An Unsupervised Word Level Language Identification of English and Kokborok Code-Mixed and Code-Switched Sentences

[1]Enjula Uchoi, [2]Lenin Laitonjam

[1]Lecturer, [2]Supervisor,
[1]Department of Computer Science and Engineering,
[1]Dhalai District Polytechnic, Agartala, India.

Abstract: Considering the increasing uses of multilingual text, the need for an automatic word level language identification model is raised. In this regard, we present an unsupervised model for word level language identification of English and Kokborok code-mixed and code-switched sentences. Several works have already been reported for various languages, including various Indian languages. But, to the best of our knowledge, ours is the first language identification work dedicated to low resource English-Kokborok language pairs. The proposed model combines a frequency lexicon based, character n-gram language model and a language dependent morphological dictionary-based model for correctly classifying each word. The model which is suitable for low resource languages that do not have a large number of annotated dataset is able to achieve a good performance with word accuracy level of 84%.

IndexTerms **- Word level language identification, code-mixing, code-switching, dictionaries, affixes, kokborok, English.**

I. INTRODUCTION

Automatic language identification based on code-mixing and code-switching in Social media is gaining more and more attention for those researchers who are basically working and doing research work in the field of Natural Language Processing. In this generation multilingual speakers are much higher in number than the monolingual speakers, we can see in Social media many of the speakers are multilingual speakers using more than two languages in their conversation in Facebook, tweeter and in many others sources. Taking this into knowledge and knowing the importance of using multilingual language and the advantages of using more than two languages in conversation can help us in different ways. So in this paper, our main objectives were to identify automatically the mixed sentence of the language of English and Kokborok using the Dictionaries based lookup, Affixes and Language model based on character n-gram Markov model. The Kokborok language is a low resource language and it is far lacking back as compared to the English language on the bases of linguistics research, source data available, etc. There are no separate scripts for the Kokborok Language but rather it was written using the Roman script and Bengali language so it is one of the challenging tasks. There are so many phonetic types and context-dependent words as well as ambiguity words. The spelling is the same but they are different in meaning as well as in pronunciation. Many researchers have done based on their own native languages like in Tamil, Assamese, Hindi, Bengali, Nepali, Spanish and Manderin(King and Abney,2013 [1];Das and Gamback,2014[2];Manas Jyoti Bora et al.,2018 [3]; Gokul Chittaranjan and Yogarshi Vyas,2014 [4] ;Utsab Barman and Amitava Das,2014 [5];Menal Dahiya,2017 [6];Yogarshi Vyas and Spandana Gella,2014 [7]). This has given us more motivation and challenges to do this type of task on this low resource based on Kokborok language. The language is our identity and the identity should be recognized in any field especially in technology. Many researchers have done and experimented and got good accuracy result using dictionaries based, character n- gram by using different types of supervised machine learning method(Das and Gamback,2014[2];Gokul Chittaranjan and Yogarshi Vyas,2014[4];Utsab Barman and Amitava Das ,2014[5];Levi King et al., 2014[8];Veena and Kumar[9];Rampreeth Ethiraj et al.,2015[10]). But our approach in this paper is different as compared to the other papers because this is based on Kokborok language, the native language of Tripura Borok. Few papers are being done on the basis of POS tagging (B.G.Patra et., 2012 [11], Stemming (B.G. Patra et al., 2012[12] and Morphological analyzer (Kumber Debbarma et al., 2012 [13]) in Kokborok language, But based on the mixed sentence no one has done so far and ours result shows quite good and by combining with the dictionaries and language model. This can be improved by using some more supervised techniques of different machine learning methods for name entity recognition and by applying Part of speech (POS) tagging to identify the word level language of Kokborok language.

