|  |  |
| --- | --- |
| Project Title: | Named Entity and Rare Word Recognition |
| Supervisor: | Bethany Chan |
| Student 1 (Leader) | Wang Dezhao |
| Student 2 | Wei Jian |
| Student 3 | Liu Wenjin |
| Student 4 |  |
| Student 5 |  |

**Aim**

|  |
| --- |
| The purpose of the project is to develop a NER algorithm based on an existing algorithm[1] trained initially on an existing dataset to tackle the recognition problem for a given specific domain dataset with greater accuracy. The NER algorithm tries to label name entities like a person, organization, location and so on.  [1] Lin, Dekang; Wu, Xiaoyun (2009). [Phrase clustering for discriminative learning](http://www.aclweb.org/anthology/P/P09/P09-1116.pdf) (PDF). Annual Meeting of the [ACL](https://en.wikipedia.org/wiki/Association_for_Computational_Linguistics)and IJCNLP. pp. 1030–1038. |

**Brief Literature Review**

|  |
| --- |
| The Named Entity Recognition(NER) task extracts chunks of text from a sentence and classifies them into pre-defined categories such as the names of persons, locations, and organizations.[1]  NER systems have been created that use linguistic grammar-based techniques as well as statistical models such as machine learning. Hand-crafted grammar-based systems typically achieve better precision, but at the cost of lower recall and months of work by experienced computational linguists. Statistical NER systems typically require a large amount of manually annotated training data.  Because of efficiency considerations, we choose linguistic grammar-based techniques to train the model, and the most common model is LSTM-CRF networks. This network can efficiently use past input features via a LSTM layer and sentence level tag information via a CRF layer. [2][3] With such a CRF layer, we can efficiently use past and future tags to predict the current tag.  [1] Elaine Marsh, Dennis Perzanowski, "MUC-7 Evaluation of IE Technology: Overview of Results", 29 April 1998 PDF  [2] Han, Li-Feng Aaron, Wong, Fai, Chao, Lidia Sam. (2013). Chinese Named Entity Recognition with Conditional Random Fields in the Light of Chinese Characteristics. Proceeding of International Conference of Language Processing and Intelligent Information Systems. M.A. Klopotek et al. (Eds.): IIS 2013, LNCS Vol. 7912, pp. 57–68 [[1]](https://link.springer.com/chapter/10.1007%2F978-3-642-38634-3_8#page-1)  [3] Jenny Rose Finkel; Trond Grenager; Christopher Manning (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling (PDF). 43rd Annual Meeting of the Association for Computational Linguistics. pp. 363–370. |

**Proposed Methodology**

|  |
| --- |
| Based on the recognition result of Stanford CoreNLP, we analyse the mistakes when the existing Name Entity Recognition tools process the given dataset for a specific domain and, from the analysis, determine the research direction to take going forward.   1. We build the Stanford CoreNLP environment and train the Stanford Name Entity Recognizer with the CoNLL-2003 dataset, which is a public and common dataset. Then, we use the trained model to process the dataset X, which is the given dataset of a specific domain. 2. From an analysis previously done on the recognition results using the Stanford Name Entity Recognizer on dataset X, it was found that wrong input format (e.g. “RMB51 ,453” will be recognized as “RMB51” and “453”), long organization name (e.g. “The Institute of Chartered Secretaries and Administrators” will be recognized as “The Institute of Chartered Secretaries”) and mixed Chinese text (e.g. the Chinese name entity will not be recognized in a long English sentence) would result in a wrong recognition. 3. Based on the existing NER algorithm and dataset, we plan to improve the performance of the recognition result in 2 ways: First, we can adjust the training dataset to make it more suitable for the NER algorithm for the specific domain dataset recognition. Second, we can combine the classic LSTM-CRF algorithm with the existing algorithm to improve the performance.   Adjust the training dataset to improve the algorithm performance.   1. We clean the existing specific domain dataset to repair wrongly-formatted data and remove mixed Chinese text. 2. We label the existing specific domain dataset using the IOBES tagging scheme. 3. To find out if the dataset improves the performance of the algorithm, we train the algorithms with both the CoNLL-2003 dataset and the specific domain dataset. In addition, to get better performance, we can adjust the training dataset’s format and scale.   Change the classic LSTM-CRF algorithm   1. Compared with the Stanford CoreNLP Name Entity Recognizer which is based on the CRF Classifier, we build the LSTM-CRF model to get better performance when processing the sequence input data. 2. We adjust the structure of the LSTM-CRF model by combining the existing algorithm to improve the recognition performance.   Rare word detection   1. To detect the rare word, we decide to recognize it with word frequency. We will establish a glossary with word frequency and set a suitable threshold. 2. After classification by the LSTM-CRF model, the rare word will be labeled with PER, ORG, LOC, and MISC. [1] Then, we calculate the word frequency of the words labeled with MISC and classify the data as Rare Word or Non-Rare Word based on the threshold.   [1] Tjong Kim Sang, Erik F.; De Meulder, Fien (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. CoNLL. |

**Milestones**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Tasks*** | | ***Estimated completion time*** | ***Estimated number of***  ***learning hours*** |
| 1 | Building CoreNLP environment | Before May 20th | 60 |
| 2 | Finding wrong\_labeled cases | Before May 20th | 60 |
| 3 | Data cleaning | Before May 30th | 60 |
| 4 | IBOES labeling | Before May 30th | 450 |
| 5 | Data training | Before June 30th | 150 |
| 6 | Compare with existing models | Before July 10th | 60 |
| 7 | Adjust models | Before July 20th | 60 |
| 8 |  |  |  |
| 9 |  |  |  |
| 10 |  |  |  |
|  |  |  | ***Total: 900*** |

**Deliverables**

|  |  |
| --- | --- |
| ***Items*** | |
| 1 | Text data which has been preprocessed and labeled |
| 2 | An ner recognition software to recognize name entities and rare words. |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |