2012 International Conference on Advanced Computer Science Applications and Technologies

Identifying the Dominant Language of Web Page using Supervised *N*-grams

Choon-Ching Ng, Siau-Chuin Liew, Wan Muhammad Syahrir Wan Hussin and Tutut Herawan
Faculty of Computer Systems & Software Engineering
Universiti Malaysia Pahang
Email: choonching5u@gmail.com, liewsc@ump.edu.my, wmsyahrir@ump.edu.my, tutut@ump.edu.my

***Abstract*—Natural language processing is an emerging technology in linguistic industry and an aid to human-computer interaction in computer science. Language identification, on the other hand, is a form of pattern recognition that helps to identify predefined language of a web page and to predict the unknown language of one particular text. Written texts are constructed by common features such as character, word and *n*-gram and these characteristics are unique among languages. From the experiment result, the performance of the supervised n-gram produces an accurate identification value and outperforms the distance measurement on Arabic script web pages.**

***Keywords*-Support vector machine, supervised *N*-grams, language identification, Arabic script**

I. Introduction

In a multilingual application, language identification (LID) is the first process for determining the language of a particular text. For example, it is quite common that Internet-based businesses or communities such as eBay, Amazon, Google, Yahoo, Facebook, etc. operate sites in diverse languages [1]. The first step to be taken in a machine translation is language detection. In addition, a text-to-speech system required the desired language of source text in order to navigate the appropriate pronunciation and phrasing strategies. It becomes increasingly important in natural language processing. Parallel phoneme recognition followed by language mode is one of the most successful systems for language identification. The system depends on phonetic n-grams to classify languages [2]. Similar algorithm also being implemented in optical character recognition [3].

LID is the process of determining the predefined language automatically from a given content (e.g., English, Malay, Chinese, Japanese, Arabic, etc.). Language is an indispensable tool for human communication, and presently, the language that dominates the Internet is English. A web page is a kind of digital document displayed on a web browser. The web page can be written using diverse languages or different encoding such as Unicode [4].

Botha and Barnard have stated that the accuracies of text-based language identification are depended on several factors, including the size of the text fragment to be identified, the amount of training data available, the classification features and algorithm employed, and the similarity of the languages to be identified [5], [6]. A language identifier usually can produce higher identification accuracy with lower computational memory and shorter processing time. Mislabeled training documents will also affect the results of language identification [7]. Biemann and Teresniak have argued that supervised training has a major drawback. The language identifier will fail on languages that are not contained in its training, and it will, for the most part, have no way to cope with that and may assign the data to some arbitrary or unknown language [8]. However, there is a possibility to use unsupervised learning method for solving this issue.

In most of the studies, some common words and character and word n-grams are accepted as the features that best distinguish different languages. Since language identification is a classification problem, several statistical and machine learning techniques used in text classification have been successfully applied to this problem. For example, language identification is needed when users look for specific information that is written in Arabic scripts in which words are similar to other languages. For this reason, many researchers undertook studies to automatically identify the language in which information is written on a web document [9]. The information technology industry has been motivated to deal with problems related to language processing and identification throughout the world [10].

In this paper, integration between normalized bigrams and support vector machine is proposed to create a novel language identification method that detects the language of web pages. The first phase is to find out the interesting features of all languages of one particular script. Then, the normalized vectors are integrated with support vector machine to predict the language of a text.

This paper is organized as follows. Section 2 provides a background of LID. Section 3 discusses the experiment methodology and follows by the result analysis and discussion in Section 4. Finally, the conclusion of this work is summarized in Section 5.

II. Related Work

The field of human language technology covers a number of research activities, including the coding, identification, interpretation, translation, and generation of language. The aim of such research is to enable humans to communicate with machines using natural language skills. Language

978-0-7695-4959-0/13 $25.00 © 2013 IEEE
DOI 10.1109/ACSAT.2012.74

technology research involves many disciplines, such as linguistics, psychology, electrical engineering and computer science. Cooperation among these disciplines is needed to create multimodal and multimedia systems that use the combination of text, speech, facial cues, and gestures, both to improve language understanding and to produce a more natural language processing by animated characters [11].