II. BACKGROUND OF KOKBOROK

The word Kok-borok[1]is the combination of two different words which actually"Kok" means a language, and the "Borok" means a Human. This Kok-borok means the language which is being spoken by the people of Tripura especially the Borok people and as well as the Borok people of the other states of Mizoram, Manipur, and Assam as well as our neighboring countries like Myanmar and Bangladesh. Kokborok language, especially in Tripura, is written in Bengali script as well as in the Roman script. Most of the people who speak Kokborok preferred the Roman script rather than the Bengali script. The Kokborok language actually originated from the side of Tibeto-Burman which is the sub-group of the Sino-Tibetan language. Kokborok language is quietly similar to the Bodo language as well as with Kachari language and has 36 dialects. The sub-communities in Tripura and other states like Mizoram and Assam. their own dialects namely Debbarma, Hrangkhawl, Puran, Halam, Tripura, Reang, Jamatia, Noata, Murasing, Uchoi, Rupini which are all different from each other also with Dimasa language which is spoken in some of the states like Assam. Kokborok language. The first Kokborok script was actually known as Koloma. Tripura is the only state in India which was ruled by 184 been merged with the Indian union on 15th October 1949.The scripts that have been used for recording the

history was written down by the Koloma script that is available in the book of Rajratnakar. In the Later, the two Brahmins known as Sukreswar and Vaneswar have translated the language into Sanskrit and then again they translated the chronicle into Bengali in the14th century. This Kokborok language has phonemic tone, word order, stem homophony, and also the verb morphology as well the verb derivational suffixes formed from the verbs. As time passes the Kokborok has finally come in contact with English, Arabic, Bengali, Persian language, etc. The Kokborok words have different types of complex agglutinative structures. Since this Kokborok language does not have its own script so far and it's under process, Due to this, the Kokborok language speaking people from the native of Tripura cannot access information conveniently. Kokborok is not a single language and dialects spoken in Tripura. There are different dialects spoken in Tripura like Debbarma, Reang, Jamatia, Noatia, Lusai, Uchai, Chaimal, Halam, Kukis, Garos, etc. they have their different dialects and face many difficulties in understanding within them, so Code-mixed and Code - switching can be the key source for this aspect. This will helps us to understand stronger with each other and it is very important to explore the needs of the code-mixing and code-switching in Kokborok language with different others language, not only with English but also with Hindi, Bengali, etc. that will make the relationship stronger in between the Multilingual speakers.

III. RELATED WORK

Some of the other approaches are Nguyen and Dogruoz,2013 Their recent studies have word level to support the analysis and for processing of code-switched text and built a character n-gram model for each of the language with the maximum length of 5-gram And they use a logistic regression model and experimented using linear-chain conditional random Fields [14].Somnath Banerjee, et al., 2014, They developed a machine learning based on the CRF model and post processing heuristics for identifying the WLL identification task [15]. Levi King et al.,2014 Developed by using supervised methods on the bases of characters n-grams, due to limited monolingual text documents their capability of capturing in some of the patterns in the texts of code-switching is limited [1]. Cavnar and Trenkle, 1994 Described that using character N-gram method they achieve 100% accuracy [16]. Ali Selamat and Nicholas Akosa, 2016 Proposed algorithm based on the lexicon to perform in document level as well as in sentence level for language identification [17]. Yamaguchi and Ishii, 2012 The problem has been done based on text segmentation by obtaining minimum length in the text [18]. Tommi Vatanen et al., 2010 Identifying any short text by using character n-gram models and using the ranking technique of Cavnar and Trenkle (1994) [19]. Vyas and Gella,2014 described their task by using Part of speech tagging of Hindi and English Language with back-transliteration normalization [7]. Harshi Jhamtani et al., 2014 Has proposed to identify the Code-Switched sentences by using the techniques to find the character sequence [20]. R.V.Kumar et al.,2015 Based on the Indian language of the mixed script and label the words as language1, language2 and used the sequence level to implement the Name entities, mixed script, and punctuation using the Support Vector Machine [21]. A. Zubiaga et al., 2014 have done based on the code-mixed of the identification of the language of Catalan, Spanish, Portuguese, Basque with English from Tweet [22]. Rao and Devi,2016 have done by extracting the entity of the language of Hindi, Tamil with English or code-mixed in a social media text [23].

IV. PROPOSED MODEL

The main concept behind this Word level language code mixing and code-switching of English and Kokborok is that firstly it starts with checking languages that are based on only the current word. and check the features words in the sequences. Then it checks with the words that have the highest probability and get evaluated.

4.1 Dictionary based

In the dictionaries based method, the word has been an extract from the corpora. Then it looks up in dictionaries for words and then chooses that language which has the highest probability and if supposed the word does not present in the dictionaries, we choose English as the language.

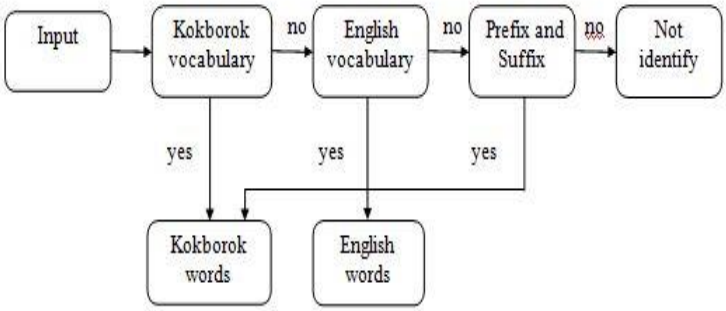

Figure 1: Flowchart for identification WLL using Prefix and Suffix.

4.2 Prefix and Suffix

Since Kokborok is very rich in affixes so the reason of applying the prefix and suffix of Kokborok language is to identify those words which contain only prefix and suffix in the Kokborok Language and this will help us in identifying the Kokborok language easily. The Kokborok prefix and suffix words are shown in Table 1

1) Algorithm: Using Kokborok Affixes as shown in Fig. 1

1. Input the mixed sentence.
2. If found in Kokborok vocabulary identity as Kokborok words else it will identify as English words.
3. If not identified in both the language, it will check the Kokborok Affixes.
4. If it identifies prefix and suffix words identify as Kokborok words.
5. Exit.

4.3 Language model

The language model is built using a character n-gram. The maximum length of each language is 4 and by applying the smoothing method and including the word boundaries so that the probabilities can be calculated. As shown in Fig.2 D. Dictionary based(DICT) + Language model(LM) The reason for the Dictionary and Language model is that if the words are in the dictionaries based, a language model we have to use the logistic regression model for the following purpose with independent and contextual dependence.

4.4 Experimental setup

4.4.1 Language identification: After the words are tokenized it is run in the language identification module and check whether the token words are Kokborok or English. The language of each token is being identified by combining the information from the dictionary and character n-gram.

4.4.2 Independent identification: Individual word independent Label which is done by the Dictionary and language model. Our task is similar with the work done by King et al.,(2014) [8] using the character n-gram, lexical and with word label probabilities. the frequency of the lexical(lex) and character(ch) are linearly interpolated and the combined(comb) probability of both is given in the equation below. (1)

$$P_{comb} = \lambda_{lex} * P_{lex} + \lambda_{ch} * P_{ch} \quad (1)$$

Where $0 \leq \lambda_{lex}, \lambda_{ch} \leq 1$; where $\lambda_{lex} + \lambda_{ch} = 1$. Those words that do not appear in the lexicon based receive a very low and non-zero lexical probability. Supposed if the word does not present in both the lexicon language then the lexical probability is not included and the implementation is done on the basis of the character model, even if $\lambda_{lex} = 1$.

