ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Automatic Language Identification from Written Texts – An Overview

H L Shashirekha[1]

Department of Computer Science, Mangalore University, Mangalore, India[1]

ABSTRACT: Language Identification is the task of automatically identifying the language(s) in which the content is written in a document (web page, text document). Due to the widespread use of internet, identification of languages has become an important preprocessing step for a number of applications such as machine translation, Part-of-Speech tagging, linguistic corpus creation, supporting low-density languages, accessibility of social media or user-generated content, search engines and information extraction in addition to processing multilingual documents. In a multilingual country like India, Language Identification has wider scope to bridge the digital divide between different language users. This paper presents a brief overview of the challenges involved in automatic language identification, existing methodologies and some of the tools available for language identification.

KEYWORDS: n-gram, Text Classification, Natural Language Processing, Language Identification

I. INTRODUCTION

The task of identifying languages dates back to the ages when people started using language for communication. Human beings are very good at identifying spoken languages but not all are good at identifying languages in written form and also one cannot expect every human being to know all the languages. Internet and the related technologies have made abundance of text data available on-line in various languages and at the same have reached common man. However, in order to use the content of these textual data, one should know the language in which it is written or it has to be translated to the local language or mother tongue of an individual, which needs a language translator. Added to this is the collection of multilingual documents available in digital form which is quite natural in a multilingual country like India. More importantly, as different languages have different grammatical structures, the language processing tools are language dependent. Hence, there is need for automated tools and techniques which can identify the language of the written text and then select the required tools for further processing of the text based on the language of the written text. The solution to this problem is the Automatic Language Identification (LI) - the task of automatically identifying the language(s) in which the content is written in a document (web page, text document). A number of applications such as Machine translation, Part of Speech tagging, linguistic corpus creation, supporting low-density languages, accessibility of social media/user-generated content, search engines and information extraction require LI as a preprocessing step in addition to processing multilingual documents. Since many of the textual data analysis tools are language dependent, LI becomes a fundamental but crucial step for such applications. Automatic LI is not only useful but also very challenging due to various languages and the availability of text data in these various languages.

OVERVIEW

This paper presents a brief overview of the challenges that need to be addressed, existing methodologies and some of the tools available for automatic language identification.

1.1 Challenges involved in Automatic Language Identification

LI is a well-established research topic with state-of-the-art algorithms achieving over 95% accuracy. Text collection (web page or text document) which goes as input to Language Identification may be written in only one language (monolingual) or multiple languages (multilingual). Processing monolingual documents is fairly simple compared to that of processing multilingual documents as the former requires the knowledge of only one language while the later requires the knowledge of several languages [15] and also other related issues. Some of the challenges of Automatic Language Identification task are discussed below:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

- Length of the text data - Documents have to be of sufficient length in order to identify the language correctly for the simple reason that more vocabulary will be included in a lengthy document. However, documents containing messages from social media such as twitter will often be short making the language identification difficult [5, 8, 14].
- Noisy text - Text data may contain different noises like abbreviation, short forms of words and tags. In some cases, the whole text will be abbreviated or written in SMS style or written in a coded language which can be understood only by its creator [8, 14]. Developing a LI system robust to all types of noise is really challenging.
- Character encoding - The text processing requires judicious choice of character encoding and its usage in all the text data. However, character encoding of the text data may be different in different texts [6]. Developing a generic system to handle all types of character encoding needs to be addressed.
- Segmentation of documents - A problem closely associated to processing multilingual documents is the segmentation of documents which identifies the regions of a document in different languages [15]. Once the region is identified, the language of the content in that region can be identified and used for further processing.
- Common words – In a multilingual country, vocabulary of a language get influenced by various other languages and in due course of time those words become part and parcel of the language. Further, in case of similar languages, certain words are used commonly in all languages makes the language identification task difficult.
- Open class languages –The existing tools/algorithms for language identification are limited to a closed class of languages which are used in the training set. Determining the language(s) of a document outside the closed group is really challenging.
- Closely related languages - Similar languages or dialects of languages form the closely related languages and share a great deal of lexical and grammatical features making the language identification task difficult.

