Named Entity Recognition (NER)

Shehab Ahmed Sayem
Registration No : 2014331048

Shadat Tonmoy
Registration No : 2014331070

Submitted To :
Dr. Farida Chowdhury
Associate Professor
Dept of CSE,SUST.

Supervised by:
Md. Mahfuzur Rahaman
Assistant Professor
Dept of CSE,SUST.

Introduction:

Day by day human activities are moving towards electronic devices like Smartphones, Tablet, Notebook etc. Monetary transactions are also moving to the same medium. Now People are becoming more comfortable with e-transactions like Mobile Payment, ATM Card, Credit Card, Online Payment etc. These media are replacing the transactions with physical money. At the end of the day, people want to know the summary of their transactions like how much is credited to or debited from his/her account. Moreover, a single person can have multiple accounts and maintaining the summary of these accounts is difficult for him.

Currently there are many applications which can help people to manage and track their transactions. But these applications are dependent on user inputs. More clearly a user needs to give their transaction information like amount, time and date, sector of the transaction. Without these inputs, these applications are not able to give the user any information. But manual user input through any I/O device is tiresome and time consuming which results in a bad user experience.

Minimizing the quantity of user input can improve the efficiency and user experience of any system. To minimize the quantity of user input we can use algorithms that can help the system to learn about the transaction information instead of depending on user input. Machine learning algorithms are the best way to train the system to learn. For training the system we need a source from where we can get transaction related information which will help us to build a training data set. Fortunately, now a day all of the e-transactions provide the transactional information to the user through SMS or E-mail. We will use these SMS to build our dataset and train the system using Machine Learning Algorithm. Here we need to extract information from these SMS. Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. We will use bidirectional LSTM, CNN with CRF to train our system so that it can recognize named entity from the SMS.

Preliminary Literature Review:

So far, we have read some papers including the one which is recognized as the State of the art in the field of NER. One of the papers has proposed a end-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF by Xuezhe et al.(2016) [2]. They introduced a novel neutral network architecture that benefits from both word- and character-level representations automatically, by using combination of bidirectional LSTM, CNN and CRF. They evaluated their system on two data sets for two sequence labeling tasks Penn Treebank WSJ corpus for part-of-speech (POS) tagging and CoNLL 2003 corpus for named entity recognition (NER). They obtained state-of-the-art performance on both datasets - 97.55% accuracy for POS tagging and 91.21% F1 for NER.

In recent years, several different neural network architectures have been proposed and successfully applied to linguistic sequence labeling such as POS tagging, chunking and NER. Among these neural architectures there are the BLSTM-CRF model proposed by Huang et al.(2015)[1], the LSTM CNNs model by Chiu and Nichols (2015)[3] and the BLSTM-CRF by Lample et al. (2016).[4] Huang et al. (2015) used BLSTM for word-level representations and CRF for jointly label decoding.

Chiu and Nichols (2015) proposed a hybrid of BLSTM and CNNs to model both character and word level representations. They evaluated their model on NER and achieved competitive performance. By using two lexicons constructed from publicly available sources, they established a state of the art performance with an F1 score of 91.62 on CoNLL-2003 and 86.28 on OntoNotes, surpassing systems that employ heavy feature engineering, proprietary lexicons, and rich entity linking information.

Another neural architecture employing CNN to model character-level information is the CharWNN architecture has also been applied to Spanish and Portuguese NER (dos Santos et al. 2015)[5] Ling et al. (2015)[6]and Yang et al. (2016)[7] also used BSLTM to compose character embeddings to words representation, which is similar to Lample et al. (2016). Peng and Dredze (2016)[8] Improved NER for Chinese Social Media with Word Segmentation.

Methodology:

We are investigating the research problem from users point of view. There are many methodologies that can be used in NER like CRF (Conditional Random Field), HMM (Hidden Markov Model), Maximum Entropy, BLSTM (Bidirectional Long Short-Term Memory). We are currently studying about these methodologies and we will combine the best ones and try to get a better accuracy.

To obtain our goal we will follow the following steps:

- Literature Study : We will study different paper related to NER to gain and improve our knowledge about different methodologies implemented in their problem domain.
- Data Collection : We will collect real life transaction related SMS as our Data Set.
- Data labeling : We will label our data for different field related to transaction information like amount, date, organization name etc.
- Implementation : We will try to implement the current state of art paper.
- Train and Test : After our implementation is done we will train the system with our dataset to check accuracy in our problem domain.
- Scope of Improvement : We will research to find the scope of improvement

Findings:

Our goal is to improve the current accuracy in the field of Named Entity Recognition. We hope that from our study and implementation we will be able to find a scope where we can improve the accuracy in our problem domain.

We think from our study we can create a method which can be implemented in various systems like Mobile Apps, Web Application etc. Using these real life system people will be benefited. They could track their expenses and also get their buying behavior so that they can manage their incomes more appropriately, which will make an impact in the socioeconomic field.

Tentative chapter outline:

.

- Chapter 1: Introduction
- Chapter 2: Literature Review
- Chapter 3: Research Methodology
  - Compare Different Methodologies
  - Implementation
  - Improvement
- Chapter 4: Results
- Chapter 5: Discussion and Recommendations for Future Research

Time line in GANTT chart

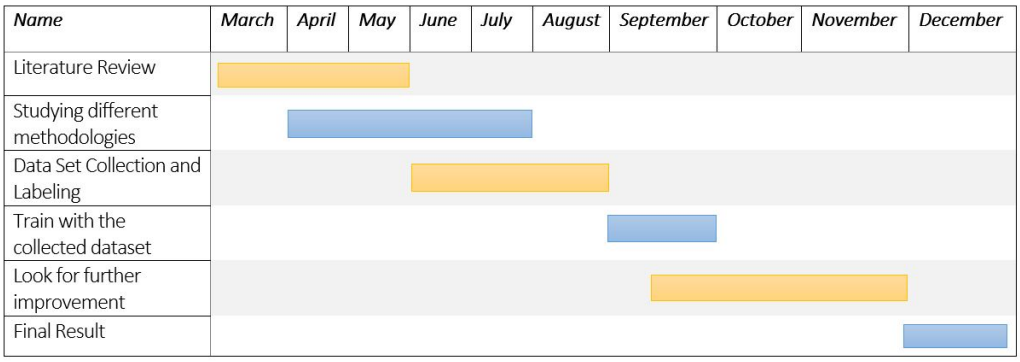

Figure 1: GANTT chart

References

[1] [Huang et al.2015] Zhiheng Huang, Wei Xu, and KaiYu. 2015. Bidirectional lstm-crf models for se-quence tagging.

[2] [Xuezhe et al.2016] Xuezhe Ma, Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF.

[3] [Chiu et al.2015] Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns.

[4] [Lample et al.2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition.

[5] [dos Santos et al.2015] Ccero dos Santos,Victor Guimaraes, RJ Niter oi, and Rio de Janeiro. 2015.Boosting named entity recognition with neural character embeddings.

[6] [Ling et al.2015] Wang Ling, Chris Dyer, Alan WBlack, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation.

[7] [Yang et al.2016] Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multitask cross lingual sequence tagging from scratch.

[8] [Peng and Dredze.2016] Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning.