PROJECT REPORT

Judicial NER

Project description:

The project involves identifying and classifying the named entities in Judicial documents(Hindi). The reason normal NER softwares don’t show a good accuracy is that Judicial documents have specific formats and it’s difficult to extract out information from the syntax and dependency analysis.

So, this project uses a rule based approach to identify the Named entities in Judicial documents.

Literature review:

- Named Entity Recognition and Resolution in Legal Text by Christopher Dozier, Ravikumar Kondadadi and Marc Light
- Towards Deep Learning in Hindi NER: An approach to tackle the Labelled Data Scarcity by Ameya Prabhu(only for looking into other approaches)

Methodology used:

We use a rule based approach to identify and label the named entities. The scope of each type of named entity is calculated first using a careful analysis of the test data and then based on the surroundings of the keyword that marks that particular named entity, we assign labels to each word. We look at the tags and the dependencies(if needed) of the neighbouring words to check for the scope.

Work/experiments done:

1. Manually annotated the data to check the scope and frequency of each type of named entity(used Bash scripts to find frequency).
2. Built up the corresponding dictionaries by scraping data using python.
3. Experimented with binary and utf-encoding to tokenize and convert data into required format.
4. Experimented with LTRC shallow parser to retrieve the tags.
5. Experimented with LTRC dependency parser.
6. Used the isc-tokenizer, tagger and parser.
7. Constructed rules based on neighbouring words, their tags and dependencies.

8. Rechecked rules to check for rule clashing and changed the order of the rules accordingly.
9. Used dependencies and unicode rules to accomplish date/time recognition.
10. Experimented with other python libraries to include features like progress bars, pie charts etc.

Analysis:

The software worked really well and successfully showed the correct tags for the named entities in most of the cases. The accuracy depends on some rules clashing if it wasn't correctly captured in the dictionary files made from the testing data and scraping.
We tested 2 files, achieved 91% accuracy for one and 93.6% for the other.

The reason for such a high accuracy is that the data occurs in very predictable format and as we make rules based on tags and dependencies which is general in most of the cases, we correctly identify the labels.

We were even able to identify errors in the manually annotated file. eg:('vidhaan sabha' wasn't marked as a named entity while 'lok sabha' was)

Conclusion:

A rule based approach is apt for this task and doesn't require really complex rules and parsing techniques as the data occurs in predictable formats only. Better dictionaries can be created if accuracy is seen to be dropping.Overall, the code successfully works especially for large files. The efficiency can be improved though by used machine learning techniques and training the system on the data.

Future Work:

- Making better dictionaries using more data.
- Adopting a Brill's tagger based approach to re-adjust the rules and accomplish correct labelling.
- Date/time recognition rules can be made better by looking at more environments.
- Using machine learning and other data structures than simple lists to improve the efficiency.