Journal of King Saud University – Computer and Information Sciences (2016) **28**, 457–469

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Word-length algorithm for language identification of under-resourced languages

Ali Selamat [a,*], **Nicholas Akosu** [b]

[a] *UTM-IRDA Digital Media Center and Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM, Johor Bahru, Johor, Malaysia*
[b] *Software Engineering Research Group (SERG), Faculty of Computing, Universiti Teknologi Malaysia, Malaysia*

Received 22 July 2014; revised 27 October 2014; accepted 22 December 2014
Available online 28 November 2015

KEYWORDS

Language identification;
Under-resourced languages;
Resource-scarce;
Digital divide;
Spellchecker model

Abstract Language identification is widely used in machine learning, text mining, information retrieval, and speech processing. Available techniques for solving the problem of language identification do require large amount of training text that are not available for under-resourced languages which form the bulk of the World's languages. The primary objective of this study is to propose a lexicon based algorithm which is able to perform language identification using minimal training data. Because language identification is often the first step in many natural language processing tasks, it is necessary to explore techniques that will perform language identification in the shortest possible time. Hence, the second objective of this research is to study the effect of the proposed algorithm on the run-time performance of language identification. Precision, recall, and F_1 measures were used to determine the effectiveness of the proposed word length algorithm using datasets drawn from the Universal Declaration of Human Rights Act in 15 languages. The experimental results show good accuracy on language identification at the document level and at the sentence level based on the available dataset. The improved algorithm also showed significant improvement in run time performance compared with the spelling checker approach.

1. Introduction

Language identification (LID) refers to the process of determining the natural language in which a given text is written. Pienaar and Snyman (2010) observed that the language of a document can often not be determined on the basis of the file name alone. Moreover, documents on the Internet are not easily deciphered by computers with respect to language identification, because Web documents are traditionally created with the human reader in mind. Beesley (1988) noted that computers cannot use HTML code to determine the language of a web document even though XML and semantic mark-up with entries such as "xml: Lang attribute" and the <meta Lang = "fr"/> constructs have been introduced to tackle these challenges. Many documents still do not make use of metadata tags, or where such tags are used they may

* Corresponding author.
E-mail address: aselamat@gmail.com (A. Selamat).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.12.004

not be used correctly, thereby giving misleading information. According to Beesley (1988) as far as language identification is concerned the best effort is to try and deduce the information from the text itself, knowing that even when metadata are provided they may contain errors. Language identification is often the first step in many text processing systems. Whether it is a machine translation, semantic understanding, categorisation, storage, or information retrieval, text manipulation used online with mobile devices, or email interception, language identification would need to be done first. Therefore, there are serious implications and consequences for not embarking on research in language identification of under-resourced languages. We define under-resourced languages as those languages that do not have (or not enough) digital resources that can be employed for extensive research. The native speakers of such languages either do not use computers or if they do it is usually via a foreign language. This research is focused on languages with little or no digital resources, hence the name 'under-resourced languages'. These are mainly minority languages i.e., languages spoken by a few, but which are gaining importance due to an increasing and widespread use of the Internet and the possibility of such languages being used for communication over the Internet. So far, not much research has been done on identification of these languages probably because they were previously perceived as being less important than the popular languages. In this research we have taken advantage of the fact that the UDHR corpus is a multilingual corpus covering several languages (including some under-resourced languages) thereby making it possible to get a kind of kick-off resource base for this class of languages. Most resource-scarce languages cannot be identified automatically because no research has been done in this area, which means that criminals can use these languages for purposes of information hiding. There are several other consequences. For example, accessibility to Web documents is often hindered due to linguistic diversity on the Internet. Easy worldwide information exchange is one of the core advantages of the Web.

According to Kralisch and Mandl (2006), the language-related link following behaviour reveals important insight into the role of language when accessing information on the Web. Such insight into the role of language helps realise the goal of expanding language participation in Internet communication, thereby reducing the language "digital divide." To bring any language into the fold of natural language processing, some measure of research into its nature needs to be carried out. For many minority languages, however, such a study has yet to be done (Pienaar and Snyman, 2010). Such research would necessarily include or even begin with language identification of the languages in question. In addition, the study of any language on the digital stage needs a significant amount of digital resources. Where such resources are not available, research into these languages becomes difficult. Since language identification is often the first step in many natural language processing tasks (Newman, 1987), it is considered the place to begin. For example, it is only after language identification has been done that an appropriate translator can be selected for a meaningful translation wherever this is required.

Initially the digital divide was perceived as an issue of inadequate access to Information and Communication Technology (ICT) facilities. However, as the accessibility problem was being tackled it was soon realised that language would pose an even bigger problem with respect to information sharing among the peoples and strata of society. Erard (2003) emphasised the need for encoding of languages that are to be used on the Internet, noting that very few languages have so far been encoded which means that all the other languages are left out of the digital information bracket. On the other hand, Martindale, 2002 points out the special difficulties of digital communication in South Africa, a country with 11 official languages which necessitates the creation of websites in each separate language. The author concludes that the problem needs to be addressed by creating automatic translation programmes (Al-Salman, 2008; Bajwa et al., 2012) to facilitate information exchange. We have already noted that for any meaningful translation to happen, language identification must be performed first. It is clear that the relevance and gravity of effect of the various aspects of the language digital divide vary from country to country and from society to society. The implication of inability to identify any language automatically is that such languages become 'invisible' in any multilingual environment like the Internet. Even if documents in these languages are available, other participants do not know what to do with them. The language digital divide really means a division between those languages that are recognisable and those that are not recognisable by computers. By recognisable we mean ability to identify it automatically so that documents written in the language can be treated appropriately as far as natural language processing is concerned.

Language identification of resource-scarce languages using the spelling checker technique was proposed by Pienaar and Snyman (2010). Their experiments demonstrated substantial benefits in the identification of the South African languages using second-generation spelling checkers. In this research we propose an algorithm that improves the algorithm used by Pienaar and Snyman (2010). The proposed method involves pre-processing of input documents, tokenization, and generation of wordlist models using word-length aggregation, aimed at improving computational time gains and efficiency. The proposed models are targeted at solving the current problems of computational complexity, and time-consuming and multilingual identification. The techniques proposed hold the potential of applicability to any other languages as long as they are written in orthographical forms that permit tokenization. Using the lexicon-based approach for language identification as proposed in this research could pave way for further research and generate more digital resources for under-resourced languages. For example, the resulting word list models derived from training data in standard corpora can be further developed into pronouncing dictionaries (Carnegie Mellon University, 2008), thereby enabling applications and research in speech technology. In this research we undertake to find out how this technique will perform with respect to other languages, including languages of the same family. The languages featured in the study include four Nigerian languages (Hausa, Igbo, Tiv, and Yoruba), two South African languages (Ndebele and Zulu), Swahili in East Africa, two Ghanaian Languages (Akuapem and Asante), two South East Asian languages (Bahasa Melayu and Bahasa Indonesia), Croatian, Serbian, and Slovakian. This selection was deliberate in including two Asian languages which are strictly not under-resourced but are closely related languages. The same can be said of Serbian and Croatian which were only included in order to test the performance of our system on closely related languages. The English language is possibly the