Indonesian Journal of Electrical Engineering and Computer Science
Vol. 18, No. 1, April 2020, pp. 544~548
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v18.i1.pp544-548

Identification of language in a cross linguistic environment

Merin Thomas[1], Latha C. A[2], Antony Puthussery[3]
[1]Research Scholar, Regional Research Center, Visvesvaraya Technological University, India
[2]Head of the Department (CSE), AMC Engineering College, Visvesvaraya Technological University, India
[3]Assistant Professor, Department of Science and Humanities, CHRIST (Deemed to be University), India

Article Info

Article history:

Received Jul 30, 2019
Revised Sept 4, 2019
Accepted Oct 12, 2019

Keywords:

Cross linguistic
Multilinguistic
Sentimental analysis

ABSTRACT

World has become very small due to software internationationalism. Applications of machine translations are increasing day by day. Using multiple languages in the social media text is a developing trend. Availability of fonts in the native language enhanced the usage of native text in internet communications. Usage of transliterations of language has become quite common. In Indian scenario current generations are familiar to talk in native language but not to read and write in the native language, hence they started using English representation of native language in textual messages. This paper describes the identification of the transliterated text in cross lingual environment. In this paper a Neural network model identifies the prominent language in the text and hence the same can be used to identify the meaning of the text in the concerned language. The model is based upon Recurrent Neural Networks that found to be the most efficient in machine translations. Language identification can serve as a base for many applications in multi linguistic environment. Currently the South Indian Languages Malayalam, Tamil are identified from given text. An algorithmic approach of Stop words-based model is depicted in this paper. Model can be also enhanced to address all the Indian Languages that are in use.

Corresponding Author:

Merin Thomas,
Research Scholar, Regional Research Center,
Visvesvaraya Technological University, India.
Email: merin.jisso@gmail.com

1. INTRODUCTION

Natural language processing has been an interesting area of research in machine learning. Artificial intelligence provided to the machines enables them to cope up with the native languages used by the humans. Complexity of the native languages is one of the most challenging problems to deal with the Natural Language processing. To design intelligent machines machine learning technique neural network can be used [1]. Unlike computer language keywords, meaning of the keyword changes with sentences in native languages where ambiguity is at the peak. Semantic analysis can be done with the help of corpus associated with the language.

India is a multilinguistic Country where in each state speaks different language. Language boundary and cultural differences make its beauty in diversity. With 22 major languages, written in 13 different scripts, with over 720 dialects, India stands to be one of the largest multilinguistic countries in Asia. Malayalam, Tamil and Telugu are the prominent languages in South India. Malayalam is native language of the state Kerala spoken by 38 million people, Kannada, the native language of Karnataka and Tamil, native language of Tamil Nadu and also official language of two other countries Singapore and Sri Lanka. Tamil is spoken by a total 70 million people. Apart from these languages, English has become the common language spoken in India.

In the earlier stages of computers only English language were widely used in the documents, emails and messages. To make computer adaptable to all sectors of people, even somebody who does not know English, only way out was to make computer enabled with native languages. Introduction of fonts in native

languages enabled its usage widely. Usage of Computers and Mobile Apps became widespread. Machines Translations played a vital role in converting any language to any other language. Translations are done with the help of dictionaries or wordnet of language, exact word by word translations may end in changing the entire meaning of the context.

ഉപപുമാവ് is the Malayalam word that represents a kind of breakfast in South India, many of the translators available to provide translation to English interpret the word as subheading or upright flour. The major Challenge dealing with the native language is the degree of ambiguity that add on to every context making the accuracy to fall below considering word by word translations. Different architectures for performing translation task is Rule Based Machine Translation and Statistical Machine translation [2].

Current generation saw the need of transliteration than translation. People were accustomed to English Writing and Reading than their native language, where in English words took most its place. Prominence of English words in our day to day life is so high that it became convenient in substituting the words in native language. But people who were well versed in speaking native language but not that well versed in writing or reading started using native language typed in English which is called as Transliterations for ease of communication. Fact that humans are more comfortable in their Natural Language when it comes to expression of words has its application in this context. There are basically three approaches for transliteration. They are based on grapheme, Phoneme and Hybrid Approaches. In grapheme approach, it directly transforms grapheme from source to target. In Phoneme model the key is pronunciation of source language. Hybrid model uses both the grapheme and phoneme model information.

