Improving Language Identification of Web Page Using Optimum Profile

Choon-Ching Ng[1,⋆] and Ali Selamat[2]

[1] Faculty of Computer Systems & Software Engineering,
Universiti Malaysia Pahang,
Lebuhraya Tun Razak, 26300 Gambang, Kuantan Pahang, Malaysia
choonching@ump.edu.my
[2] Faculty of Computer Science & Information Systems,
Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
aselamat@utm.my

Abstract. Language is an indispensable tool for human communication, and presently, the language that dominates the Internet is English. Language identification is the process of determining a predetermined language automatically from a given content (e.g., English, Malay, Danish, Estonian, Czech, Slovak, etc.). The ability to identify other languages in relation to English is highly desirable. It is the goal of this research to improve the method used to achieve this end. Three methods have been studied in this research are distance measurement, Boolean method, and the proposed method, namely, optimum profile. From the initial experiments, we have found that, distance measurement and Boolean method is not reliable in the European web page identification. Therefore, we propose optimum profile which is using N-grams frequency and N-grams position to do web page language identification. The result show that the proposed method gives the highest performance with accuracy 91.52%.

Keywords: N-grams profile, rank-order statistics, distance measurement, Boolean method, optimum profile.

1 Introduction

Language identification is used frequently in a number of applications, such as machine translation, information retrieval, speech recognition, and text categorization. Among researches of text-based language identification, N-grams is perhaps the most widely used and studied [1]. The N-grams method, which is a sub-sequence of N objects from a longer sequence, when rank-order statistics on N-grams profile are adopted and the distance measurement is used to identify the predefined language of a particular content.

It is being argued that text-based language identification is a completely solved problem. However, we have found that improvements are still needed

⋆ Corresponding author.

J.M. Zain et al. (Eds.): ICSECS 2011, Part II, CCIS 180, pp. 157–166, 2011.

because of several problems arise when dealing with web page language identification. Firstly, the web pages contain multiple languages which may produce faulty output in related language systems. Secondly, web page language identification is difficult due to plethora of internation terms and proper names occuring in the internet. Other issues are web page format, encoding, spelling and grammar errors [2,3].

This paper is organized as follows: Related works on language identification is described in Section 2. Next, data preparation and language identification using the distance measurement, Boolean method, and optimum profile are explained in Section 3. The experimental results based on confusion matrix and accuracy are detailed out in Section 4. The conclusion of the research is given in Section 5.

2 Related Works

Human usually don't have any need for language identifiers, however the field of human language technology covers a number of research activities, such as 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 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 more natural language processing by animated characters [4,5].

Language technologies play a key role in the age of information [5]. 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 [6]. The technology convergence in the processing of text, speech, and images has lead to the particular ability to make sense of the massive amounts of information now available via computer networks. 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, and so on. Translation of texts or speech from one language to another is needed to access and interpret all available material and present it to the student in his or her native language. As a result, it increases academic interests of the student [6,7].

Some works have been reported to detect the language of a particular web page. They are decision tree neural networks [3], discrete HMMs [2], short letter sequences (N-grams) [8], and metadata description [9]. A variety of features have been used for language identification. These includes: the presence of particular characters [10,11], written words [12,13], and N-grams [1,14].

3 Method

In this section, we describe the distance measurement, Boolean method, and optimum profile as shown in Figure 1. First of all, data sets of languages have been collected from news website. Then, these data sets were saved in unicode form by setting the file name corresponding to the target language. Many types of encoding have been used on the web document to ensure that character processing is not miscalculated. We have converted the identified encoding into the unicode encoding as the latter is able to accommodate all encoding types by the use of specific numeric number. In this work, we have collected European web pages as experimental data sets such as Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish and Swedish, and the corresponding annotation are bul, cze, dan, dut, eng, est, fin, fre, ger, gre, hun, iri, ita, lat, lit, mat, pol, por, rom, slo, sle, spa, and swe, respectively. Each language consists of 1000 web pages, 100 units were used for training, 500 units were randomly selected from the remaining data sets as testing data. Threshold has been set to top 100 units of language features. N-grams were mixed by unigram, bigrams, and trigrams, but statistical anaysis was done independently.

