Language Identification for Multilingual Machine Translation

Arun Babhulgaonkar and Shefali Sonavane

***Abstract*—Machine translation is the process of translating a text in one natural language into another natural language using computer system. Translating a document containing a single source language contents is easy but when the information in the source document is given in multilingual format then there is a need to identify the languages that are involved in such multilingual document. Language identification is the task in natural language processing that automatically identifies the natural language in which the content in given document are written in. Language identification is the fundamental and crucial step in many NLP applications. In this paper, n-gram based and machine learning based language identifiers are trained and used to identify three Indian languages such as Hindi, Marathi and Sanskrit present in a document given for machine translation. It is observed that, support vector machine based language identifier is more accurate than any other technique and it achieves 89% accuracy that is 18% more than traditional n-gram based approach. The inclusion of language identification component in machine translation improved the quality of translation.**

***Index Terms*—Language Identification (LI), Language Modeling (LM), Machine Learning, Machine Translation (MT).**

I. Introduction

EVERY human being may not be expected to know all the natural languages. Abundance of useful information is available on the web but may be in foreign language. For making it available in native language of user, first the source language of the information must be identified. Language identification (LI), also called as language guessing, is the task in natural language processing (NLP) that automatically identify the natural language in which the content in given document are written in. Before going for any particular natural language application one must identify the language of the content. Natural languages have different grammatical structures hence many task of NLP such as POS tagging, information extraction, machine translation, multilingual documents processing are language dependent. Language identification is fundamental and crucial stage in many NLP applications. Hence, there is need to develop an automated tool and techniques for language identification before application of further processing. For example, in case of machine translation to convert a foreign language text into required language text, the language in which the original text is written must be identified. Once it is identified then using a machine translation system it can be translated in required target language. Due to diversity of documents on the web, LI is a vital task for web search engines during crawling and indexing of web documents. For cross-lingual applications there is an increasing demand to deal with multilingual documents. Computationally, language identification problem is viewed as a special case of text categorization or classification.

This paper enlists the challenges in automatic language identification that is required as a pre-processing task for MT. Many methods of LI are based on n-gram modeling. State-of-the-art techniques use machine learning based algorithms for extracting linguistic features along with traditional n-gram modeling for language identification. Language identification is helpful to remove the language barrier among the users along with machine translation. Although a lot of research is going on for language identification for foreign languages, a little attention is paid for automatic identification of Indian languages.

Rest part of the paper is organized as follows: In section II, role of language identifier in machine translation is presented. In section III, challenges involved in language identification are discussed. In section IV, methodology used for experimentation is discussed. In section V, research work related to language identification is given. In section VI, experimentation and result analysis done is discussed and finally in section VII concluding remark is given.

II. Language Identification for Machine Translation

With the rapid growth of internet and success of search engines, everybody try to find ones required information on the web. But sometimes the useful information is available in foreign language that may not be known to the user. Here, automatic machine translation application of natural language processing is used to translate this information into user's native language. Difficulty arises when the information in the source document is given in multilingual format. In this scenario, single machine translation model trained on a single source to target language pair is not sufficient. For a MT system trained for a particular source language, the contents in the document that are in other language work as noise and due

Arun Babhulgaonkar is a research scholar in Department of Computer Science & Engineering, Walchand College of Engineering, Sangli, Maharashtra, India.
(e-mail: arbabhulgaonkar@dbatu.ac.in).

Shefali Sonavane is working as Associate Professor in Department of Info. Tech., Walchand College of Engineering, Sangli, Maharashtra, India.
(e-mail: shefali.sonavane@walchandsangli.ac.in).

978-1-7281-4988-2/20/$31.00 ©2020 IEEE

to this the overall translation quality also gets degraded. In this scenario, the language identification is used first to separate the contents in the source document into segments of languages involved. Then, each segment is translated using a separate machine translation engine trained for that particular source language. Fig. 1 illustrates this.

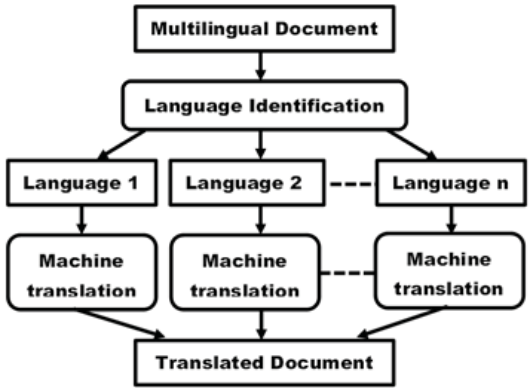

Fig. 1. Role of language identification in machine translation.

