5.1 Present Work

As compared to English a lot is yet to be done in NER for Telugu. Various approaches available for NER have been discussed this far. In this thesis, a platform independent rule based named entity recognizer and classifier has been built for Telugu. The system is built using newspaper and Telugu-wiki corpus. Several experiments have been conducted and results explained as well. Since Telugu language suffers from the lack of labelled data, this system can be used to produce training data which can be used for machine learning techniques or hybrid techniques and can further produce more reliable annotated data. Though this is a rule based system, various domains of data such as history, scientific documents, novels, stories, literature, articles etc have been looked at. We can say this system can work across all these domains with satisfactory results. An attempt is made to identify multi-word named entity which helps in identifying the multi word organization names which is otherwise not possible with token level named entity recognition. This system may be executed on entire sentences database to get a named entity tagged data with precision in the range of 79%-94%.

There are some observations and recommendations while building a named entity recognition and classification system for Telugu.

- Common and proper noun disambiguation can be made more accurate by efficient verb (action verbs which are specific to living beings) classification.
A single verb root can lead to the formation of a few hundred thousand word forms. If we can identify and classify verb forms in more detail we would be able to disambiguate common noun and proper noun more efficiently.

- If we can identify domain based loanwords and loanwords which we use in our daily life but do not generate NEs, performance of the system can be improved greatly.

### 5.2 Future Enhancements

This work can be extended to correct identification of named entity boundaries and to identify nested named entities. The output of this work can be used as training data for any machine learning technique since machine learning techniques require large amount of tagged data to get very good results.

1. Domain specific terms, brand names, and so on are not handled.
2. This approach can divide the other categories into proper meaningful sub categories.
3. Need separate features for identification domain specific terms.
4. In Feature uses machine learning algorithms for disambiguation.