups with group centers and clustering precisions. As shown in Table 1, the texts with languages Tamil and Hindi are identified in this step with accuracy 90.50% and 95.50% for the testing dataset.

The average accuracy of 95.14% is achieved for six clusters with the centers: 1246.56, 878.73, 1370.18, 2820.96, 1874.38, 107.67 for testing data.

The obtained centers are used to cluster the test data. In the second stage, the texts from each cluster are separately classified. To this aim, the training texts (without elimination of common characters such as '?', '.', '!') are analyzed and decomposed using the Wavelet packet transform into 32 sub-bands. We have selected the Wavelet packet transform to analyze the texts. We employ a Wavelet packet transform with Daubachies kernel and five levels of decomposition. The Daubachies kernel is selected as it has valuable characteristics, i.e., vanishing moment and orthogonality conditions. We focussed on the sub-band energy to extract a feature as the classical multidimensional scaling (CMDS) representation of sub-band energy shows a separable cluster of different languages. This fact is shown in Figure 3 for seven languages from the dataset. The CMDS is a geometrical representation of data structure. Experimentally, we observed that the magnitude of the median provides more discrimination than the mean feature. We applied the logarithm to the median's magnitude to expand the distance between the languages with a close feature.

The multi-layer perceptron neural network with the parameters specified in Table 2; which is used to classify the languages based on the extracted features. The network has 32 input nodes as the length of the feature vector is 32. The number of neurons in the hidden layer is set equal to input layer. The network is trained 10 times and the results are averaged because the MLP may provide different results with different initial point. Table 3 shows the classification results for different states where 5, 7, and 12 spaces are inserted in-between two consecutive words of testing dataset. Finally, the accuracy of the system (clustering and classification) is given in Table 4. Figure 4 shows the structure of the experiment.

TABLE 1. Clustering the data into six clusters

Cluster members	Cluster centre	Accuracy (%)
ar, arz, ps	1246.56	87
ru, be, bg	878.73	94.83
fa, ckb	1370.18	77.50
ta	2820.96	90.50
hi	1874.38	95.50
en, fr, it, az, ca, cs, de, eo, es, fi, gl, he, hr, id, it, nl, pl, pt, ro, tr	107.67	98.55

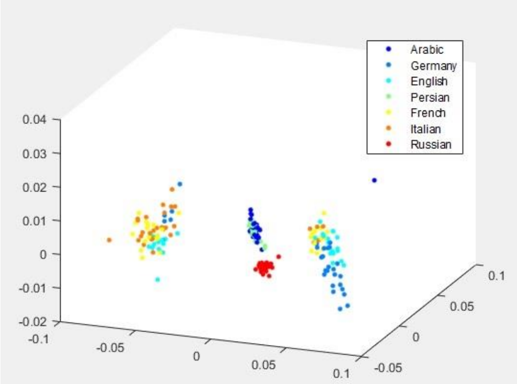

Figure 1. The CMDS representation of sub-band energy (with correlation distance) for seven languages

TABLE 2. The parameters of the neural network

Parameter	Value
Input layer neurons	The size of features
hidden layer neurons	The size of features
Output layer neurons	The number of languages
Hidden neurons activation function	Hyperbolic tangent sigmoid transfer function
Output neurons activation function	TSoft max transfer function
Data division	Random (80% train, and 20% validation data)
Maximum number of epochs	5000
Training method	Scaled conjugate gradient

TABLE 3. The accuracy of classification for testing data

Cluster members	**1 space**	**5 spaces**	**7 spaces**	**12 spaces**
ar, arz, ps	71.95	73.35	72.60	71.52
ru, be, bg	64.85	67.98	72.65	71.10
fa, ckb	75.65	76.88	76.77	77.12
en, fr, it, az, ca, cs, de, eo, es, fi, gl, he, hr, id, it, nl, pl, pt, ro, tr	62.72	67.72	69.38	69.50

TABLE 4. The accuracy of language identification

	1 space	**5 spaces**	**7 spaces**	**12 spaces**
Proposed method	66.88	70.70	72.20	72.02