Transliteration can be dated back to 1994 where major work was in the area of Arabic-English [3]. A generative model for back transliteration from English to Japanese was proposed in 1997[4].Mathematical approximation technique using statistical model was used in English Korean Transliteration in the year 2000 [5]. An automatic character alignment method for English word and Korean transliteration is discussed in [6]. In year 2002, a hybrid model [7] was built on phonetic and spelling mappings using Finite state machines. Transliteration of Arabic names in to English was done by this method. In 2004, a new framework allowing direct orthographic mapping (DOM) between two different languages, through a joint source-channel model, also called n-gram transliteration model (TM) was introduced [8]. It generates probabilistic orthographic transformation rules using a data driven approach. Phonemic interpretation, level is skipped, so that the error rate in transliteration is reduced significantly.

Sample Transliteration:

ഉപപുമാവ് will be typed in English as Uppumavu.

Cross Linguistic is the usage of multiple languages in the same text. This effect is due to the influence of other languages especially English in their native language. Cross linguistic and Transliterations are the two issues that have to be addressed in the analysis of Social media text. When it comes to data analysis than language boundaries meaning of the data matters. In application like analysing the review of the products, on mining the web, we may have to analyze reviews in different languages, transliterations about the same product etc. So it is important to identify to which language the text belongs to in order to understand meaning in the text. Identification of the language in the Indian scenario is one of the toughest jobs when concerned with number of existing languages. Usage of transliteration in the social media text had made the problem even worse.

2. ALGORITHM FOR LANGUAGE IDENTIFICATION IN CROSS LINGUAL AND TRANSLITERATION TEXT

In this paper we describe algorithmic stop words based model for the identification of particular language in a text of conversation. Stop words are basically the most common words used inside a language. Procuring of the appropriate data set, is achallenging task. Social media text can be either transliterated or it can be mixture of multiple languages. The algorithm identifies the language with respect to three languages used inside the text, Malayalam, Tamil and English.Several machine learning algorithms are used for the categorization of languages in a multilinguistic approach.Category of the classification algorithm ranges from the simple naive bayesian approach to complex deep learning algorithms.Hybrid methodology is also followed to bring out best features among supervised algorithm and unsupervised algorithm.Stop word based model is simple method that divides the text in to language bags based on the stop words.

Algorithm for Stop word-based Language Detection Model

1.Remove the content E from I, Where I is the input text and
E={;,:,"<",<,>,.,?,/,{,},{,[,],|,\,!,@,#,$,%,^,&,*,(,),_,-,,=,+,emotional icons}

2.Divide the sentence S in to set W, where W is set of unique words in S

3.Invert the case of W to form set w, where w ∈ lowercase(W)

4. For each w_i element of w, s_i element of S, where S is the set of transliterated stopwords of all languages
for i= 1 to n
strcmp (w_i, S_i) = k_i, k_i is the match for each language L_i.

4. Find $K=\sum k_i$
5. If K_i = count (Z_i), Z_i is the count of matched stopwords of language L_i
6. Find M=max (K_i)

7.Language L is identified as the one with largest M value.

3. DATA SET

Transliterated text of stop words of languages Tamil, Malayalam, English were collected. Sample of more than 1000 stop words in each languages were used. Transliteration of stop words are used to train the model. Cross linguistic input text was collected from Twitter, Facebook and Whatsapp.

Table 1. Sample stop word samples of English, Tamil and Malayalam

i	oru	aksharam
me	endru	oru
my	mattrum	paranju
myself	indha	enna
we	idhu	roopa
our	Naalai	sarkar
ours	endra	sammanam
ourselves	kondu	bharya
you	enbadhu	adheham
you're	pala	thanne
you've	aagum	samsthana
you'll	alladhu	keralam
you'd	avar	makkal
your	naan	ninnu
yours	ulla	nair
yourself	andha	vare
yourselves	ivar	cheythu
he	ena	muthal
him	mudhal	dey
his	enna	puthiya
himself	irundhu	uthgadanam
she	sila	manthri
she's	en	ennal
her	pondra	aa
hers	vendum	mathram
herself	vandhu	innu
it	idhan	kottayam
it's	adhu	ninnum
its	avan	kuduthal
itself	thaan	ippol
they	palarum	eppol
them	ennum	niryathanayi

4. EXPERIMENTAL RESULTS

Program was executed for basically two kinds of input. One input with pure Malayalam and Tamil text written in English or can be called as transliterated text of English and Tamil. Other one with combination of two languages. Output was compared with actual results to record the performance index.