Feature selection determines the appropriate features or attributes to be used in language identification. It is based on N-grams frequency (NF) and rank-order statistics [15]. For NF, it is based on the occurences of the particular N-grams (ngm) in a document, not the whole data set. The number of a particular ngm contributed to the document is an important factor in a language identification. For example, the ngm 'ber' appears in Malay more frequently than in English, so a Malay document has higher occurences of that ngm. The formula of NF_L is given by equation 1, where T_d is the total N-grams in that

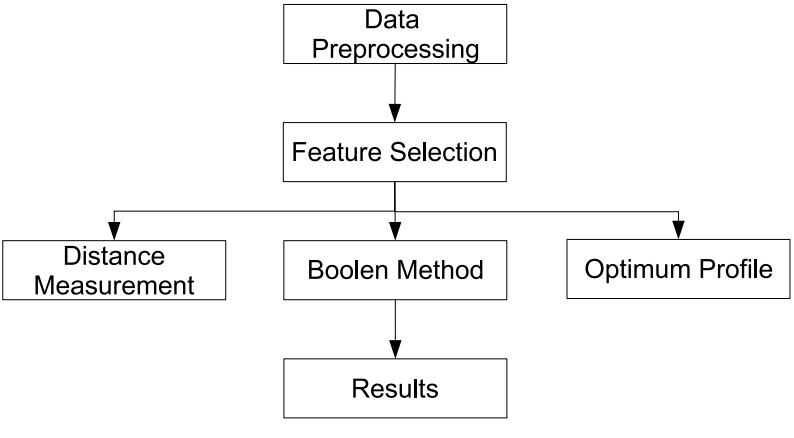

Fig. 1. Research framework of proposed method

document d, L is the target language, and D is the data set. Each unigram, bigrams, and trigrams calculation is done separately and also the highest NF is selected as features. Finally, ngm is sorted based on rank-order statistics. Details of experimental setup and measurement have been described in the paper of Selamat and Ng [3,16,17].

$$NF_L\left(ngm\right) = \sum_{d=1}^{D} \left(\frac{\sum ngm_d}{T_d} \right) \tag{1}$$

3.1 Distance Measurement

Distance measurement has been proposed by Cavnar and Trenkle [1], they have used rank-order statistics on N-grams profiles in order to find out closest profile as winner of language identification and text categorization. Figure 2 illustrates the distance measurement of N-grams profile. First, training profile of each language is generated from the desired data set by using N-grams frequency. After

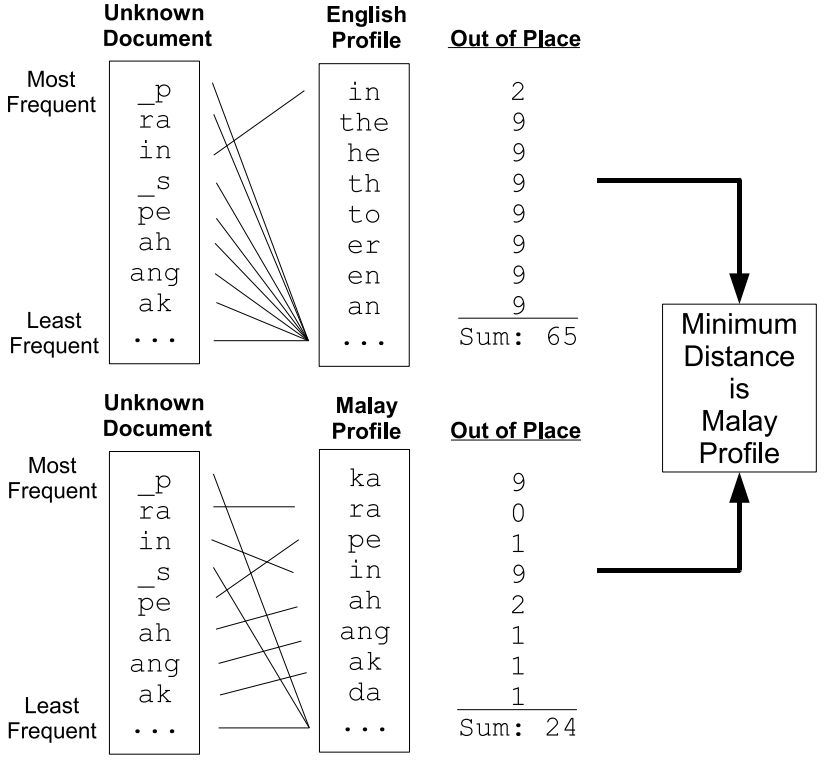

Fig. 2. Distance measurement of N-grams profile (note: it is assumed that for those not found N-grams in this figure is assigned with a maximum value nine) [1]

that, rank-order statistics are applied on the language profile to sort the N-grams from most frequent to least frequent. Same process goes on unknown document or target document. Then, unknown document profile is compared with all profiles of desired languages. Out of place is the distance between desired N-grams and target N-grams. Finally, minimum distance of one particular profile is selected as winner based on the sum out of place.