Language technologies play a key role in the age of information [12]. Today, almost all device systems combine language understanding and generation that allow people to interact with computers using text or speech to obtain information, to study, to do business, and to communicate with each other effectively. The technology convergence in the processing of text, speech, and images has led to computer's ability to make sense of the massive amounts of information now available via internet connection. For example, if a student wants to gather information about the art of getting things done, he or she can set in motion a set of procedures that locate, organize, and summarize all available information related to the topic from books, periodicals, newspapers, etc. Translation of texts or speech from one language to another is needed to access and interpret all available materials and present it to the student in his or her native language. As a result, it will increase academic interests of the student [13], [14].

The volume of social network posts available on the internet has expanded dramatically in recent years. While the vast amount of user-generated text from social media is available online, effective language-processing technology is difficult without knowing which language is being processed. Bergsma et al. [15] have illustrated the language identification performance on twitter text in low-resource languages and created language-specific twitter collections in non-Latin scripts for the experimental purpose. Multilingual posts can potentially affect the outcomes of content analysis on microblog platforms. Carter et al. [16] have proposed a character n-gram distance metric in order to identify the five microblog characteristics of language identification, which include blogger, link, mention, tag and conversation. They have releaved that there are four main categories of errors include fluent multilingual posts, prior effects, named entity errors, and language ambiguity. Mayer [1] has presented a simple, stable, and highly runtime efficient algorithm for language identification, which does not need any human-labeled training data, provided there already exists a strong language identifier based on a site language or user language. Although the proposed method works well on short text, the result is arguable due to the strong dependency on site language. Centroid-based classification is a machine learning approach used in the text classification domain. Conventional machine learning approaches to language identification perform very well on long documents using standard language, but relatively poorly on short documents with non-standard orthography.

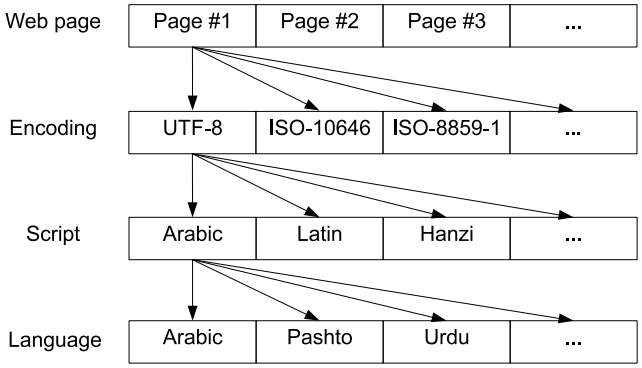

Figure 1. Encoding, script and language identification [11].

This is getting worse on social media sources like Twitter. Vogel and Tresner-Kirsch [17] have proposed the modified LIGA algorithm and shown that a modied algorithm achieves 99.8% accuracy disambiguating among six European languages. However, the analysis of the results does not include the data sets from different languages such as Urdu, Arabic, Persian, etc. Takçı and Güngör [18] have proposed a novel method named as inverse class frequency to increase the quality of the centroid values. They also used a feature set formed of individual characters rather than words or n-gram sequence in order to decrease the training and classification times. However, the stated method does not include the non-European languages. Lui and Baldwin [19] have demonstrated the problem of negative transfer in training a system using language labeled data from a variety of domains. They have developed a method for identifying features that are strongly predictive of language across multiple domains by examining the difference in information gain of each feature, but the performance does not include the analysis on minor languages.

Many countries have diverse multilingual population and multiple official languages (e.g., India, Singapore, Malaysia, etc.). According to *Internet World Stats*, internet use had a more explosive growth between 2000 and 2007 in Middle Eastern countries (e.g., Iran, Saudi Arabia, Syria, Yemen, etc.) than in the rest of the world [20], [10]. In addition, the Summer Institute of Linguistics has reported that there are 69 kinds of languages (including English) spoken or used by more than 10 million people in the world [13]. It is very time-consuming if language identification is done manually. Figure 1 depicts the general framework of web page LID [11]. It is a relationship between web page, encoding, script and languages. With the increasing number of web pages on the Internet, it has become a need to provide an intelligent method to effectively classify those information.

