Notification No. 513/2022

Date of Award: 17-05-2022

Name of Scholar: Mohd Zeeshan Ansari

Name of Supervisor: Prof. Tanvir Ahmad

Name of the Department: Computer Engineering

Topic of Research: **Analysis and Design of Mixed Lingual Information Extraction Models and Methods**

Findings

Information Extraction systems are often used to extract meaningful entities and their relationships from unstructured text. Named Entity Recognition (NER) as a subtask of Information Extraction can identify generic entities like locations, people, and organizations in a piece of text. In languages without a rich vocabulary, extracting names or even nouns is challenging. Moreover, variations in language orthography and syntactic and semantic structures hinder language modelling and categorization. The work presented in this thesis is based on the extraction of structured information from Hindi-English mixed lingual text. The presented work augments Named Entity Recognition tasks with Language Identification, Word Sense Disambiguation and Transliteration tasks. Unique approaches for discriminating between texts in Hindi and English languages are proposed. The work includes a novel framework for categorizing interlingual homographs in mixed-lingual text. An end-to-end mechanism for transliterating Hindi into Roman script is offered. An innovative methodological approach to enhance mixed lingual Named Entity Recognition by incorporating linguistic information into current systems has been proposed.

This research aims to investigate and resolve the challenges in the development of the mixed lingual information extraction system capable of automatically detecting multilingual named entities in Hindi English text. It provides a collection of models and methods that augments automatic language identification and transliteration with named entity recognition to generate the mixed lingual text using English and Hindi Wikipedia and classify named entities. The contribution to the thesis can be summarized under the subtasks of (1) language identification of mixed lingual text (2) Homograph Language Identification using Word Sense Disambiguation (3) Hindi-English Transliteration and (4) Mixed Lingual Named Entity Recognition.

Language identification being the subtask in this work employs one method based on the probabilistic modelling technique for generating word embeddings which show comparable performance with existing models based on pre-trained embeddings. The alternative method proposed is based on the BERT modelling for both input representation and classification. The effect of vocabulary generation using methods in classification task are inspected and found effective. In the second approach, a large number of language lexicons are prepared as two-dimensional measurable language strength to show the likelihood that a word belongs to the Hindi or English languages. These language lexicons were created by utilizing the complimentary English and Hindi Roman vocabulary. After the extraction of significant features from the vocabulary, classifiers are trained to obtain probabilistic scores. The scores generated resemble the Hindi linguistic power of each word in a two-dimensional space. The visualization and study of lexicons generated show that they have acquired language characteristics such that it has high values for Hindi words and low values for English words.

The language prediction using lexicons proved useful in the task. For Hindi-English language identification subtask the proposed BERT based model generates the best F1-score of 0.84 as compared to other proposed and existing models.

The rationale for incorporating the lexical semantics is essential to enhance the performance of the generic language identification task. The induction of word sense disambiguation leads to a novel framework for the determination of the language of homographs via the use of supervised learning techniques. In the first approach, the Lesk algorithm is customized to meet the classification needs. The findings indicate that small window sizes between one and four are extremely effective, moreover, larger window sizes do not enhance the results. The second approach is based upon classical machine learning models. Finally, empirical results indicate that the Random Forest classifier enhanced with a char N-grams language model outperformed the other suggested classifiers, with an F1-Measure of 97.94 per cent. Moreover, the machine learning methods outperform the thesaurus-based customized Lesk's algorithm which is in line with the state-of-the-art methods.

The transliteration method presented in this work is, subsequently, utilized to construct the multilingual Named Entity Recognition framework. The transliteration model, in particular for Hindi to English, is constructed using a sequence-to-sequence neural network built on top of an encoder-decoder architecture. The transliteration generation results show that the character error rate improves when compared to the basic model. The proposed work when compared to existing Hindi to English, Arabic to English, and Chinese to English transliteration models and it is observed that the proposed model outperforms them all with the character error rate of 20.1 per cent except for Chinese to English, where the character error rate is 16.2 per cent.

For the mixed language named entity recognition task, the Hindi-English named entity annotated corpus is created from Wikipedia with minimal human intervention. To extract the named entities, we use the wikilinks of Indian context Wikipedia category pages. These named entities are in turn utilized to extract article text from Hindi and English Wikipedia pages. Transliteration is employed to transform Devanagari Hindi to Roman Hindi to achieve the mixed lingual corpus. The named entity annotated corpus is subsequently language classified at the word level to generate language labels so that the corpus becomes ready for multitask learning. Consequently, the multilingual multi-task architecture is developed ideally for low-resource situations. Experiments reveal that our approach is capable of effectively transferring additional language information to augment the named entity recognition models. The findings indicate that our performance on the Wikipedia dataset is equivalent to that on the CoNLL'03 dataset. We achieved an F1-Score of 88.46 per cent, which is significantly higher than that of the CoNLL'03 baseline.