Scientific Report

Basant Agarwal
Indian
Norwegian University of Science and Technology (NTNU)
Dr. Heri Ramampiaro

01/11/2016- 31/10/2017
I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

Today, social media websites (such as Twitter, Facebook, Flicker etc.) allowed users to create content and interact with other users. A large number of businesses are based on, or being benefited by analytics of these user-generated content on social media. It has attracted the attention of Natural Language Processing (NLP) and Machine Learning research community to analyze this user-generated content due to its numerous applications and utility in commercial interest. It is very challenging to analyse this contents due to various reasons like short communication, irregular language usage etc. Recently, deep learning approaches have gained significant success to deal such noisy and irregular text in understanding the meaning of the text.

Deep learning approaches have been very successful in recent time in the area of Natural language processing. Deep learning methods are increasing interest in scientific community due to its ability to learn deep semantic representation of text. Deep learning is revolutionizing in natural language technologies since it is offering an effective way of representing text and allows system to learn features automatically from data without the need of explicitly designing them.

We focused on the problem of paraphrase detection using deep learning techniques. Measuring the semantic relatedness of two pieces of text in order to determine if they are paraphrased, is a fundamental problem in language processing tasks. It is very challenging to measure relatedness among pair of sentences when the text is grammatically irregular, noisy, excessive use of jargons, short and there is limited amount of labeled training data. We propose a deep learning approach for paraphrase detection. We propose a hybrid deep neural architecture composed of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), which is further enhanced by a novel word-pair similarity module. Proposed paraphrase detection model is composed of two main components i.e. Pair-wise word similarity matching module and Sentence Modeling. Proposed Pair-wise similarity matching model is used to extract fine-grained similarity information between pair of sentences. We use convolutional neural network to learn the patterns in the semantic correspondence between each pair of words in the two sentences that are intuitively useful for paraphrase identification task. The idea to apply convolutions over pair-wise word to word similarity matrix to extract the important word-word similarity pairs is motivated by the way of applying convolutions over text to obtain the most important part of the sentence. We train hybrid neural network composed of CNN and LSTM network on paired sentences to learn the sentence representation that captures rich semantics.

We also focused on another application of deep learning for NLP i.e.
sentiment analysis for financial text. The objective of this problem was to predict the sentiment of given financial text into the range of $[-1,1]$. Learning good representations of given financial text during training at all is a difficult challenge in natural language processing. Representation of out-of-vocabulary words has been an open question of research. Generally, a unique representation or pre-trained general embeddings is used to represent the out-of-vocabulary words. In this work, we provide a novel method to assign domain-specific vector representation for the out-of-vocabulary words. We develop a novel model that transform the pre-trained general embeddings into the domain-specific embedding space. We also proposed a new method to increase the domain-specific embeddings space, which improved the performance of the fine-grained sentiment analysis by reducing the out-of-vocabulary words. We evaluate the performance on a benchmark dataset of SemEval-2017 shared task on financial sentiment analysis.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

- Basant Agarwal, Heri Ramampiaro, Helge Langseth, Massimiliano Ruocco, “Paraphrase Detection Model for User-Generated Short Text in Twitter using CNN and LSTM Neural Network”, (Under review)
  
  Abstract:
  The ability to detect similar sentences written in natural language, is crucial for several applications such as: text mining, text summarization, plagiarism detection, authorship authentication and question answering as it requires a deep understanding of the semantics of text. Given two sentences, the objective is to detect whether they are semantically identical. Existing paraphrase systems perform well on clean-text corpus but they deliver quite low performance against noisy text as it is quite challenging to detect the paraphrase sentences on user generated noisy and short text on microblogs such as Twitter due to language irregularity. To cope with this challenge, in this paper, we propose a novel deep neural network based approach that rely on coarse-grained sentence modelling based on Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) as well as on a novel fine-grained word-level similarity matching model for paraphrase detection task. Experimental results show that the proposed approach outperforms existing state-of-the-art approaches on user-generated noisy Twitter SemEval 2015 dataset and achieves highly competitive performance demonstrated on the a popular Microsoft Paraphrase Corpus.
III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

- Attended several seminars organised by the Data and Artificial Intelligence (DART) group at Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

**Place:** INRIA, Paris  
**Duration:** 30 August to 5th September 2017  
**Host:** Dr. Éric Villemonte de La Clergerie

During the fellowship, I had the chance to visit ALMAnaCH research group, INRIA Paris, France from 30 August to 5th September 2017. I was hosted by Dr. Éric Villemonte de La Clergerie at INRIA Paris. This research exchange visit was very useful. Researchers of the ALMAnaCH group discussed their work which was related to the our research work. Dr Éric described his research work and demonstrated his recent models based on syntactic deep structures. Dr Éric arranged some meetings with his research group members, where we discussed research activities in more details. I also gave a seminar on my research work and results, and received valuable feedbacks from the research members of the group. I could learn current research being conducted in the ALMnaCH research group and exchange the research ideas with other researchers of the group.

\[29/10/17\]

[Signatures]