3.2 Boolean Method

Boolean method has been used to measure matching rates between target profile and training profile. It is different with distance measurement which is depends on N-grams frequency. Instead, this method returns value of one if one particular N-gram from the target profile is found on the desired profile. Otherwise, it returns value of zero if there is no match. After that, matching rate is derived by dividing the total Boolean value to total number of distinct N-grams in the target profile. Finally, the maximum matching rate is selected as winner among the training profiles.

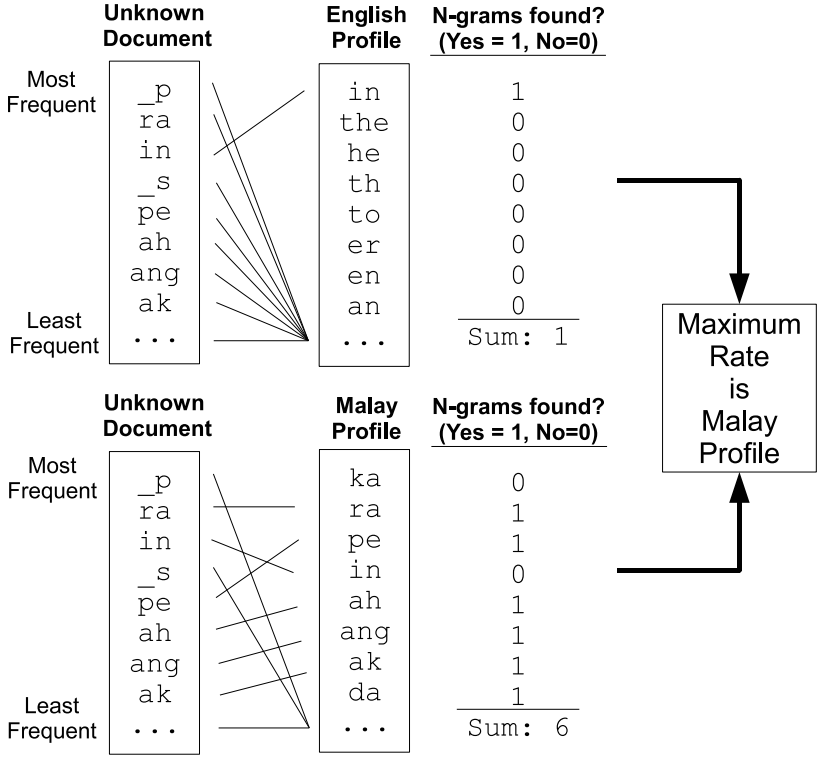

Fig. 3. Boolean method of N-grams profile (note: it is assumed that for those not found N-grams in this figure is assigned with a zero value) [18]

3.3 Optimum Profile

Figure 4 shows the example of proposed method N-grams optimum profile. This method makes use of N-grams frequency and N-grams position. Accumulated N-grams frequencies is the first identifier of language identification and it is followed by converge point which is to determine the fastest convergence of N-grams position. A random double is added to converge point to increase the level of discriminant. Indonesian N-grams profile consists of 'ka', 'ra', 'in', 'kan', and 'pe'; however Malay N-grams profile is comprised of 'ka', 'ra', 'pe', 'in', and 'ah'. Each N-grams frequency of Indonesian is 50, 60, 10, 0, 20, and Malay is 50, 60, 20, 10, and 0; while the accumulated frequencies are 50, 110, 120, 120, 140, 50, 110, 130, 140, and 140, respectively. Converge point of Indonesian and Malay are 4 and 3, respectively. Winner of this example is Malay due to the converge point is smaller than Indonesian.

