Word-Level Language Identification in Bilingual Text and Back-Transliteration

Menal Dahiya

***Abstract*— With the advances in online social networking and communication, the challenge lies in providing the user the platform that allows communication with various languages. Once, the platform is provided, this needs to be pre-processed and transformed that system can understand. Transliteration is the conversion of a text from one script to another. In this paper, we describe a system for word-level language identification of mixed text. The proposed system uses a method based on list searching and minimum edit distance. The performance of the proposed system is carried on the test sets provided by the shared task on language identification for English Hindi (En-Hi) pair. The experimental results show a consistent performance with high precision.**

***Index Terms*— Back-Transliteration, Language Labeling, Mixed-text.**

I. INTRODUCTION

Language identification at the document level has been considered an almost solved problem in some application areas, but language detectors fail in the social media context due to phenomena such as utterance internal code-switching, lexical borrowings, and phonetic typing; all implying that language identification in social media has to be carried out at the word level [1].

Most of the languages are written using indigenous scripts, i.e. Hindi is written in Devanagari. However, often the web-sites and the user generated content (such as tweets and blogs) in these languages are written using Roman script, after transliteration. Transliteration, the process of converting words into Roman script, is being on the web abundantly on the Web not only for documents, and user queries that are used to search for these documents. In this paper we are proposing the solution of language identification in bilingual text (Hindi and English), and back-transliteration of Hindi Words e.g Desh ki population' this is a Hindi English mixed message, where Desh ki is in Hindi and the population is in English. We are also targeting Named Entity in Indian Languages by marking proper nouns [2]. A challenge which is faced while processing translated queries is because of extensive spelling variation. For in-stance, the word Dhanyavad ("thank you" in Hindi and many other Indian languages) can be written in Roman script as "dhanyavaad, dhanya-vada, dhanya-bad, dhany-vad, danya-vad, danya-vaad ", this unique situation is tackled which is found in Web search for users of many languages around the world because this is a problem of higher importance. We have used ML approaches and a List to label the words [3].

Manuscript received May, 2017.

Menal Dahiya, Assistant Professor, Dept. of Computer Science, Maharaja Surajmal Institute, Janakpuri, Delhi, India.

II. RELATED WORK

Automatic language identification research has focused on identifying both spoken languages as well as written texts. Language identification of speech has been studied by House and Neuburg (1977), where the authors assumed that the linguistic classes of a language are probabilistic functions of a Markov chain. Language identification of written texts has been studied at document-level as well as at word-level perspectives.

The N-gram-based approach for chops texts up in equally-sized character strings, N-grams, of length n [4]. The assumed that is used is that every language uses certain N-grams more frequently than other languages, thus providing a clue on the language the text is in. This idea works due to Zipf's law stating that the size of the r-th largest occurrence of the event is inversely proportional to its rank r. Experimental studies in suggest that using trigrams (at the character level) generally yields the best results.

In the out-of-placement measure is used to compare unlabeled text against the model. This measure sorts the N-grams in both the model as well as the unlabeled text separately based on their occurrence counts and compares the model's occurrence list with the texts list. But, it was shown that accuracy in results is found in the use of a cumulative frequency based more time e client. The out-of-placement measure works well when sufficient training data is available whereas the cumulative frequency measurement works equally well with little data at hand. Therefore we will use the cumulative frequency measurement in our experiments [5].

Automatic language identification (LID) is the process of using a computer system to identify the language of a spoken utterance. Formal evaluations have indicated that the most successful approach to automatic language identification relies on using the phonotactic content of a speech signal to discriminate among a set of languages. Systems which are based on phonotactic characteristics, such as PPRLM (Parallel Phone Recognition and Language Modeling), set of phone recognizers is typically employed to generate a parallel stream of what we call as phone sequences and a bank of n-gram language models to capture the phonotactics [6]. Although phone based systems provide the best LID performance, their heavy computational demands may preclude their use in low cost, real-time applications. An alternative approach to LID uses Gaussian mixture models (GMMs) to classify languages using the acoustic content of

the speech signal. Although GMM systems are quite efficient, they do not provide the superior performance of phone based LID systems. Recently a variation of the phonotactic approach was proposed in which a Gaussian mixture model, rather than a phone recognizer, was used to tokenize the incoming speech [7]. This approach produced a GMM LID system whose performance was competitive with phone-based approaches but whose operation was much faster.

The present work reports on the performance of GMM-based LID systems that use shifted-deltacepstral (SDC) coefficients as a means of incorporating additional temporal information about the speech into the feature vectors. The use of temporal information spanning a large number of frames is motivated by the success of phonetic approaches that naturally base their tokenization over multiple frames. It will be shown that GMM-based LID systems that use SDC feature vectors perform as well as PPRLM and at a greatly reduced computational cost.

Stochastic process also is called the first order Markov process if its state ck in time k depends only on previous state c_{k-1} in time k − 1 (Formula 1).

$$P(c_k | c_0, c_1 \ldots c_{k-1}) = P(c_k | c_{k-1}). \quad (1)$$

In general, the n-th order Markov process which is used is described in Formula 2.

$$P(c_k | c_0, c_1 \ldots c_{k-1}) = P(c_k | c_{k-n}, \ldots, c_{k-1}). \quad (2)$$

The character sequence ck−n. . . ck−1 is named as Markov process prefix (also the term context is used), c_k is usually as named suffix.