Machine translation block in Fig. 1 is further expanded in Fig. 2. Language model, reordering model and translation model are main working components of a translation system. Translation model is trained on bilingual parallel corpus of source and target language. The language model is trained on monolingual corpus of target language of translation. Reordering model takes care of reordering of words according to grammatical structure of target language.

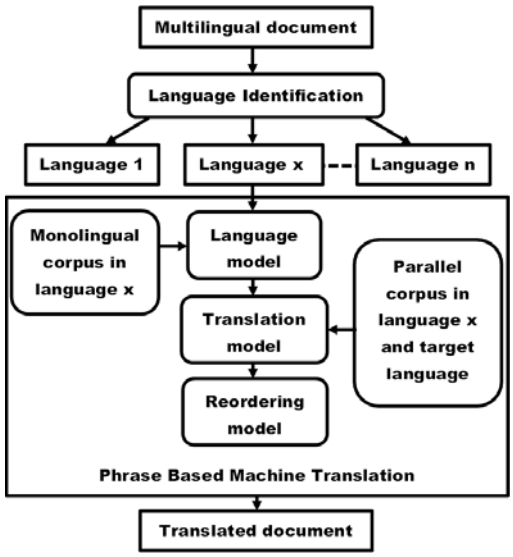

Fig. 2. Pharse based machine translation.

After obtaining translation of language in each segment to the target language, the final translated document is constructed by combining together all the segments. This way by obtaining translations of all the languages present in the multilingual source document, the translation quality is also increased.

III. Challenges Involved in Language Identification

Text document given for language identification task may be monolingual or multilingual. Processing monolingual documents are fairly easy as it requires the knowledge of only one language but to process multilingual document one need to have knowledge of many languages and also their interdependency issues. Major challenges of LI are enlisted below:

1) Length of the text data: For successfully identifying the language, text data in the document must be sufficiently large enough to apply n-gram like techniques. However, messages in social media application like whatsapp, twitter are very small to work with.

2) Noisy text: Use of short forms of words and tags known as abbreviations are considered as noise in the text. Due to unavailability of space and typing inconvenience user write whole text in abbreviated or coded form that creator only can understand. To work with such a noisy data is really challenging task.

3) Character encoding: Many natural languages use different character sets. Character encoding, also known as character set, represents individual characters using an encoding system made up of other symbols. Many encoding systems such as ASCII, Unicode are available. Languages in a single text document may contain various encodings. To deal with such encoding variations is a tricky task and need a generic system.

4) Segmentation of documents: A multilingual document contains intermixed text written in multiple languages alternately. The big problem in such document is the segmentation of text in various languages involved, for separating out the contents of the document. Once the segments are identified, the content can be collected for further processing.

5) Common words: Due to cultural contact between two language community words get adopted by one language from another language. Such words are called as a loanword or borrowing. In a multilingual country like India, adoption of words from other languages is very common. This inheritance of words makes some words available in many languages that make the LI a difficult task.

6) Open class languages: Most of the automatic language identification tools and algorithms available today are applicable to a closed class of languages. For languages outside this class LI is really challenging.

7) Languages with same origin: Many closely related languages share scripts and grammatical features as they are derived from the same origin. For example, Hindi and Marathi language both uses Devnagri script. Discrimination of language of the word in such scenario is the big challenge in the language identification task.

IV. Methodology

One n-gram based and four machine learning based language identifiers are constructed for Hindi, Marathi and Sanskrit language. Final decision about the language is taken

by combining the decisions of all the classifiers using majority voting of ensemble learning.

A. n-gram based language Identification

The nature of a natural language how the words appear one after other in a sentence can be captured with language modeling. The language model finds the probability distribution over a sequence of words that shows the tendency of frequently following words. n-gram is the leading method of language modeling. It shows how many words/characters are considered for predicting the next word/character in a sentence. n may be 2 words for 2-gram, 3 words for 3-gram model and so on. Language model can be constructed over sequence of characters or words. For example, the words in sentence "my name is Arun" can be modeled using word and character n-grams as:

Character based modeling
Using Unigram model : (m), (y), (n), (a), (m)...
Using Bigram model : (my), (yn), (na), (am)..
Using Trigram model : (myn), (yna), (nam), (ame)..