1.2 Popular algorithms for LI

LI may be considered as a special case of multi label text classification with a predefined set of labels representing the languages of the documents in the training set and LI as the task of assigning a subset of the predefined labels (languages) to a text document under consideration. This scenario holds good for the classification of multilingual documents. However, for LI of monolingual documents, multiclass text classifiers with 'k' classes can be considered depending upon the number of languages to be identified. Many machine learning and statistical approaches in combination with linguistic approaches (for feature extraction) are explored by many researchers to address the issues related to the identification of their native languages as well as languages in their neighboring regions. A brief survey of some of the popular algorithms is given below:

A small, fast and robust N-gram based method proposed by W. B. Cavnar and J. M. Trenkle [1] for text classification has been applied by many researchers successfully for LI [3, 8, 9, 10, 11, 13]. N-gram is an N-character slice of a longer string and N-grams of different lengths will be used. The general frame work of an N-gram based method is to compute the language profile for each language in the training set and the target profile for each test document under consideration. The system then computes a distance measure between the target profile and each of the language profiles in the training set and finally selects the language whose profile has the smallest distance to the target profile. Several researchers have explored the variations of the N-gram based method for LI. Bashir Ahmed et. al [3] have used an ad-hoc cumulative frequency addition of N-grams (2-grams to 7-grams) for the identification of short texts of 12 languages. They claim that the speed of their method is comparable with Naïve Bayes method for classification and the accuracy is comparable with rank-order statistics method. Erik Tromp and MykolaPechenizkiy [8] propose a graph-based N-gram approach for the identification of languages in relatively short and ill-written texts. While most of the approaches give importance only for word occurrence, this approach gives importance for word ordering in addition to word occurrence which is represented as a graph. They experimented on the collection of Twitter messages written in six languages and found that their method is significantly more accurate than the existing N-gram based approaches and less prone to overfitting and domain-specific jargon.

The literature review reports some works on the identification of languages in web pages. An improved N-grams approach based on a combination of original N-grams (ONG) approach and a modified N-grams (MNG) approach for language identification of web documents is proposed Ali Selamat [9]. They select the features based on a distance measurement from the original N-grams approach and the features based on a Boolean matching rate from the modified

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

N-grams approach for the identification of 12 languages in Roman and Arabic scripts web pages. The results illustrate that the improved N-grams approach has been able to improve the language identification of the contents in Roman and Arabic scripts. An improved N-gram algorithm is proposed by Yew Choong Chew et al., [10] for LI of web pages for Asian languages based on non-Latin script. The performance of the algorithm was evaluated based on a written text corpus of 1,660 web pages spanning 182 languages from Asia, Africa, America, Europe and Oceania and achieved an accuracy rate of 94.04%.

Even though lot of work have been done in the identification of various languages, very less work in reported for automatic identification of Indian languages [5, 6, 12, 11]. Deepamala. N and Ramakanth Kumar. P [11] has proposed an N-gram algorithm for the identification of documents with Kannada, Telugu and English sentences by processing n-gram of only the last word of the sentence instead of complete sentence and they have used it as a preprocessing step for sentence boundary detection and found encourage results. Identification of languages having common script has its own challenges. Sreejith C et al [12] have proposed an N-gram based algorithm for distinguishing between Hindi and Sanskrit texts which are having a common script. They have used character based unigram, bigram and trigram based training profiles and has achieved 99% accuracy. KaviNarayana Murthy and G. Bharadwaja Kumar [5] formulate LI as a two class pairwise classification problem using Multiple Linear Regression (MLR) for the classification of small text samples of Indian languages. This paper also throws light on the scripting languages and issues related to identification of Indian languages. The challenges in automatic LI Romanized text is addressed by KosuruPavan et al., [6] in the form of a system called RoLI. RoLI is an N-gram based approach which also exploits sound based similarity of words and has achieved an average accuracy of 98.3% despite the spelling variations on five Indian languages: Hindi, Telugu, Tamil, Kannada and Malayalam.

Abdelmalek Amine et. al., [7] have proposed hybrid algorithm based on the combination of k-means and artificial ant class algorithm for the identification of languages in a multilingual text. They operate on the N-grams of characters as attributes, and cluster together similar texts and discover the number of languages in a completely unsupervised manner. Rafael DueireLins, Paulo Gonçalves Jr. [2], propose a recognition strategy based on linguistic features called closed word classes that include Adverbs, Articles, Conjunctions, Interjections, Numerals, Prepositions and Pronouns and experimented on four languages viz., Portuguese, Spanish, French and English. Bruno Martins and Mário J. Silva [4] discusses an N-gram based algorithm complemented with heuristics and a new similarity measure to identify the language of a given Web document. The algorithm was tested on 23 different languages, constructed from textual information extracted from newsgroups and the Web and is being used as part of Portuguese Web search engine (www.tumba.pt). Marcos Zampieri [13] has proposed simple bag-of-words approach for identification of language varieties and has performed experiments using Multinomial Naïve Bayes (MNB), Support Vector Machines (SVM) and J48 classifier. The results show that their method has performance comparable to state-of-the-art methods based on n-gram models. MarcoLui et al., [15] presents a system for language identification in multilingual documents using a generative mixture model inspired by supervised topic modeling algorithms, combined with a document representation for monolingual documents. The results illustrates that the proposed system outperforms alternative approaches from the literature on synthetic data, as well as on real-world data from related research on linguistic corpus creation for low-density languages using the web as a resource.