Learning Phase – Creating Models of Language Categories a training text document (representative of particular language) is processed as a stream of characters. This stream is divided into Markov processes with length k characters (k is the order of Markov process). Each Markov process those are unique are together stored with information about its number of occurrences. After processing the whole document, all counts of Markov processes are converted into probabilities using Formula 3 (k-th order Markov processes).

$$p(w_1 w_{k+1}) = \frac{T(w_1 \ldots w_{k+1}) + 1}{T(w_1 \ldots w_k) + |A|} \quad (3)$$

Where |A| is the size of an alphabet, T ($w_1 \ldots w_k$) is number of occurrences of Markov process prefix,

T ($w_1 \ldots w_{k+1}$) is number of occurrences of the whole Markov process and p ($w_1 \ldots w_{k+1}$) is the computed probability [8].

III. Proposed System and Results

Language identification (LI) is an important task in natural language processing. Several machine learning approaches have been proposed for addressing this problem, but most of them assume relatively long and well written texts. We propose a graph-based N-gram approach for LI called LIGA which targets relatively short and ill-written texts.

We have studied the problem of language identification on relatively short texts typical for social media like Twitter. Earlier works on language identification showed promising and highly accurate results for well-constructed, sufficiently long enough texts. However, the results were shown to deteriorate considerably when texts become a few hundred characters long. Besides that, our experiments suggested that LIGA is less likely to be sensitive to use of jargon or to domain boundaries.

An open issue not addressed in related work on language identification is dealing with an absence of certain N-grams or words that nuances the certainty of classification. We overcome this limitation in the proposed LIGA approach. When an unlabeled text contains a number of N-grams not present in a learned model, the learned model will not be condent about the label. This suggests assigning condense scores to the assigned labels that can be utilized in the further natural language processing routine or in the mechanism suggesting updates to or relearning of the language identification models.

We evaluated the performance of our approach only on short texts extracted from Twitter with respect to the goals of this study. However, it would be interesting to compare these results with results obtained from using longer texts or texts extracted from other sources. In addition to regarding other sources, using more data and especially incorporating more languages, gives stronger results and a broader comparison.

IV. Conclusion and Future Work

In this paper, we described a method of labeling and mapping words from a mixed bilingual text, and back transliterate Hindi words into native script using list based searching and other models. This model can be further used in shared task for building new design / product in field of artificial reality as well. The model is aiming a very important part of language processing which is rarely addressed .We will be making a smarter model soon with improved efficiency and increased accuracy.

In future, we are targeting a predictive model now, which can work on informal words (nt= NOT), as they are not found in dictionary. Also we are working on words which are combination of two or more individual words, for both languages. (E.g. Parmeshware= param + eshwara, lejayenge= le+ jayenge). We plan to do this parallel to increase our processing speed. We will be coming with tries model for faster search, organized data-structure, and training for new words.

REFERENCES

[1] S. Das and A. Kumar, "Performance Evaluation of Dictionary Based CLIR Strategies for Cross Language News Story Search," http://www.isical.ac.in/~fire/2013/slides/clinss_sujoy_fire13.pdf, 2013.

[2] B. King and S. Abney, "Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods," *Proceedings of NAACL-HLT*, pp. 1110-1119, 2013.

[3] P. J. Antony and K. P. Soman, "Machine Transliteration for Indian Languages: A Literature Survey," *International Journal of Scientific Engineering Research*, Vol.2, Issue.12, PP. 1-8, December 2011.

[4] W. B. Cavnar and J. M. Trenkle, "N-gram-based Text Categorization," http://odur.let.rug.nl/~vannoord/TextCat/textcat.pdf, 1994.

[5] S. Gella, J. Sharma and K. Bali, "Query Word Labeling and Back Transliteration for Indian Languages: Shared Task System Description," http://spandanagella.com/files/FIRE_2013.pdf, 2013.
[6] FIRE 2013 Shared Task detailed description: FAQ retrieval using noisy queries, http://www.isical.ac.in/˜fire/faq-retrieval/2013/faq-retrieval.html, 2013.
[7] U. Z. Ahmed, K. Bali, M. Choudhury and B.V. Sowmya, "Challenges in Designing Input Method Editors for Indian Languages: The Role of Word-Origin and Context," *IJCNLP Workshop on Advances in Text Input Methods*, Association for Computational Linguistics, November 2011.
[8] P. Vojtek and M. Bielikov´a, "Comparing Natural Language Identification Methods based on Markov Processes," http://www2.fiit.stuba.sk/~bielik/publ/abstracts/2007/slovko2007vojtek-bielik.pdf, 2007.

Ms Menal Dahiya is Assistant Professor of Computer Science at Maharaja Surajmal Institute (Affiliated to GGSIP University, Dwarka) and received P.hD from Maharshi Dayanand University, Rohtak in the Department of Computer Science and Applications. She received her M.Phil in Computer Science from Chaudhary Devi Lal University, Sirsa, India in 2007. Before she had studied at Guru Jambheshwar University of Science and Technology (GJU), Hisar and Kurukshetra University, India. Her main research interest are Neural Network, Wireless Security and Wireless Communication. Several of her research papers have been published in International and National peer-reviewed journals indexed in Scopus, Copernicus, ICI and in UGC approved list.