**Amrita Vishwa Vidyapeetham  
Amrita School of Engineering, Coimbatore  
Department of Computer Science and Engineering  
19CSE453 NLP  
  
  
  
  
  
  
  
  
  
  
Code-Mixed Translation: Parallel Corpus Creation**

|  |  |
| --- | --- |
| **Roll No** | **Name** |
| CB.EN.U4CSE19305 | ASHWANTH K |
| CB.EN.U4CSE19317 | D T SUVANESH |
| CB.EN.U4CSE19318 | DHARUN NARAYANAN L K |
| CB.EN.U4CSE19327 | K R RITHICK |

Abstract

Code-mixing, the use of two or more languages in a single sentence, is generated by multi-lingual speakers globally. Digital communication media such as Facebook, Twitter, Instagram, and others have allowed people to have many more informal interactions than ever before in the last decade. Because of the informal nature of interactions, a new hybrid language code-mixed data has evolved, with no properly specified framework. As a result, there is an increasing demand for code-mixed hybrid language translation into standard languages. Existing machine translation algorithms, on the other hand, fail to adequately translate code-mixed text because of a lack of gold parallel data. With the ease of internet access, the number of social media users in India topped 518 million in 2020. Furthermore, by 2040, the number of people using social media in the country is speculated to reach 1.5 billion. At this rate, India will give rise to enormous amounts of codemixed data due to the bilingual (mother tongue and English) nature of social media conversation. A proper translation of this data into pure English form can be used in sentiment analysis, which is essential in many real-world applications such as stance detection, review analysis, recommendation systems, and so forth. Hence, we are planning to create parallel corpus for the code-mixed data of Tamil and English and use it to translate the code-mixed data into a single language.

Keywords: Code-mixed, Parallel-corpus, Tamil