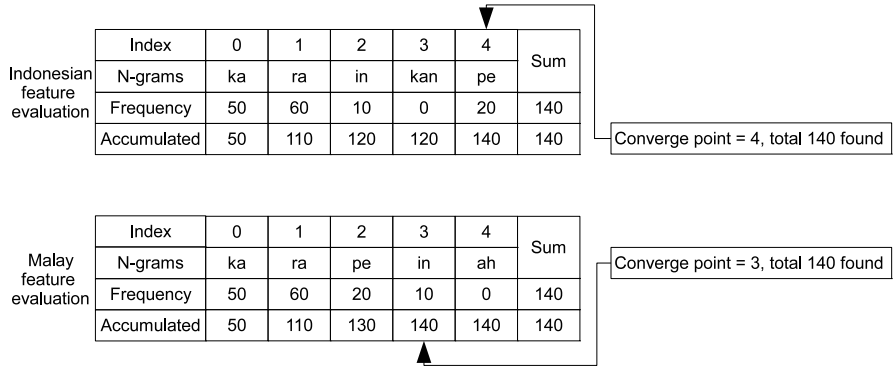

Fig. 4. Proposed N-grams Optimum Profile

4 Experimental Results

In this following subsections, we discuss the confusion matrix and accuracy of European web page language identification. Three methods have been evaluated which are distance measurement, Boolean method, and optimum profile. European data sets have been used in the experiment with total 23 languages. Threshold was set to top 100 features of each language.

4.1 Confusion Matrix of Web Page Language Identification

Table 1 shows the confusion matrix of European web page language identification using distance measurement. It is observed that the Bulgarian, Czech, Estonian, and Greek give worst results with correctly predicted samples are 0, 66, 1, and 4, respectively. Other languages more than 321 samples were correctly predicted. Finnish and Hungarian have achieved the best identification results which are 100% correctness.

Table 1. Confusion matrix of distance measurement on European web page language identification

		Predicted Language																						
		bul	cze	dan	dut	eng	est	fin	fre	ger	gre	hun	iri	ita	lat	lit	mat	pol	por	rom	slo	sle	spa	swe
Desired Language	bul	0	0	0	0	0	8	0	0	0	377	0	115	0	0	0	0	0	0	0	0	0	0	0
	cze	0	66	0	0	0	7	0	0	0	3	0	0	0	0	0	0	16	0	0	163	245	0	0
	dan	1	0	471	1	0	0	0	0	2	11	4	0	0	0	0	0	0	0	0	0	0	0	10
	dut	0	0	0	396	0	46	0	0	0	53	2	0	0	0	0	0	2	0	0	0	0	0	1
	eng	7	0	0	0	473	0	0	0	0	18	2	0	0	0	0	0	0	0	0	0	0	0	0
	est	8	3	23	1	4	1	21	2	0	233	117	5	0	6	1	29	36	6	2	0	1	0	1
	fin	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	fre	0	0	0	0	0	0	0	495	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
	ger	0	0	0	0	0	0	0	0	499	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	gre	55	12	3	1	0	59	219	1	6	4	3	31	2	22	0	38	9	0	2	14	18	0	1
	hun	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0
	iri	1	4	0	0	0	95	0	0	0	2	0	391	0	1	0	1	0	0	0	1	2	0	2
	ita	17	1	0	0	0	26	1	0	0	0	21	0	422	2	0	0	3	0	0	3	3	0	1
	lat	20	13	0	1	1	36	0	3	0	42	4	13	0	347	2	3	4	0	0	9	0	0	2
	lit	0	0	0	0	0	0	0	0	0	1	0	0	0	3	494	0	1	0	0	1	0	0	0
	mat	3	16	0	0	0	14	1	1	0	27	5	1	0	0	0	432	0	0	0	0	0	0	0
	pol	2	3	0	0	0	34	0	0	0	34	1	0	0	0	0	0	417	0	0	9	0	0	0
	por	8	0	0	0	0	6	0	0	0	15	1	1	1	0	0	0	4	463	0	0	0	1	0
	rom	2	1	0	1	1	15	2	1	0	9	7	0	12	0	0	2	4	0	442	0	0	0	1
	slo	6	17	0	2	0	59	0	0	1	77	0	0	0	0	1	0	13	0	0	321	3	0	0
	sle	0	0	0	0	0	1	0	0	0	2	0	1	0	0	0	0	25	0	0	49	422	0	0
	spa	1	0	0	0	14	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	482	0
	swe	9	4	1	1	0	12	0	1	2	82	21	0	0	0	0	5	6	0	0	1	0	0	355

Table 2. Confusion matrix of Boolean method on European web page language identification