III. *N*-GRAMS SUPPORT VECTOR MACHINE

Figure 2 illustrates the adopted processes in this work involve extracting region of interest (ROI), *n*-gram rank

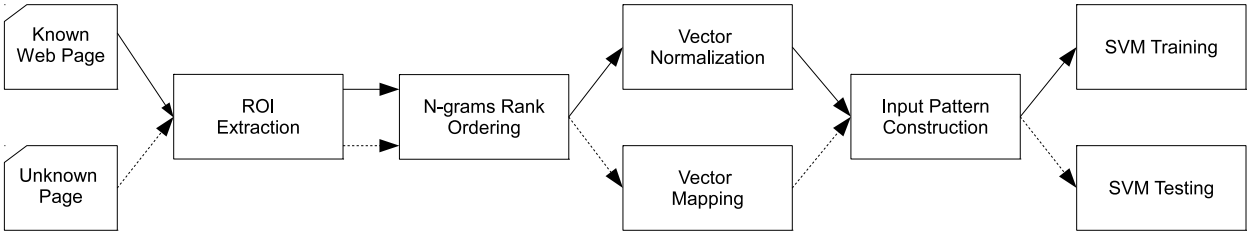

Figure 2. The process of language identification.

ordering, vector normalization or mapping, input pattern construction and LID using support vector machine.

A. ROI Extraction

In data set preparation, one thousand web pages of each language have been collected from the news website. This research is focusing on identifying languages of Arabic script such as Arabic, Persian, Urdu and Pashto. These data sets are between two and ten kilobytes, which were saved in Unicode form by setting the file name corresponding to the language. During the noise filtering process, irrelevant strings have been removed, which are not in the range of Arabic script in Unicode (0600-06FF). Finally, the ROI is ready to be processed by *N*-gram. Details of data set preparation can be found in [21].

B. N-gram Rank Ordering

A large number of approaches to LID have been proposed in the literature. The primary focus is on methods that do not require extensive linguistic expertise. Such approaches, which are statistical in nature, have been shown to perform competitively and are much easier to develop when appropriately codified linguistic knowledge is not available. Statistical language models can be built from counts of words or letters, or from *n*-gram statistics. The sub sequence of characters is referred to as an *n*-gram that is commonly used to predict desired language of a text, moreover, due to their uniqueness, they can be utilized as the feature of other language applications like speech recognition for a further increase to the reliability measures. The *n*-gram based models outperform a word based model for small text fragments and do equally well for larger fragments [22]. In another study *n*-gram achieved better results than string kernels [23].

C. Vector Normalization / Mapping

During the SVM training, vector normalization has been used to find out the standard *n*-gram to appear among languages. Meanwhile, vector mapping is to select the normalized standard *n*-gram for input pattern construction. For vector normalization, it is to filter out and discard the meaningless *n*-gram to a certain level of threshold. If one language is lacked of that *n*-gram, then it will be discarded from the input pattern.

D. Input Pattern Construction

In general, input to be used by machine learning method is in the matrix form. Therefore, the input pattern is constructed based on the normalized vector. Size of the input pattern is equivalent to size of normalized vector. Input patterns of training and testing set are generated separately. Values are the corresponding *n*-gram based rank order statistics that standardize between minus one and one.

E. Language Identification using SVM

As a popular tool for discriminative classification, SVM has been widely used in several areas for pattern recognition and brought significant improvements. LIBSVM is an integrated software or library for support vector classification, regression and distribution estimation (one-class SVM). It also supports multi-class classification and has been selected as the machine learning method in this work [24]. As discussed previously, the size of the feature space grows exponentially with *n*,which leads to long training times and extensive resource usage as *n* becomes large; therefore, the SVM experiments has been limited to *n*=2 or so-called bigrams. A language model was built with samples of size from a training set. These samples contained a frequency count of each *n*-gram combination in the character string. Thus, the feature dimension of the SVM is equal to the number of *n*-gram combinations. Samples of the testing set are created using the same size of the character window as used to build the model. After training the SVM language model, the test samples can be classified according to language. A comparison between the SVM classification and conventional method, distance measurement (DM), proposed by Cavnar and Trenkle [25] has been investigated.