Our approach of the character model is simply based on the character n-gram Markov chain, as done by Dunning(1994) [24]. This is done by estimating the initial and transition probability that is by counting n-gram at the starting of tokens and after other n-grams. The 2-gram and the 4-gram models have been trained and their performance compared is shown below.

$$P_{initial}(<C1C2) = \frac{count(<C1C2)}{count(tokens)} \quad (2)$$

$$P_{transition}(C2,C3,C4 \mid C1,C2,C3) = P_{transition}(C_4 \mid C_1,C_2,C_3). \; P_{transition} = \frac{count(<C1,C2,C3,C4)}{count(C1,C2,C3)} \quad (3)$$

Using the language model, the probability of a word can be calculated by multiplying the initial and transition word of a probability (4)

$$P(<C_1C_2C_3>) = P_{initial}(<C_1C_2) * P_{transition}(C_3 \mid < C_1C_2) * P_{transition}(> \mid C_1C_2C_3) \quad (4)$$

Since this is going to apply in the large corpus so there is a chance of occurring many data sparsity and in some cases the character combination may not be seen in some cases, this may lead the result of the probability into zero. To overcome these types of problems a lambda smoothing is being used by assigning a tiny value of non-zero probability to n-gram involving unseen characters. The lambda value is fixed to 0.001.

$$P_{initial}(<C1C2) = \frac{count(<C1C2 + \lambda)}{count(tokens + \lambda(N + 1)} \quad (5)$$

where N= number of different that is observed in n-gram

$$P_{transition}(C4 \mid C1C2C3) = \frac{count(<C1C2C3C4) + \lambda)}{count(C1C2C3) + \lambda(D + 1)} \quad (6)$$

where D= number of differences that is observed in n-gram. For unseen n-gram the probability is assumed to be in with same probability(7)

$$P_{transition}(C4 \mid CCC_{unseen}) = 1 (S + 1) \quad (7)$$

Where S= is the character size

Table 1: Suffix and Prefix words

Suffix words	Prefix words
ana , anw	a
Bukumuini , brebre , bai, bwswk, brubru ,bo , bubuk	Bu ,bw , bi, ba, buse , bo
Chumu ,chom	-
di, dudu , drop , drudru , de	-
Glak , gra , ga , go , ganag , gara , galak , gwja	-
hiya	-
-	I , iri
Ja , jago , jak , jadi , jaba	Jwk , jwla
Khlai, kha, khu , kwrwi , khan , kwrwng , khamun , khai , kho	Ko , kw , ku , ke , ka ,ke ,ki , k
Li , lan , lai , lolo , lwlwk , liya , lainai	-
Mani , mathanglainai, marii , mariui , ma , mung	Ma , masema , mu , mw
Nai , na , naide ,niya	Nw ,nu , nase , nwse
o	o
Roro , rok , rokni	Ri , ring , rise , rose
Sa , si , sinai	se
Ta , thang , thokthok , thok , twi	-
ui	u
Witongo , witongmani , wi	-
Ya , yakhu , yade , yung , yana	-

4.4.3 Contextual identification:

Context-dependent is basically done by using the previous and current token and by taking the current token with the next token which is the result of the language model. The purpose of using this context dependent was to identify the similarity words which are present in both the languages because there is a chance that the same words of frequency value may be the same and this may lead us to misclassify the words to whom it belongs to. Those words which are similar in both the languages are e,g ***'no','o','da(day)','sa(say)','do','ring','rose'*** and many others. So when this type of ambiguity occurs, it is very tough to identify those languages and sometimes it gets misclassified. To overcome these types of problems in the language identification we proposed the sequence of language models using the supervised method of machine learning with discriminative Hidden Markov Model. The transition probability is assumed to be the same for both the language. The main parameter that we need to specify is the continuing probability of the same language P_{count}. By which we can infer the probability of switching the languages.

$$P_{switching} = 1 - P_{count} \quad (8)$$

The purpose of using the Viterbi algorithm is to find the optimal language in the sequences of tokens based on P_{count} ,,$P_{Switching}$.

4) Examples:1.Norokni ***class o*** khorok bwswk students tong"

<EN>How many students are there in your class<EN>

5) Examples: 2. Nini ***table no*** kisa buskang tidi.

