Report on “Understanding word contexts and its application towards Natural Language Processing”

Submitted by
Kishore Kashyap
Department of IT, Gauhati University
Under the mentorship of
Dr. Maunendra Sankar Desarkar
Department of CSE, IIT, Hyderabad

Faculty Internship (Under TEQIP 3)
24 June 2019

Abstract: In this report it is tried to summarise the work done during Faculty Internship program from 14th June to 23rd June, 2019 under the mentorship of Dr. Maunendra Sankar Desarkar, Assistant Professor, Department of CSE, IIT, Hyderabad. During the period we explored various Natural Language Processing (NLP) task using Statistical and Deep Learning Techniques. The final work was done to understand word context, word embeddings and their possible applications to various NLP related tasks.

Introduction:

The main purpose of this internship was to explore advanced techniques that falls under the domain of expertise of the chosen Faulty from Indian Institute of Technology, Hyderabad. In addition with this it was also aimed to find out future research collaborations between the two faculty members of the mentee-mentor institutes. During this period we started working on a predefined problem (Microsoft AI Challenge, 2018) with one of the Ph.D scholars of Department of CSE, IIT, Hyderabad. This starting point lead us to concentrate more on the core methods and techniques required for solving those problems.

Initial Study:

In the initial phase, we were exposed to the problem of ranking passages based on a given query. The problem statement was:

“Given a user query and candidate passages corresponding to each, the task is to mark the most relevant passage which contains the answer to the user query.”

In the data, there were 10 passages for each query out of which only one passage was correct. Therefore, only one passage was marked as label 1 and all other passages for that query were marked as label 0. Our goal was to rank the passages by scoring them such that the actual correct passage gets as high score as possible.

Dataset: The organiser of the competition provided three types of data sets to the participants -- i) the labelled train data for training the models and doing validations ii) the unlabelled eval1 data against which we had to submit our predictions during the contest and iii) the unlabelled eval2 data against which final predictions are submitted.

As the competition was already over and was open only for experimentation, we could not able to submit our experimental result for the online scoring. For the next two days we tried to evaluate our results in offline mode.

Works done:

1. For initial two days we were trying to generate our own evaluation set and "ground truth" file (reference file). As there was no clear guidelines for creating the reference file, we tried different formats for the same. At last, after studying the evaluation python file the organising committee supplied, we were able to re-create the reference file. This file was the used to evaluate our model performance in offline mode. This file and our own evaluation dataset was then shared with the Ph.D. scholar.

2. With the locally created reference file, we tested the baseline model which uses Okapi BM25 algorithm [1]. The evaluation metric used was Mean Reciprocal Rank (MRR) [2] which can be defined as:

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

Where, N is the total number of queries and $rank_i$ is the position of the first correct label of each query. When we ran the baseline model, we got a MRR score of 0.433378. We then changed the base algorithm from BM25 to BM25+. Using this algorithm we got MRR value as 0.447298. This was 3.21% increase in performance.

3. After this initial explorations, we tried to implement Deep Learning model for the same ranking task. But, while studying the required models, it was found that understanding of basics of word context android embeddings were necessary. We then solely concentrated on word vector models such as word2vec [3] and GloVe [4].

4. We then created word embedding model for Assamese language [5] using ‘Gensim’ [6]. This word vector was created cleaned corpus from Wikipedia dump of Assamese language. This became the first word vector model for Assamese language. Anyone can download the file from Github repository https://github.com/KashyapKishore/word_embedding4assamese. This word vectors will create a new direction for word context research for Assamese NLP. Possible applications may be development of Context based Spell Checker for Assamese, Prediction of next word in an Assamese language etc.

Conclusion and Future Work:

This work is a preliminary work towards larger collaborative research that we can plan for next level of NLP research in English and Indic Languages.

References:

[1] S. E. Robertson and H. Zaragoza. 2009. “The Probabilistic Relevance Framework: BM25 and Beyond”. Foundations and Trends in Information Retrieval 3(4): 333-389.

[2] E. Voorhees. Proceedings of the 8th text retrieval conference, 1999. TREC-8 Question Answering Track Report.

[3] Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space”. arXiv:1301.3781 [cs.CL]

[4] PJeffrey Pennington, Richard Socher, Christopher D. Manning, “GloVe: Global Vectors for Word Representation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, October 25-29, 2014, Doha, Qatar

[5] https://en.wikipedia.org/wiki/Assamese_language (last accessed on 25/06/2019)

[6] Radim Rehurek, Petr Sojka, “Software Framework for Topic Modelling with Large Corpora” IN PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS, 2010, pg 45–50