Word based modeling
Using Unigram model : (my), (name), (is), (Arun)
Using Bigram model : (my name), (name is), (is Arun)
Using Trigram model : (my name is), (name is Arun)

In word based n-gram language modelling, probability of next word w_n is calculated using previous n-1 words $w_1 w_2 \ldots . w_{n-1}$ in the sentence as:

$$p(w_n | h) = p(w_n \mid w_1 w_2 w_3 \ldots . w_{n-2} w_{n-1}) \quad (1)$$

Where, *h* represents contextual history of words in the sentence. The probability distribution of n-grams is obtained by using maximum likelihood estimation as:

$$p(w_n | w_1, \ldots . , w_{n-2}, w_{n-1}) = \frac{C(w_1, \ldots . w_{n-2}, w_{n-1}\, w_n)}{C(w_1, \ldots . , w_{n-2}, w_{n-1})} \quad (2)$$

Thus, n-gram modelling finds probability of subset of n character or words sequence in a long sentence in the text. Here, n is the number of characters or words used and it may vary from 1 to any number of words. For language identification purpose, sequences of characters or words in the sentence of all languages are modelled separately using n-grams. The n-gram technique captures the profile of a natural language during the training. Thus, profiles of all the languages present in training data set are stored in trained n-gram models separately. The profile of the language in test document is then obtained using n-gram technique. The distance between the profile of test document language and profiles of languages in training documents is measured using similarity metric and finally the language in training dataset whose profile is having minimum distance is selected as the language of the test document.

B. Classification based language identification

A multilingual document contains intermixed text written in multiple languages alternately. Language identification is used for the segmentation of text in various languages involved in document for translation. Language identification is considered here as a classification task. Each language is considered as a class label. A text document is processed and a predefined language class is assigned to it. Features of the classifier are extracted from the text and the classifier is trained to identify the languages. The segments in the multilingual document are given to the classifier and according to features involved in it the language identifier labels it with respective language. Fig. 3 illustrates the classification approach of language identification. The characters, words, keywords, grammatical structure of language differentiate one natural language from other natural language. The language specific words and keywords are extracted as features of the natural language. Many words are common in languages having same origin. As Marathi and Hindi are originated from Sanskrit language, many overlapping words are there in these languages. Each and every word of language may not be used as discriminator. Term frequency (TF) and Inverted Document Frequency (IDF) of words is calculated and the language keywords are obtained. Weights are assigned to the keywords according to their frequency distribution. Importance of the word in the language is obtained using maximum likelihood estimation. For language classification purpose four supervised techniques such as logistic regression, Naïve Bayes classifier, SVM and k-Nearest Neighbours are used during this experimentation.

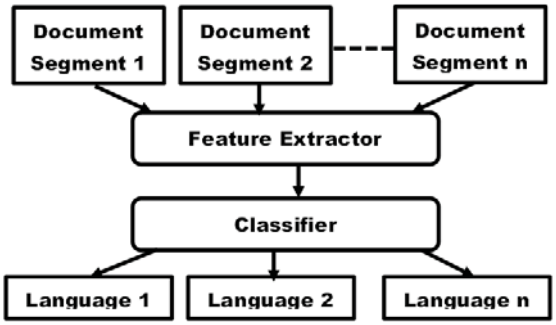

Fig. 3. Language identification using classification approach.

V. Related Work

W. B. Cavnar and J. M. Trenkle [1] introduced the use of n-gram for text categorization. This method became base of many approaches proposed by researchers for LI. Bashir Ahmed et. al [2] improved results for documents containing short strings using an ad-hoc cumulative frequency addition of n-grams for language identification. They improved the classification speed as compared to Naïve Bayes classifier and improved the accuracy as compared to rank-order statistics approach. Bruno Martins and Mário J. Silva [3] developed a n-gram based algorithm working on the inverse of similarity heuristic metric to recognize language of a web document. They implemented a crawler for a search engine after