<EN> Bring your table little forward<EN>

6) Examples: 3. Nung no ang nainai just wait" <EN> I will see you just wait <EN>

From the above examples 1 and example 2 of mixed sentences of English and Kokborok, the second and third word in both the phrases can be misclassified without using the contextual information, because both of the words are the same for both languages. Since Kokborok language is written in Roman script and the word of ***class*** and ***table*** is being borrowed from English language. Whereas in example 3, the second word ***'no'*** is there for both the languages but in this case the word ***'no'*** means ***'you'*** is English language is actually representing the Kokborok words, which can be misclassified if the contextual dependent is not applied to it.

4.5 Data collection

We have collected the corpus for both the dictionary and the character model from the Crubadan corpus (Scannell, 2007) [25]. In this corpus a large number of many languages are there and also include many minority languages like Kokborok. The dataset which is available in the Crubadan corpus is not sufficient for our task so we have a test with a new dataset of Kokborok Bible. The Sources are very few as compared to other languages. So in this paper, the data of mixed sentences has being collected from the WhatsApp internet forum from the groups of United Tiprasa Forum(UTF), Krishna Nagar (EU), Tipra Kwsrang Bodol, Tribal Engineering Society, NIT Students Borok, Information Yarung, Tripura Uchoi Youth Association(TUYA), Platinum Jubilee, Uchoi Christian Fellowship(UCF), Depachara Baptist youth, Global Tiprasa Engineers, Depachara Baptist Church And other organization.Total 7000 mixed sentences of English and Kokborok has being collected from the month of August 2018 to February 2019 from the WhatsApp groups. The total number of 34 prefixes and 77 suffixes of Kokborok has been collected from the Kokborok- English dictionary, learn Kokborok ,A Kokborok grammar and from the Tripura university of Kokborok Department. As shown in Table 2, Table 3 and Table 4.

Table 2: Kokborok Data

Data Name	Total words
Kokborok Bible	718,883
Mixed sentences of English and Kokborok	59,419

Table 3: Kokborok Affixes

Name Affixes	Number of Affixes
Prefix	34
Suffix	77

Table 4: Character 2-Gram Model

Languages	Number of Tokens
English(en)	1767141
Kokborok(trp)	711509

V. RESULTS AND DISCUSSION

After observing the Total Token (TT), True Count (TC) and Error Count (EC) in Dictionary (DICT) , Affixes, Dictionary with Affixes and Dictionary with Language model(LM), our system has achieved higher accuracy of 84%. As shown in the Table 6. And the Bar Chart and Line graph is shown in Fig.2 and Fig.3. 1) Error Analysis: Some of the errors that has been observed using this classifier in word level language identification of English and Kokborok. After observing the Total Token (TT), True Count (TC) and Error Count (EC) in Dictionary (DICT), Affixes, Dictionary with Affixes and Dictionary with Language model (LM), our system has achieved higher accuracy of 84%. As shown in the TABLE V. Most of the name of the place, organization, e.g like ***Office-o***, ***Bazar-o***, ***Month-o, Seven-o***, in Kokborok language are joined with the suffix 'o' and like in the action words in Kokborok e,g ***chalainai do***, ***thanglainai do***, ***jotono no, colour no*** is the joined words with suffix ***'no'*** and ***'do'*** this types of words is misclassified in the word level language identification. Another joined word like ***paaaa....please***, ***o almost*** and there are many ambiguity words which are joined with the Kokborok word as well as in English word which are misclassified in our task shown in Table 5.