IV. Results

Figure 3 shows the experimental results on 100 testing web pages that vary on training data size. Initially, only one web page has been selected randomly as training data in which each language consists of ten kilobytes (KB). Therefore, the total training data is 40 kilobytes for four languages. Accuracy of SVM and DM is 79% and 0.50%, respectively. Then, the SVM produced a rapid fluctuation to 97% on 238KB if compared with DM. Between 505KB and 2415KB, SVM achieved between 95% and 99%, meanwhile

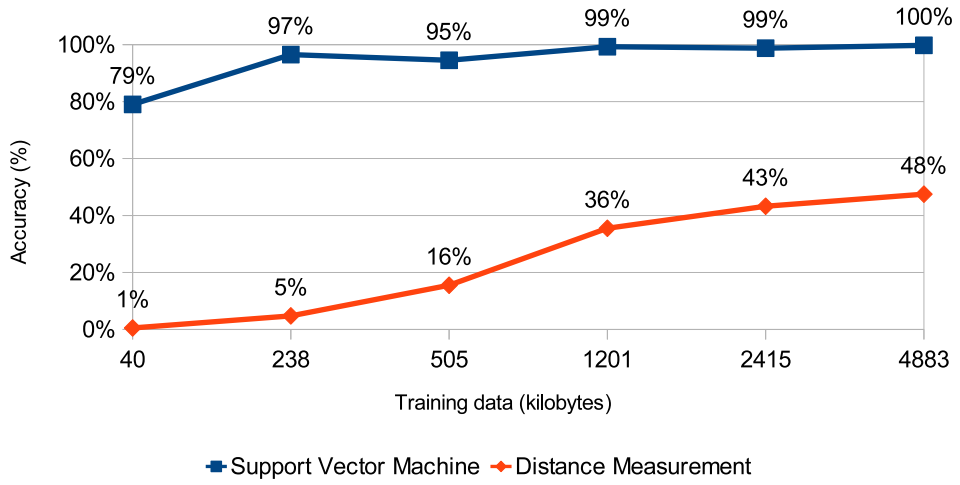

Figure 3. Identification results on 100 testing web pages.

DM only presented between 16% and 43%. At the training size of 4883KB, SVM has outperformed than DM with only one misclassification web page and the accuracy is almost 100%.

In this work, the effect of the training data size has been examined. From the experiment, it has been found that the amount of training data required for asymptotic performance depended most strongly on the size of the training data. Integration between the bigrams and SVM has been proven as a reliable supervised method on Arabic script web page LID. Although DM's accuracy is gradually increased when training data is getting larger, but the computation cost is not worth.

Supervised learning is the machine learning task of predicting a class from the labeled data. It performs very well if the given training data is sufficient. Supervised language identification has been proven that is workable on Arabic script web pages. However, it cannot correctly predict the arbitrary or unknown language. Therefore, in the future work, the proposed method can be extended to semi-supervised method for increasing the identification accuracy on real time application.

V. Conclusion

Language identification is important in a vast variety of the natural language processing system. If an information retrieval system needs to identify documents automatically with little or no human oversight, then it is important to produce a reliable system that is capable of operating at a high level of accuracy. This paper aims to optimize the performance of language identification based on supervised machine learning method instead of conventional statistical approach. Therefore, an improved method has been proposed which is the integration between bigrams and support vector machine. A result given has shown that it is significant in Arabic script web page language identification; it has surpassed the conventional method distance measurement in terms of accuracy. In the future work, an insight on implementing the proposed method will be explored into South African, Indian and European languages, which have more than 10 languages in one particular script.

Acknowledgment

This work is supported by the Department of Research & Innovations, Universiti Malaysia Pahang, under the vot RDU 120333. The authors are grateful to the anonymous reviewers for their valuable and insightful comments.

References

[1] U. Mayer, "Bootstrapped language identification for multi-site internet domains," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 579–585.