successfully testing the algorithm on information extracted from web pages of 23 different languages [4,5]. KosuruPavan et al. [6] developed a system named as RoLI to handle the challenges of Romanized text in Indian languages. They used classic Soundex algorithm to deal with the phonetic and Levenshtein distance measure for spelling differences of Romanized text in Hindi, Telugu, Malayalam, Tamil and Kannada language. Abdelmalek Amine et. al. [7] proposed clustering based unsupervised classification of multilingual text. They combined artificial ant class algorithm and k-means along with n-gram modelling of text to form a hybrid algorithm for automatic language identification. Erik Tromp and Mykola Pechenizkiy [8] used a graph based approach called as LIGA to capture language grammar. They improved the results over prevalent n-gram based approaches for short and ill-written texts in social media like Twitter. They represented word occurrence in terms of a graph and considered the importance of word ordering also. They claimed that their LIGA approach is less prone to overfitting and jargon. Language identification of Latin script based languages is successfully implemented but still many new challenges are there for Indian languages. Deepamala N. and Ramakanth Kumar P. [9] has suggested new approach of using n-grams at the end of the sentence only instead of using n-grams over entire sentence for identification of Kannada, Telugu and English sentences in a document. Sreejith C. et al. [10] have used n-gram technique to distinguish Hindi and Sanskrit languages that share a Devnagri script. They developed character based n-gram models from unigram to trigram to obtain language profiles. Marcos Zampieri [11] represented the words in the text using bag-of-words approach and used three classification algorithms like Multinomial Naïve Bayes, SVM and J48 classifier. KheireddineAbainia et al. [12] experimented on noisy data of discussion forums. They proposed two methods such character-based method and term-based method for language identification of noisy text. Marco Lui et al. [13] used generative mixture model inspired from supervised topic modelling algorithms. They experimented on real-world data as well as synthetic data and claimed that their approach outperforms the prevalent approaches. Arkaitz Zubiaga et al. [14] presented a benchmark tweeter dataset for identification of multilingual and similar languages for further research. Marcos Zampieri et al. [15] presented a system for native language identification by using an ensemble of multiple SVM classifiers. Alina Maria et al. [16] also used SVM ensemble learning to identify five languages of the Indo Aryan origin. They trained the models on characters and words of Braj Bhasha, Hindi, Awadhi, Magahi and Bhojpuri.

VI. Experimentation and Result Analysis

Dataset used:

For experimentation, three monolingual corpora of Indian languages such as Marathi, Hindi and Sanskrit are used. Each corpus is containing 1,00000 lines. This big corpus is partitioned into small documents of 100 lines each. Thus, 1000 documents for each language, totally 3000 documents are obtained. 2400 documents are used for training purpose, 300 documents are used for validation and 300 documents are used for testing purpose. The corpus is obtained from Leipzig Corpora Collection [4]. First the corpus is tokenized by applying freely available tokenization tool from IIT, Bombay. Language models are constructed using SRILM language modelling toolkit. All the classifiers are trained and tested using WEKA tool. WEKA provides a good set of kernel functions for SVM classifier. After thorough trials of many kernel functions, Radial Basis Function kernel (RBF) is used for SVM classifier. Phrase based machine translation systems are developed using MOSES decoder [5]. In Table I, the statistics of training, test and validation dataset after all pre-processing is given. Unique words count is vocabulary of each dataset. Number of tokens is total words in each dataset.

TABLE I
Dataset Used for Experimentation

Dataset type	Unique words Count			Number of tokens		
	Hin	Mar	Sans	Hin	Mar	Sans
Training set	98838	86787	94889	1499789	1003947	714304
Validation set	10702	11189	12118	144378	124997	89288
Test set	11049	10998	11979	144879	124839	88129

Hin: Hindi, Mar: Marathi, Sans: Sanskrit

The confusion matrix of language identifiers for all the languages used in experimentation i.e. Hindi, Marathi and Sanskrit, is given in Table II.

TABLE II
Confusion Matrix of Language Identifiers

True language	Predicted as			
	Hindi	Marathi	Sanskrit	
n-gram based Language Identifier				
Hindi	78	13	9	100
Marathi	15	75	10	100
Sanskrit	19	8	73	100
	112	96	92	
Logistic Regression				
Hindi	81	12	7	100
Marathi	9	83	8	100
Sanskrit	8	13	79	100
	98	108	94	
Support Vector Machine				
Hindi	89	7	4	100
Marathi	5	91	4	100
Sanskrit	9	4	87	100
	103	102	95	
Naïve Bayes Classifier				
Hindi	79	12	9	100
Marathi	16	77	7	100
Sanskrit	13	11	76	100
	108	100	92	
k-Nearest Neighbours				
Hindi	82	11	7	100
Marathi	9	84	7	100
Sanskrit	9	11	80	100
	100	106	94	

The results are evaluated using precision, recall and f-measure accuracy metrics. The precision metric gives what fraction of languages assigned class i are actually of class i. It is calculated as:

$$Precision = \frac{True\ class\ i}{Total\ predicted\ as\ class\ i} \quad (3)$$

Recall metric shows what fraction of languages in class i are classified correctly. It is calculated as:

$$Recall = \frac{True\ class\ i}{Total\ actually\ class\ i} \quad (4)$$

The f-measure is calculated using precision and recall as:

$$\text{f} - \text{measure} = \frac{2*Precision*Recall}{Precision+Recal} \quad (5)$$

Accuracy of the classifier is given in terms f-measure. It conveys the balance between the precision and recall. Micro-averaged precision and recall is obtained by taking average of respective measure for each language class for each identifier. Results are given in the Table III.