Table 5: Identification of Errors word

Mixed sentence of English and Kokborok	Word Errors
Office-o/trp belaikha/trp *pressure*/en rwjago/trp	***Office-o***/en
O/trp ***meeting***/en ***ni***/trp tamo/trp outcome/en ongkha/trp	***O***/trp ***meeting***/en ***ni***/en
Please/en bono/trp kisa/trp naidi/trp ***do***/trp ***please/en do***/trp	***do***/trp ***pleas***e/en ***do***/en
o/trp ***almost***/en done/en kisa/trp naising/trp di/trp	***o-almos***t/en

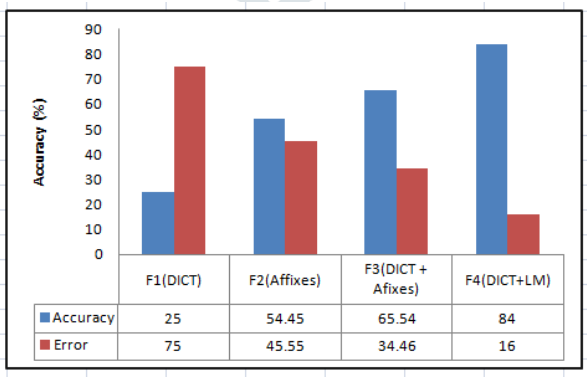

Fig. 2. Result of language identification of code-mixing and code-switching of english and kokborok

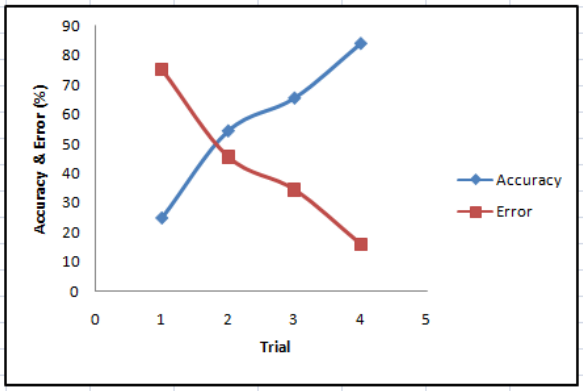

Fig. 3 Line graph of Accuracy and Error.

Table 6: Result

Train	TT	TC	EC	Accuracy %	Error%
DICT	3575	8939	26816	25.00	75.00
Affixes	35755	19469	16286	54.45	45.55
DICT +Affixes	35755	23434	12321	65.54	34.46
DICT + LM	35755	30034	5721	**84.00**	16.00

VI. CONCLUSION AND FUTURE SCOPE

The main aim of this research was to identify and understand the natural way of using code-mixing and code switching in Kokborok and English. This is done by using the word uni-gram which is one of the most basic features . This character n-gram has also proved to be one of the useful methods for implementing the task for identifying the language. Actually, the Prefixes and suffixes are not the same as the character n-grams but our expectation was that this prefix and suffix will help us in capturing the text of similar features. Our results reveal that the language model is performing much better than the dictionaries based method and by adding the context-dependent method it shows the improvement in the performance. The challenging part was the highly informal spelling occurrences in the online environments and as well as the occurrences of named entities. Our approach is language-independent and context-dependent and this system can even be implemented for other low resources languages and even for developing the other Indian languages also. In this current stage, our achievement was with the total accuracy of 84%.In the future, our planning was to carry out some more experiments by using different machine learning.This should clearly explain the main conclusions of the work highlighting its importance and relevance.