[2] Y. Deng and J. Liu, "Automatic language identification using support vector machines and phonetic n-gram," in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*. IEEE, 2008, pp. 71–74.

[3] A. Popat and E. Brevdo, "Image-domain script and language identification," Jul. 31 2012, uS Patent 8,233,726.

[4] J. D. Allen, *The Unicode standard 5.0*. Addison-Wesley, 2006.

[5] G. Botha and E. Barnard, "Factors that affect the accuracy of text-based language identification," *Computer Speech & Language*, 2012.

[6] P. Sibun and J. C. Reynar, "Language identification: Examining the issues," in *Proceedings of the Symposium on Document Analysis and Information Retrieval 1996*, 1996, pp. 125–135.

[7] J. Zou, G. Chen, and W. Guo, "Chinese web page classification using noise-tolerant support vector machines," in *Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 2005, pp. 785–790.

[8] C. Biemann and S. Teresniak, "Disentangling from babylonian confusion-unsupervised language identification," in *Proceedings of Computational Linguistics and Intelligent Text Processing*, vol. 3406. Springer, 2005, pp. 762–773.

[9] A. Xafopoulos, C. Kotropoulos, G. Almpanidis, and I. Pitas, "Language identification in web documents using discrete hmms," *Pattern recognition*, vol. 37, no. 3, pp. 583–594, 2004.

[10] P. Payack, "The number of words in the english language," *Global Language Monitor http://www. languagemonitor. com/wst_page7. html (accessed on 28.06.2012)*, 2006.

[11] C. Ng and A. Selamat, "Improve feature selection method of web page language identification using fuzzy artmap," *International Journal of Intelligent Information and Database Systems*, vol. 4, no. 6, pp. 629–642, 2010.

[12] P. Constable and G. Simons, "Language identification and it: Addressing problems of linguistic diversity on a global scale," in *Proceedings of the 17th International Unicode Conference, SIL Electronic Working Papers*, San Jos, California, 2000, pp. 1–22.

[13] M. Z. Abd Rozan, Y. Mikami, A. Z. Abu Bakar, and O. Vikas, "Multilingual ict education: Language observatory as a monitoring instrument," in *Proceedings of the South East Asia Regional Computer Confederation 2005: ICT Building Bridges Conference*, vol. 46, Sydney, Australia, 2005, pp. 53–61.

[14] P. McNamee and J. Mayfield, "Character n-gram tokenization for european language text retrieval," *Information Retrieval*, vol. 7, no. 1, pp. 73–97, 2004.

[15] S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson, "Language identification for creating language-specific twitter collections," in *Proceedings of the 2012 Workshop on Language in Social Media*. Association for Computational Linguistics, 2012, pp. 65–74.

[16] S. Carter, W. Weerkamp, and M. Tsagkias, "Microblog language identification: overcoming the limitations of short, unedited and idiomatic text," *Language Resources and Evaluation*, pp. 1–21, 2012.

[17] J. Vogel and D. Tresner-Kirsch, "Robust language identification in short, noisy texts: Improvements to liga," in *Third International Workshop on Mining Ubiquitous and Social Environments*, 2012, pp. 1–9.

[18] H. Takçı and T. Güngör, "A high performance centroid-based classification approach for language identification," *Pattern Recognition Letters*, 2012.

[19] M. Lui and T. Baldwin, "Cross-domain feature selection for language identification," in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, 2011, pp. 553–561.

[20] M. M. GROUP *et al.*, "Internet world stats," *Usage and Population Statistics. URL: http://www. internetworldstats. com/stats. htm (accessed on 21.06. 2012)*, 2009.

[21] A. Selamat and C. Ng, "Arabic script web page language identifications using decision tree neural networks," *Pattern Recognition*, vol. 44, no. 1, pp. 133–144, 2011.

[22] G. Grefenstette, "Comparing two language identification schemes," in *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95)*, vol. 1, Rome, Italy, 1995, pp. 263–268.

[23] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *The Journal of Machine Learning Research*, vol. 2, pp. 419–444, 2002.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[25] W. Cavnar and J. Trenkle, "N-gram-based text categorization," in *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, USA, 1994, p. 161175.