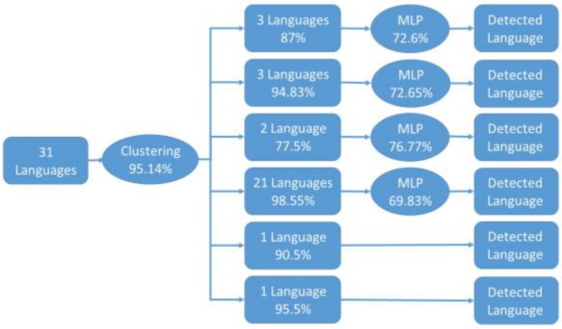

Figure 4. Structure of the designed system

5. CONCLUSION

LID plays an important role in most of the text processing applications. As this task is the first step to almost any text processing technique, the errors made in this task will propagate and deteriorate the results in the latter stages. In this paper, a new signal processing based technique was proposed to identify the text languages without any dictionary necessity. The proposed method includes the preprocessing, clustering, feature extraction, and classification stages. The proposed method was tested on our dataset with 31 different languages. Similar languages with the same origin exist in our dataset. The accuracy of 72.20% was achieved for text language identification.

6. REFERENCES

1. Cai, W., Cai, Z., Liu, W., Wang, X. and Li, M., "Insights in-to-end learning scheme for language identification", in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, (2018), 5209-5213, DOI: 10.21437/Interspeech.2019-1386
2. Sharma, A.R. and Kaushik, P., "Literature survey of statistical, deep and reinforcement learning in natural language processing", in 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE, (2017), 350-354, DOI: 10.1109/CCAA.2017.8229841
3. Ambikairajah, E., Li, H., Wang, L., Yin, B. and Sethu, V., "Language identification: A tutorial", ***IEEE Circuits and Systems Magazine***, Vol. 11, No. 2, (2011), 82-108, DOI: 10.1109/MCAS.2011.941081
4. Kralisch, A. and Mandl, T., "Barriers to information access across languages on the internet: Network and language effects", in Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), IEEE, (2006), 54b-54b, DOI: 10.1109/HICSS.2006.71
5. Radford, W. and Gallé, M., "Discriminating between similar languages in twitter using label propagation", ***arXiv preprint arXiv:1607.05408***, (2016), DOI: arXiv:1607.05408
6. Castro, D., Souza, E. and De Oliveira, A.L., "Discriminating between brazilian and european portuguese national varieties on twitter texts", in 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), IEEE, (2016), 265-270, DOI: 10.1109/BRACIS.2016.056
7. Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A. and Tiedemann, J., "Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task", in Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), (2016), 1-14, DOI:
8. Zissman, M.A. and Berkling, K.M., "Automatic language identification", ***Speech Communication***, Vol. 35, No. 1-2, (2001), 115-124, DOI: 10.1016/S0167-6393(00)00099-6
9. Kosmajac, D. and Keselj, V., "Slavic language identification using cascade classifier approach", in 2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH), IEEE, (2018), 1-6, DOI: 10.1109/INFOTEH.2018.8345541
10. Martins, B. and Silva, M.J., "Language identification in web pages", in Proceedings of the 2005 ACM symposium on Applied computing, (2005), 764-768, DOI: 10.1145/1066677.1066852
11. Bangalore, S. and Rambow, O., "Corpus-based lexical choice in natural language generation", in Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, (2000), 464-471,
12. Ng, C.-C. and Selamat, A., "Improving language identification of web page using optimum profile", in International Conference on Software Engineering and Computer Systems, Springer, (2011), 157-166, DOI: 10.1007/978-3-642-22191-0_14
13. Dunning, T., "Statistical identification of language": Computing Research Laboratory, New Mexico State University Las Cruces, NM, USA, (1994).
14. Cavnar, W.B. and Trenkle, J.M., "N-gram-based text categorization", in Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, Citeseer, (1994), DOI: 10.1.1.53.9367
15. Bhargava, A. and Kondrak, G., "Language identification of names with svms", in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, (2010), 693-696, DOI: 10.5555/1857999.1858101
16. Lui, M. and Baldwin, T., "Langid. Py: An off-the-shelf language identification tool", in Proceedings of the ACL 2012 system demonstrations, (2012), 25-30.
17. Bobicev, V., "Native language identification with ppm", in Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, (2013), 180-187.
18. Duvenhage, B., "Short text language identification for under resourced languages", ***arXiv preprint arXiv:1911.07555***, (2019), DOI: https://arxiv.org/abs/1911.07555
19. Carter, S., Weerkamp, W. and Tsagkias, M., "Microblog language identification: Overcoming the limitations of short, unedited and

idiomatic text", ***Language Resources and Evaluation***, Vol. 47, No. 1, (2013), 195-215, DOI: 10.1007/s10579-012-9195-y