REFERENCES

[1] B. King and S. Abney, “Labeling the languages of words in mixed language documents using weakly supervised methods,” in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 1110–1119.
[2] A. Das and B. Gamb¨ack, “Identifying languages at the word level in code-mixed Indian social media text,” in Proceedings of the 11th International Conference on Natural Language Processing, 2014, pp. 378–387.
[3] M. J. Bora and R. Kumar, “Automatic word-level identification of language in assamese English Hindi code-mixed data,” in 4th Workshop on Indian Language Data and Resources, Proceedings of the Eleventh
[4] G. Chittaranjan, Y. Vyas, K. Bali, and M. Choudhury, “Word-level language identification using crf: Code-switching shared task report of msr India system,” in Proceedings of The First Workshop on Computational Approaches to Code Switching, 2014, pp. 73–79
[5] U. Barman, J. Wagner, G. Chrupała, and J. Foster, “Dcu-uvt: Word level language classification with code-mixed data,” in Proceedings of the First Workshop on Computational Approaches to Code Switching, 2014, pp. 127–132.
[6] M. Dahiya, “Word-level language identification in bilingual text and back-transliteration,” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 6, no. 6, 2017.
[7] Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury, “Pos tagging of english-hindi code-mixed social media content,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 974–979.
[8] L. King, E. Baucom, T. Gilmanov, S. K¨ubler, D. Whyatt, W. Maier, and P. Rodrigues, “The iucl+ system: Word-level language identification via extended markov models,” in Proceedings of the First Workshop on Computational Approaches to Code Switching, 2014, pp. 102–106.
[9] P. Veena, M. Anand Kumar, and K. Soman, “Character embedding for language identification in hindi-english code-mixed social media text,” Computación y Sistemas, vol. 22, no. 1, pp. 65–74, 2018.
[10] R. Ethiraj, S. Shanmugam, G. Srinivasa, and N. Sinha, “Nelis-named entity and language identification system: Shared task system description.” in FIRE Workshops, 2015, pp. 43–46.
[**12]** B. G. Patra, K. Debbarma, S. Debbarma, D. Das, A. Das, and S. Bandyopadhyay, “A light weight stemmer in kokborok,” in Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012), 2012, pp. 318–325.
[13] K. Debbarma, B. G. Patra, D. Das, and S. Bandyopadhyay, “Morphological analyzer for kokborok,” in Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, 2012, pp. 41– 52.

[14] D. Nguyen and A. S. Do˘gru¨oz, "Word level language identification in online multilingual communication," in Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 857–862.
[15] S. Banerjee, A. Kuila, A. Roy, S. K. Naskar, P. Rosso, and S. Bandyopadhyay, "A hybrid approach for transliterated word-level language identification: Crf with post-processing heuristics," in Proceedings of the Forum for Information Retrieval Evaluation. ACM, 2014, pp. 54– 59.
[16]W.B.Cavnar, J.M.Trenkleetal., "N-gram-based text categorization,"in Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, vol. 161175. Citeseer, 1994.
[17] A. Selamat and N. Akosu, "Word-length algorithm for language identification of under-resourced languages," Journal of King Saud University Computer and Information Sciences, vol. 28, no. 4, pp. 457–469, 2016.
[18] H. Yamaguchi and K. Tanaka-Ishii, "Text segmentation by language using minimum description length," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers Volume 1. Association for Computational Linguistics, 2012, pp. 969– 978.
[19] T. Vatanen, J. J. V¨ayrynen, and S. Virpioja, "Language identification of short text segments with n-gram models." in LREC, 2010.
[20] H. Jhamtani, S. K. Bhogi, and V. Raychoudhury, "Word-level language identification in bi-lingual code-switched texts," in Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing, 2014.
[21] R. V. Kumar, M. A. Kumar, and K. Soman, "Amrita cen nlp@ fire 2015 language identification for Indian languages in social media text." in FIRE Workshops, 2015, pp. 26–28.
[22] A. Zubiaga, I. San Vicente, P. Gamallo, J. R. P. Campos, I. A. Loinaz, N. Aranberri, A. Ezeiza, and V. Fresno-Fern´andez, "Overview of tweetlid: Tweet language identification at sepln 2014." in TweetLID@ SEPLN, 2014, pp. 1–11.
[23] P. R. Rao and S. L. Devi, "Cmee-il: Code mix entity extraction in Indian languages from social media text@ fire 2016-an overview." in FIRE (Working Notes), 2016, pp. 289–295.
[24] T. Dunning, Statistical identification of language. Computing Research Laboratory, New Mexico State University Las Cruces, NM, USA, 1994.
[25] K. P. Scannell, "The cr´ubad´an project: Corpus building for under resourced languages," in Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, vol. 4, 2007, pp. 5–15.