TABLE III
PERFORMANCE OF LANGUAGE IDENTIFIERS

Language Identifier	Precision	Recall	f-measure
n-gram based	75.71	75.33	75.52
Logistic regression	81.18	81.00	81.09
Support Vector Machine	89.07	89.00	89.03
Naïve Bayes classifier	77.59	77.33	77.46
k-Nearest Neighbours	82.12	82.00	82.06

It is observed that Support Vector Machine based language identifier is more effective for language identification purpose. After developing language identifier, a separate machine translation system is trained for Hindi to English, Marathi to English and Sanskrit to English.

A source document containing intermixed contents in Hindi, Marathi and Sanskrit language is given for translation and a translated document in English is obtained. It is observed that the BLEU score of translation output is improved from 13.71 to 25.51 points due to inclusion of language identification component in machine translation system.

VII. CONCLUSION

This paper presents a brief overview of the need and challenges involved in automatic language identification for machine translation task. Language identification and machine translation is very essential to make cross lingual information available to mass. As Marathi, Hindi and Sanskrit are very closely related languages, getting the distinguishing features that classify them is a difficult task. It is also observed that, most of the misclassified instances are short and noisy. As all these languages share the same script, classification of named entities is also difficult.

It is observed that, machine learning based language identification is more effective than traditional n-grams based approach. Support Vector Machine based language identifier achieves 89% accuracy and it is 18% more than traditional n-grams based approach. Segmentation and translation of individual languages in a multilingual document really improved the quality of machine translation.

REFERENCES

[1] W. B. Cavnar and J. M. Trenkle, "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] Bashir Ahmed, Sung-Hyuk Cha, and Charles Tappert, "Language identification from text using n-gram based cumulative frequency addition," in *Proceedings of Student/Faculty Research Day, CSIS*, Pace University, 7 May 2004.

[3] Bruno Martins and Mário J. Silva, "Language identification in web pages," in *Proceedings of the SAC'05*, Santa Fe, New Mexico, USA, 13-17 March 2005.

[4] D. Goldhahn, T. Eckart and U. Quasthoff, " Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages," in *Proceedings of the 8th international language ressources and evaluation (LREC'12)*, 2012.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, and R. Zens, "Moses: open source toolkit for statistical machine translation," in *Proc. ACL Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180, 2007.

[6] KosuruPavan, Niket Tandon, Vasudeva Varma, "Addressing challenges in automatic language identification of romanized text," in *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, Macmillan publishers, India, 2010.

[7] Abdelmalek Amine, ZakariaElberrichi, Michel Simonet, "Automatic language identification: an alternative unsupervised approach using a new hybrid algorithm," in *Proceedings of IJCSA.* vol. 7, no. 1, pp. 94-107, 2010.

[8] Tromp E. and Pechenizkiy M., "Graph-based n-gram language identification on short texts," in *Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning*, Benelearn, pp. 27-34, 2011.

[9] Deepamala N, Ramakanth Kumar P. "Language identification of Kannada language using n-Gram," *International Journal of Computer Applications,* vol. 6, no. 4, pp. 24-28, May 2012.

[10] Sreejith C, Indu M, Dr. Reghu Raj P. C., "N-gram based algorithm for distinguishing between Hindi and Sanskrit texts," in *Proceedings of the Fourth IEEE International Conference on Computing, Communication and Networking Technologies*, July 4 - 6, 2013.

[11] M. Zampieri, "Using bag-of-words to distinguish similar languages: How efficient are they? " in *Proceedings of the IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI),* pp. 37-41, 19-21 Nov. 2013.

[12] KheireddineAbainia, SihamOuamour, HalimSayoud, "Robust language identification of noisy texts - proposal of hybrid approaches," in *Proceedings of 11th International Workshop on Text-based Information Retrieval (TIR)*, Munich, Germany, September 2014.

[13] Marco Lui, Jey Han Lau and Timothy Baldwin, "Automatic detection and language identification of multilingual documents," *Transactions of the Association for Computational Linguistics*, pp. 27–40, 2014.

[14] Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno, "Tweetlid: a benchmark for tweet language identification," *Language Resources and Evaluation*, vol. 50, no. 4, pp. 729–766, 2016.

[15] Marcos Zampieri, Alina Maria Ciobanu, and Liviu P. Dinu, "Native language identification on text and speech," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics,* Copenhagen, Denmark, pp. 398–404, September 2017.

[16] Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Santanu Pal, Liviu P. Dinu, "Discriminating between Indo-Aryan languages using SVM ensembles," in *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects*, Santa Fe, New Mexico, USA, pp. 178–184, 20 August 2018.