Chapter 5

Conclusion and Future Directions

In natural language texts, identification and classification of proper nouns is a challenging but useful task, known as NER. While for many languages, it continues to be an active area of research, for many other languages, this work has barely started. Our work involves a computational model that performs NER in text in a resource-poor Indo-Aryan language, namely Assamese, which has received little attention in computational linguistic research. NER is difficult as ambiguity exists among the different classes of NEs such as *Washington* can be a person name as well as a location name. Indian languages also face the problem of ambiguity. We have identified various issues in NER in Indian languages and Assamese in particular. Though Assamese is a relatively free-word order language, the predominant word order is SOV. Assamese language does not have the concept of capitalization which is an important clue to identify the NEs. Our work involves the successful implementation of the following. We implement a method of suffix stripping using rules for

our morphologically-rich, agglutinating, and free-word order language. The main aim of this step is the identification of the root words which are location NEs. We obtained an accuracy of 89%, but found that it produce serrors when the last character of a word matches a suffix in the suffix list. To remove such errors, a gazetteer or dictionary of location names was introduced. Though preparing a gazetteer list manually is a time-consuming process, we collected some amount of data for our experimental work. We have also derived some handcrafted rules for different classes of NER and found that handcrafted rules work well provided the rules are carefully prepared. Our hand coded rules result in an accuracy of 70-75%. We have also implemented NER using a gazetteer list which results in an accuracy of 75%-85% and have seen that the performance of the system increases as the size of the gazetteer list increases. In addition to the rule-based approach, we also implemented two ML approaches namely CRFs and HMM which are existing approaches, but the work is for a new language where an annotated corpus was not available. We have prepared a manually-tagged corpus for evaluating the performance of the approach. We used different language dependent and independent features to identify the NEs. The CRF-based NER approach gives an accuracy between 70% and 85% whereas HMM gives an accuracy of 75%-85%. We see that each of the approaches has its own merits and demerits. We have analyzed the errors encountered by the different approaches. We have handled the problems by using a hybrid approach which is a combination of the rule-based, gazetteer-based and ML approaches. The proposed hybrid system has achieved an overall improvement in Assamese NER performance. It is capable of recognizing four different types of NEs including person, location, organization, and miscellaneous which includes the date, time, and year. Our experimental results show that the hybrid approach outperforms the pure rule-

based approach, gazetteer-based approach, and the pure ML-based approach, with an F-measure of 80%-90%. Our results compare well with existing work on other Indian languages though the comparison is not absolute as the sizes and the qualities of the data vary for different authors. For future work, it would be interesting to explore additional rules and features for Assamese and also to implement them in other especially of Northeast India, using our hybrid approach.