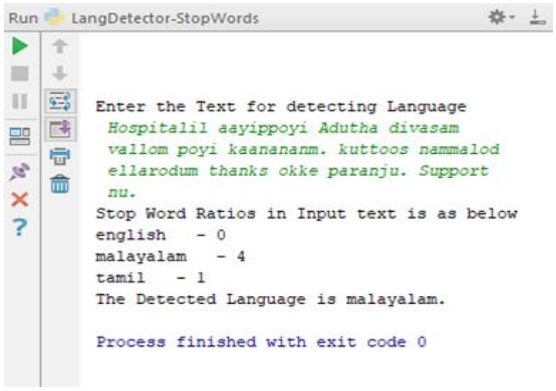

Figure 1. Transliterated input text of pure Malayalam

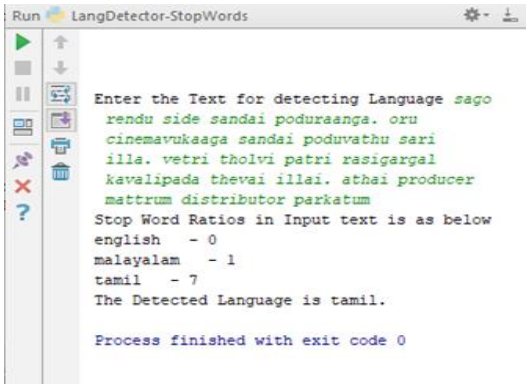

Figure 2. Transliterated input text of pure Tamil

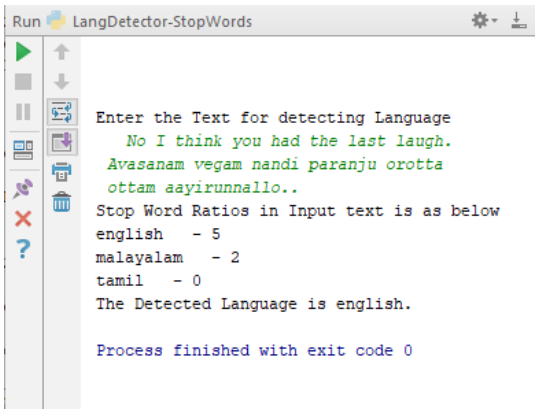

Figure 3. Transliterated input text with English with Tamil

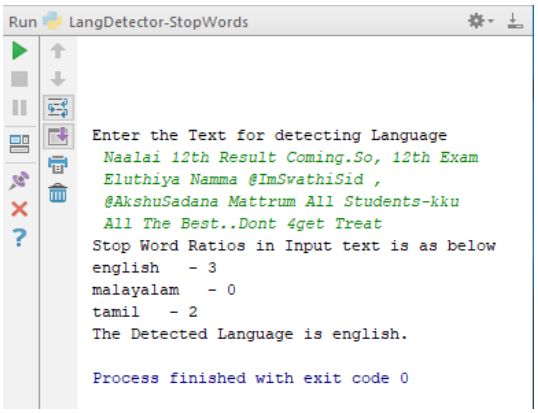

Figure 4. Transliterated input text with English mixed with Malayalam

5. CONCLUSION

Performance of the algorithm is satisfactory since 80% of accurate results predicted proved correct. Performance of the algorithm depends on extensive list of stop words. In Combinational sentences correctness proved to be the least since language detection is purely based on whether word is present in stop word list or not. With the list of appropraiate stops words this work can be extended to other native languages in India.

