INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student : Neelakshi Sarma

Roll Number : 156101011

Programme of Study : Ph.D.

Thesis Title: Automatic Language Identification in Online Multilingual Conversations

Name of Thesis Supervisor(s) : Dr. Sanasam Ranbir Singh and Prof. Diganta Goswami

Thesis Submitted to the Department/ Center : CSE

Date of completion of Thesis Viva-Voce Exam : 25/02/2021

Key words for description of Thesis Work : Language Identification, Social Media, Multilingual, Code-mixed

SHORT ABSTRACT

With the abundance of multilingual content on the Web, Automatic Language Identification (ALI) is an important pre-requisite for different Natural Language Processing applications. While ALI of well-edited text over a fairly distinct collection of languages may be regarded as a trivial problem, ALI in social media text is considered to be a non-trivial task due to the presence of slang words, misspellings, creative spellings, and special elements such as hashtags, user mentions, etc. Additionally, in a multilingual environment, phenomena such as code-mixing and lexical borrowing make the problem even more challenging. Further, the use of the same script to write content in different languages whether due to transliteration or due to shared script between languages imposes additional challenges to language identification. Also, many existing studies in ALI are not suitable for low resource languages due to either of the two reasons. First, the languages may actually lack the resources required like dictionaries, annotated corpus, clean monolingual corpus, etc. Second, the languages may consist of the basic resources in the native scripts, but due to the use of transliterated text, the available resources are rendered useless. Considering the challenges involved, this thesis work aims to address the problem of automatic language identification of code-mixed social media text in transliterated form in a highly multilingual environment. The objective is to use minimal resources so that the proposed techniques can be easily extended to newer languages with fewer resources.

Although the language identification techniques explored in this study are generic in nature and not specific to any languages, to conduct various experimental investigations, this study generates three manually annotated and three automatically annotated language identification datasets. The datasets are generated by collecting code-mixed user-comments from a highly multilingual social media environment. Altogether, the datasets are composed of six languages - Assamese, Bengali, Hindi, English, Karbi and Boro. Apart from dataset generation, this thesis work makes four important contributions. First, it studies the language characteristics of user conversations in a highly multilingual environment. Interesting observations with regards to language usages and factors influencing language choices in a multilingual environment are obtained from this study. Second, a technique for sentence-level language identification is proposed taking advantage of the social and conversational features in user conversations. The proposed technique outperforms the baseline set-ups and enhances language identification performance in a code-mixed noisy environment. Third, a word-level language identification framework is proposed that makes use of sentence-level language annotations instead of traditionally used word-level language annotations. The proposed method focuses on learning word-level representations by exploiting sentence-level structural properties to build suitable word-level language classifiers. The proposed technique substantially reduces the manual annotation effort required while yielding encouraging performance. Fourth, a word-level language identification technique is proposed that makes use of a dynamic switching mechanism to enhance word-level language identification performance in a highly multilingual environment. The proposed switching mechanism attempts to make the correct choice between two different classification outcomes when one of the outcomes is incorrect. The proposed framework yields better performance than the constituent classifiers trained over a set of non-complementary features. The proposed set-up also outperforms the baseline set-ups using mini-mum annotated resources and no external resources thus making it suitable for low resource languages.

The various automatic language techniques proposed in this study make use of minimal resources. Information obtained from the same set of sentence-level annotated data is used to train both sentence-level as well as word-level classification models. As such, the proposed techniques are also deemed suitable for automatic language identification of low resource languages. The proposed techniques are also able to enhance language identification performance in a code-mixed noisy environment.