Identifying languages from noisy text is a real challenge in LI as most methods work on clean texts and/or long texts, but often present a failure when the text is corrupted or too short. KheireddineAbainia et al., [14] have proposed a hybrid approach for the identification of languages of noisy short texts and experimented on the collection of texts from several discussion forums containing several types of noises pertaining to 32 languages. Their hybrid approach which is defined as a combination of term-based and character-based methods are quite interesting and present good language identification performances in noisy texts.

II. TOOLS FOR LI

Research in LI has resulted in the availability of a number of tools for the identification of language(s) automatically [16]. Table 1 lists a few tools for LI available commercially as well as freely.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Table 1. List of tools for LI

System	Number of languages	Availability
Textcat Several versions of textcat are also available	69	Free
SILC/Alis	28	Commercial
Xerox MLTT Language Identifier	47	Commercial
Collexion	15	Commercial
Stochastic Language Identifier	13	Free
Rosette Language Identifier by Basis Technology	30	Commercial
Language Identification program by Ted Dunning	2	Free
Lextek Language Identifier	Many	Commercial/free
Langwitch by Morphologic	7	Commercial
Languid	72	GPL
Lid	23	Commercial
C# package for language identification of Microsoft	52	Free

III. CONCLUSION

This paper provides a brief overview of the challenges involved in automatic LI, existing methodologies and some of the tools available for automatic LI. It can be observed that most of the methods use N-gram model or variation of N-gram model in combination with other techniques for feature extraction and then use machine learning techniques for the identification of languages. Even though lot of work has been done on LI, identification of Indian languages is not addressed much. LI has wider scope to bridge the digital divide between different language users especially in a multilingual country like India.

REFERENCES

1. W. B. Cavnar and J. M. Trenkle, (1994), "N-Gram-Based Text Categorization", In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13 April 1994.
2. Rafael DueireLins, Paulo Gonçalves Jr., (2004), "Automatic Language Identification of Written Texts", ACM Symposium on Applied Computing, March 14-17, 2004, Nicosia, Cyprus.
3. Bashir Ahmed, Sung-Hyuk Cha, and Charles Tappert, (2004), "Language Identification from Text Using N-gram Based Cumulative Frequency Addition", Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 7th, 2004.
4. Bruno Martins and Mário J. Silva, (2005), "Language Identification in Web Pages", SAC'05 March 13-17, 2005, Santa Fe, New Mexico, USA.
5. KaviNarayana Murthy and G. Bharadwaja Kumar, (2006), "Language Identification from Small Text Samples", Journal of Quantitative Linguistics, 2006, Volume 13, Number 1, pp. 57 – 80.
6. KosuruPavan, NiketTandon, VasudevaVarma, (2010), "Addressing Challenges in Automatic Language Identification of Romanized Text", Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, Macmillan Publishers, India.
7. Abdelmalek Amine, ZakariaElberrichi, Michel Simonet, (2010), "Automatic Language Identification: An Alternative Unsupervised Approach Using a New Hybrid Algorithm", IJCSA 7(1), 2010, pp. 94-107.
8. Tromp, E. &Pechenizkiy, M. (2011), "Graph-Based N-gram Language Identification on Short Texts", In Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning (Benelearn 2011), pp. 27-34.
9. Selamat, Ali, (2011), "Improved N-grams Approach for Web Page Language Identification", Transactions on Computational Collective Intelligence V, Lecture Notes in Computer Science, Volume 6910, 2011, pp. 1-26.
10. Yew C. Chew, YoshikiMikami, Robin L. Nagano, (2011), "Language Identification of Web Pages Based on Improved N-gram Algorithm", International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011, pp. 47-58.
11. Deepamala. N, Ramakanth Kumar. P, (2012), "Language Identification of Kannada Language using N-Gram", International Journal of Computer Applications, 6(4),"pp. 24-28, May 2012.
12. Sreejith C, Indu M, Dr. Reghu Raj P C, (2013), "N-gram based Algorithm for distinguishing between Hindi and Sanskrit texts", Proceedings of the Fourth IEEE International Conference on Computing, Communication and Networking Technologies, July 4 - 6, 2013.
13. Zampieri, M., (2013), "Using bag-of-words to distinguish similar languages: How efficient are they?" IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI), 2013, pp. 37-41, 19-21 Nov. 2013.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

14. KheireddineAbainia, SihamOuamour, HalimSayoud, (2014), "Robust Language Identification of Noisy Texts - Proposal of Hybrid Approaches", 11th International Workshop on Text-based Information Retrieval (TIR) 2014, Munich, Germany, September 2014.
15. Marco Lui, Jey Han Lau and Timothy Baldwin, (2014), "Automatic Detection and Language Identification of Multilingual Documents", Transactions of the Association for Computational Linguistics, pp. 27–40.
16. http://odur.let.rug.nl/vannoord/textcat/competitors.html