20. Bergsma, S., McNamee, P., Bagdouri, M., Fink, C. and Wilson, T., "Language identification for creating language-specific twitter collections", in Proceedings of the second workshop on language in social media, (2012), 65-74,
21. Shekhar, S., Sharma, D.K. and Beg, M.M.S., "Language identification framework in code-mixed social media text based on quantum lstm — the word belongs to which language?", ***Modern Physics Letters B***, Vol. 34, No. 06, (2020), 2050086, DOI: 10.1142/S0217984920500864
22. Gupta, Y., Raghuwanshi, G. and Tripathi, A., "A new methodology for language identification in social media code-mixed text", in International Conference on Advanced Machine Learning Technologies and Applications, Springer, (2020), 243-254, DOI: 10.1007/978-981-15-3383-9_22
23. AlyanNezhadi, M. M., Forghani, M. and Hassanpour, H., "Text language identification using signal processing techniques", in 2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS), IEEE, (2017), 147-151, DOI: 10.1109/ICSPIS.2017.8311606
24. Pradeep, J., Srinivasan, E. and Himavathi, S., "Neural network based recognition system integrating feature extraction and classification for english handwritten", ***International Journal of Engineering, Transactions B: Applications***, Vol. 25, No. 2, (2012), 99-106, DOI: *10.5829/idosi.ije.2012.25.02b.03*
25. Akbari Foroud, A. and Hajian, M., "Discrimination of power quality distorted signals based on time-frequency analysis and probabilistic neural network", ***International Journal of Engineering***, ***Transactions C: Aspects***, Vol. 27, No. 6, (2014), 881-888, DOI: *10.5829/idosi.ije.2014.27.06c.06*
26. Hamidi, H. and Daraee, A., "Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases", ***International Journal of Engineering, Transactions A: Basics***, Vol. 29, No. 7, (2016), 921-930, DOI: *10.5829/idosi.ije.2016.29.07a.06*

Persian Abstract

چکیده

چکیده: تشخیص زبان متن یک مرحله مهم قبل از هرگونه پردازش زبان طبیعی است. در این مقاله، یک روش مبتنی بر پردازش سیگنال برای تشخیص زبان متن پیشنهاد شده است. توالی کاراکترها در یک کلمه و ترتیب کلمات، زبان متن را مشخص می‌کند. توالی کاراکترها در متن می‌تواند یک امضا برای متن باشد که بتوان بدون فهمیدن معنای آن‌ها، زبان متن را تشخیص داد. این امضا می‌تواند به کمک روش‌های پردازش سیگنال از طریق تبدیل متن به سری زمانی استخراج شود. اگرچه پژوهش‌ها و نرم افزارهای تجاری متعددی برای تشخیص زبان متن وجود دارد، ولی آن‌ها به یک دیکشنری استاندارد برای هر زبان نیاز دارند. در این مقاله، یک روش بدون نیاز به دیکشنری با سه مرحله اصلی ۱) پیش پردازش، ۲) خوشه‌بندی و در نهایت ۳) دسته‌بندی پیشنهاد شده است. در اولین مرحله، متن به یک سری زمانی با کمک کدگذاری UTF-8 تبدیل شده است. در مرحله دوم، به منظور گروه‌بندی زبان‌های مشابه یکدیگر، خوشه بندی سری‌های زمانی انجام شده است. در مرحله سوم، سری‌های زمانی هر خوشه به ۳۲ زیرباند توسط تبدیل موجک تجزیه شده است و از زیرباندها ۳۲ ویژگی استخراج شده است. سپس از شبکه عصبی پرسپترونی چند لایه برای دسته‌بندی ویژگی‌های استخراج شده استفاده شده است. روش پیشنهادی بر روی پایگاه داده خودمان با ۳۱۰۰۰ متن از ۳۱ زبان مختلف آزمایش شده است. روش پیشنهادی دارای دقت ٪۷۲.۲ برای تشخیص زبان متن است.