REFERENCES

[1] Mishra, Chandrahas, and D. L. Gupta. "Deep Machine Learning and Neural Networks: An Overview."*International Journal of Artificial Intelligence (IJ-AI)*, 6.2 (2017): 66.
[2] Alqudsi, Arwa, Nazlia Omar, and Rabha W. Ibrahim. "Rule Based and Expectation Maximization algorithm for Arabic-English Hybrid Machine Translation." *International Journal of Artificial Intelligence (IJ-AI)* 5.2 (2016).
[3] Arbabi, M., Fischthal, S. M., Cheng, V. C., And Bart, E. "Algorithms for Arabic Name Transliteration". *IBM Journal of Research and Development*, 38, 2, 183, 1994.
[4] Knight, Kevin and Graehl, Jonathan. "*Machine Transliteration*". In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. 1997, pp. 128-135.
[5] Jung, S. Y., Hong, S., & Paek, E. "*English to Korean transliteration model of extended markov window*". In Proceedings of the 18th conference on Computational linguistics, 2000, pp. 383-389.
[6] Kang, B. J., & Choi, K. S. "*Automatic transliteration and back-transliteration by decision tree learning*". In Proceedings of the 2nd International Conference on Language Resources and Evaluation,2000, pp. 1135–1411.
[7] Y. Al-Onaizan and K. Knight,"*Machine Transliteration of Names in Arabic Text*", Proc. of ACL Workshop on Computational Approaches to Semitic Languages, 2002.
[8] Jong-Hoon Oh Key-Sun Choi" Machine Learning Based English-to-Korean Transliteration using Grapheme and Phoneme information" *Ieice Trans.Inf.& Syst.*, VOL.E88-D, NO.7, julyb2005, pp 1737-1748.
[9] Ali, Aasim, Shahid Siddiq, and Muhammad Kamran Malik. "Development of parallel corpus and english to urdu statistical machine translation." *Int. J. of Engineering & Technology IJET-IJENS* 10 (2010): 31-33.

[10] DEEP, Kamal; KUMAR, Ajit; GOYAL, Vishal. "Development of Punjabi-English (PunEng) Parallel Corpus for Machine Translation System". *International Journal of Engineering & Technology*, [S.l.], v. 7, n. 2, pp. 690-693, may 2018. ISSN 2227-524X.
[11] Du, Jiali, Pingfang Yu, and Minglin Li. "Machine Learning from Garden Path Sentences: The Application of Computational Linguistics." *International Journal of Emerging Technologies in Learning (iJET)*, 9.6 (2014): 58-62.
[12] Prabhu Palanisamy, Vineet Yadav and Harsha Elchuri, (2013). "*Serendio Simple and Practical Lexicon Based approach to Sentiment analysis*," Volume 2: Seventh International Workshop on Semantic Evaluation, Atlanta, Georgia, pages 543-548.
[13] I.Hemalatha, Dr.G. P Saradhi Varma, Dr.A. Govardhan (2013), "Sentiment analysis tool using machine learning algorithm", volume 2, Issue 2, *International journal of emerging Trends and Technology in Computer*.
[14] Peter D. Turney, (2002), "*Thumbs up and thumbs down? Semantic Orientation Applied to Unsupervised Classification*", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, 417-424.
[15] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. "*Thumbs up?: sentiment classification using machine learning techniques*". In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79-86.
[16] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques", *International Journal of Engineering and Technology*, 7.6 (2016): 1-7.
[17] Sida Wang and Christopher D Manning. 2012. "*Baselines and bigrams: Simple, good sentiment and topic classification*". In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pages 90-94.
[18] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and ´ Christian Janvin. 2003. "A neural probabilistic language model". *The Journal of Machine Learning Research*, 3: 1137-1155.
[19] Ronan Collobert, Jason Weston, Leon Bottou, Michael ´ Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. "Natural language processing (almost) from scratch". *The Journal of Machine Learning Research*, 12: 2493-2537.
[20] Satyanarayana P, Charishma Devi, Sowjanya P, Satish Babu, Syam Kumar, "Implementation of conventional communication system in deep learning". *International Journal of Engineering & Technology*, v.7 (1.1) pp. 696-698, 2018. ISSN 2227-524X.

BIOGRAPHIES OF AUTHORS

Mrs Merin Thomas, currently working as Assistant Professor in CHRIST (Deemed to be University) in the department of Computer Science and Engineering. She has completed her masters from Visveswarya Technical University(VTU).She is pursuing her Research under Visveswaraya Technological University.

Dr Latha C A, is a doctorate from Anna University, Chennai. She has done her post-graduation from NITK Suratkal and Graduation from Mysore University in 1991. Since then, in her vast academic experience, she is contributing to technical education in most of the capacities. Dr Latha has filed for an US patent for one of her research works. She has authored a book on 'Programming in C' which is widely appreciated and used by the students. She was BoE for VTU in 2013 and currently for Dayanand Sagar University. Being a Reviewer and Technical Program Committee member for many of the IEEE International Conferences and reputed Journals, she is also been awarded, "Outstanding Reviewer award" by reputed Elsevier publishers

Antony puthussery is a currently working as Assistant Professor in CHRIST (Deemed to be University) in the department of Science and Humanities.His area of expertise includes mathematical modeling and Theorectical Graph Theory. He has several journal and conference publications to his credit especially in the area of mathematical modeling using graph.