		Predicted Language																						
		bul	cze	dan	dut	eng	est	fin	fre	ger	gre	hun	iri	ita	lat	lit	mat	pol	por	rom	slo	sle	spa	swe
Desired Language	bul	180	0	0	0	320	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	cze	0	45	0	0	1	0	0	0	0	0	0	0	16	0	0	0	0	0	5	0	432	1	0
	dan	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	dut	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	eng	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	est	0	0	21	2	17	11	247	0	0	0	0	0	119	0	62	0	0	0	8	0	0	0	13
	fin	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	fre	0	0	0	0	1	0	0	499	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ger	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	gre	2	0	0	0	4	0	0	0	0	494	0	0	0	0	0	0	0	0	0	0	0	0	0
	hun	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0
	iri	0	0	0	0	125	0	0	0	0	0	0	375	0	0	0	0	0	0	0	0	0	0	0
	ita	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0
	lat	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0
	lit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0
	mat	0	0	1	0	4	0	0	0	0	0	0	0	10	0	0	485	0	0	0	0	0	0	0
	pol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0
	por	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	498	0	0	0	0	0
	rom	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	498	0	0	0	0
	slo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	499	0	0	0
	sle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0
	spa	0	0	0	0	302	0	0	2	0	0	1	0	1	0	0	0	0	0	1	0	0	193	0
	swe	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	499

Table 2 illustrates the confusion matrix of European web page language identification using Boolean method. Bulgarian, Czech, Estonian, and Spanish have achieved accuracy of identification below 50%. Total correctly predicted samples are 180, 45, 11, and 193, respectively. Irish has been predicted as English with 125 samples and the remaining are correct samples. Other languages give good results with more than or equal 485 correct samples. It is slightly better than distance measurement.

Table 3 depicts the results of European web page language identification by using optimum profile. It is noticed that the performance of identification has been increased with only two languages give worst results. They are language of Czech and Estonian. Only 45 samples of Czech and 9 samples of Estonian have been correctly predicted. The remaining 21 languages have correctly predicted more than 490 samples.

Table 3. Confusion matrix of optimum profile on European web page language identification

Desired Language \ Predicted Language	bul	cze	dan	dut	eng	est	fin	fre	ger	gre	hun	iri	ita	lat	lit	mat	pol	por	rom	slo	sle	spa	swe
bul	496	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
cze	0	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	427	28	0	0
dan	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dut	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
eng	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
est	0	0	111	1	1	9	324	0	0	0	0	0	0	2	23	0	0	0	0	0	1	0	28
fin	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
fre	0	0	0	0	0	0	0	498	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
ger	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0	0	0
gre	1	0	0	0	2	0	0	0	2	491	0	0	0	1	1	0	0	1	0	0	0	1	0
hun	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0	0	0
iri	0	0	0	0	0	0	0	0	0	0	0	496	4	0	0	0	0	0	0	0	0	0	0
ita	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0	0
lat	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0	0
lit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0	0
mat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0	0
pol	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0	0	0
por	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	499	0	0	0	1	0
rom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0	0
slo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0	0
sle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	500	0	0
spa	0	1	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	492	0
swe	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	499

4.2 Accuracy of Web Page Language Identification

Figure 5 shows the overall accuracy of European web page language identification. It is found that the accuracy of distance measurement, Boolean method, and optimum profile is 72.98%, 85.01%, and 91.52%. Distance measurement is having dimension problem in which similar distance might be found in more than one language and usually the smallest one is frequently encountered. Boolean

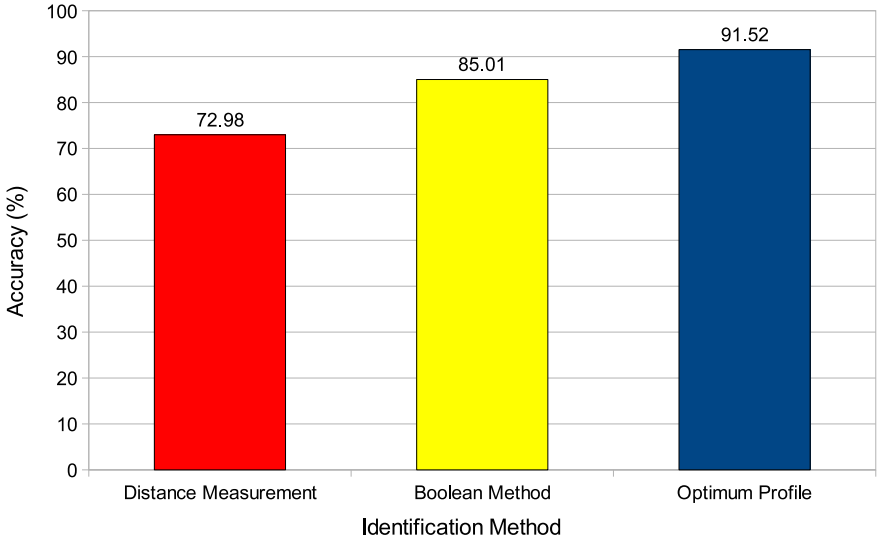

Fig. 5. Overall accuracy of European web page language identification

method is not reliable if two or more languages appear same N-grams frequency. Therefore, it has been proved that optimum profile gives the best performance in European web page language identification.

