Automatic Language Identification

Automatic Language Identification (ALI) is the task of automatically identifying languages in a given piece of text. ALI is important for different natural language applications like machine translation, information retrievel etc. For example, in machine translation, it is first important to identify the language in the source text to correctly translate it to the target language.

Automatic Language Identification can be explored at different levels:

- Document Level Language Identification : Identify the language of an entire document. e.g. in news articles or historical documents

- Sentence Level Language Identification : Identify the language of a particular sentence. e.g. in multilingual Web documents

- Word Level Language Identification : Identify the language of each word in a given sentence/document. e.g comments in Facebook, tweets in Twitter etc.

Word Level Language Identification in social media text:

Social media data processing is important in the current times for different applications like sentiment analysis, marketing and advertisements etc. However, since most of the text based applications are language dependent, one of the pre-requisites for these text processing applications is to appropriately identify the language in the underlying text.

Some of the factors that make language identification in social media more challenging than other domain like news articles, web documents are :

- Social media text are short and noisy with improper grammer and spellings
- Code-mixing i.e., mixing of different languages within the same sentence.
- Phonetic typing i.e. use of non-native script to write text in native language e.g. writing Hindi using Roman script.

In the literature, word language identification has mostly been projected as a supervised learning (classification) problem wherein an annotated corpus is used to train a machine learning model to generate predictions from new unseen text. Some of the classification models commonly used include :

- Traditional models like SVM, Logistic Regression etc. that treat the features as independent bag of words
- Sequence classification models like HMM, CRF etc
- Neural network based sequence classification models like RNNs etc.

Neural Network based approaches have become more common in the recent times because they do not require explicit feature engineering. To understand some of the aspects that these models must address for word language identification, let us consider the following examples:

Consider the words “**god**” and “**ache**” in the following examples:

- Yaar tu toh **god** hai
- I believe in **god**

- **Ache** din zaroor ayenge
- He is having stomach **ache**

We see that depending on the context, the same word belongs to different languages. In the first two examples, the word "**god**" is English although the context are in different languages. In the second set of examples, the same word "**ache**" is Hindi and English respectively depending on the context. Therefore, effectively capturing context information is important for appropriately identifying languages of word.

BERT is a deep language reprsentation model that tries to capture the context in both forward as well as the backward direction. Therefore, it will be interesting to see the performance of BERT over language identification tasks.

To evaluate the BERT over language identification, the following experiments can be conducted:

1. Train a BERT model end-to-end for language identification to find the language labels for an entire sequence of words.

2. Pretrain the BERT model and use the pre-trained embeddings into different word clasification frameworks like :

    I. Sequence classification models like LSTM models, Transformer model etc. wherein the pre-trained BERT embeddings will be used to initialize the embeddings and the model will predict the language labels for the entire sequence.

    II. Feed-forward networks like MLP and CNN wherein instead of predicting the language labels for every word in the sequence, the model will predict the label for a paricular word in a given context

3. Compare the performance obtained to that obtained using traditional word representation models like skip-gram models . Models I and II above can be used for comparison with the difference being the embeddings used.

4. *You are also encouraged to explore other models for language identification as well as other contextualized word representation models like context2vec, ElMo etc.*