5 Conclusion

Language identification is important in a vast variety of natural language processing systems. If we are to trust an information retrieval system to classify documents with little or no human oversight, we require a system that is capable of operating at a high level of high accuracy. Therefore, we have proposed optimum profile to cope with the limitations found in both distance measurement and Boolean method. From the experiments, it is concluded that optimum profile performs better than others. In the future works, the issues of multilingual web page, noise tolerent of language identifier, minority languages, and data dimensionality will be investigated.

Acknowledgments. This work is supported by the Ministry of Higher Education (MOHE) and Research Management Center, Universiti Malaysia Pahang (UMP) and Universiti Teknologi Malaysia (UTM). The authors are also grateful to the anonymous reviewers for their valuable and insightful comments.

References

1. Cavnar, W., Trenkle, J.: N-gram-based text categorization. In: Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, USA, pp. 161–175 (1994)
2. Xafopoulos, A., Kotropoulos, C., Almpanidis, G., Pitas, I.: Language identification in web documents using discrete hmms. Pattern Recognition 37(3), 583–594 (2004)
3. Selamat, A., Ng, C.: Arabic script web page language identifications using decision tree neural networks. Pattern Recognition 44(1), 133–144 (2011)
4. Muthusamy, Y., Spitz, A.: Automatic language identification. In: Cole, R., Mariani, J., Uszkoreit, H., Varile, G., Zaenen, A., Zampolli, A. (eds.) Survey of the State of the Art in Human Language Technology, pp. 255–258. Cambridge University Press, Cambridge (1997)
5. Constable, P., Simons, G.: 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 José, California, pp. 1–22 (2000)
6. Abd Rozan, M.Z., Mikami, Y., Abu Bakar, A.Z., Vikas, O.: Multilingual ict education: Language observatory as a monitoring instrument. In: Proceedings of the South East Asia Regional Computer Confederation 2005: ICT Building Bridges Conference, Sydney, Australia, vol. 46 (2005)
7. McNamee, P., Mayfield, J.: Character n-gram tokenization for european language text retrieval. Information Retrieval 7(1), 73–97 (2004)
8. Martins, B., Silva, M.J.: Language identification in web pages. In: Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 764–768 (2005)

9. Simons, G.F.: Language identification in metadata descriptions of language archive holdings. In: Workshop on Web-Based Language Documentation and Description, Philadelphia, USA (2000)
10. Hakkinen, J., Tian, J.: N-gram and decision tree based language identification for written words. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 335–338 (2001)
11. Takcı, H., Soğukpınar, İ.: Letter Based Text Scoring Method for Language Identification. In: Yakhno, T. (ed.) ADVIS 2004. LNCS, vol. 3261, pp. 283–290. Springer, Heidelberg (2004)
12. Biemann, C., Teresniak, S.: Disentangling from babylonian confusion – unsupervised language identification. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 773–784. Springer, Heidelberg (2005)
13. Hammarstrom, H.: A fine-grained model for language identification. In: Workshop of Improving Non English Web Searching, Amsterdam, The Netherlands, pp. 14–20 (2007)
14. da Silva, J.F., Lopes, G.P.: Identification of document language is not yet a completely solved problem. In: Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce, pp. 212–219. IEEE Computer Society, Washington, DC, USA (2006)
15. Ng, C., Selamat, A.: Improve feature selection method of web page language identification using fuzzy artmap. International Journal of Intelligent Information and Database Systems 4(6), 629–642 (2010)
16. Selamat, A., Subroto, I., Ng, C.: Arabic script web page language identification using hybrid-knn method. International Journal of Computational Intelligence and Applications 8(3), 315–343 (2009)
17. Selamat, A., Ng, C.: Arabic script language identification using letter frequency neural networks. International Journal of Web Information Systems 4(4), 484–500 (2008)
18. Choong, C.Y., Mikami, Y., Marasinghe, C.A., Nandasara, S.T.: Optimizing n-gram order of an n-gram based language identification algorithm for 68 written languages. International Journal on Advances in ICT for Emerging Regions 2(2), 21–28 (2009)