CLG 452: Linguistic Data 2 **(Due: 27/04/20)**

Course Project Report

Instructor: Dr. Parameshwari *Name:* YOUR NAME HERE, *Roll Number:* ROLL_NUMBER

Instructions: Read all the instructions below carefully before you start working on the report, and before you make a submission.

- Please typeset your submissions in LaTeX. Use the template provided for your answers. Please include your name and Roll number with submission.
- This report is due on 27 April at 11:55 PM.
- No extensions will be given for the submission under any circumstances.
- Submissions via any other method other than Moodle submission will be deemed invalid.
- Plagiarism of any sort will not be tolerated. Strict action will be taken for those caught in plagiarism.

1 Tags

Discuss the different kind of tags in **both** UD and AnnCorra occurring in your chosen Indian language.

1.1 UD

You can do this either in tabular form or in a list form. An example of a table is:

Tag	*Examples*	*Linguistic Cues*
k1	• Multiple Example 1 • Multiple Example 2	• Multiple Cue 1 • Multiple Cue 2
k2	Single Examples	• Multiple Cue 1 • Multiple Cue 2
k3	• Multiple Example 1 • Multiple Example 2	Single Cue 1
k4	Single Example 1	Single Cue 1

Table 1: Discussion on UD tags

If you do not want to write it in a tabular form, write a list of all the tags with their examples and linguistic cues.

(a) k1: example1, example 2. Cue 1, Cue 2.

(b) ...

1.2 AnnCorra

Do the same for AnnCorra.

2 Linguistic Challenges with Annotation

(a) Differential Object marking
(b) Non-Nominative Subjects
(c) Complex Predicates
(d) Non-finite clauses: Conditional, Concessive, Relative, participial clauses
(e) Ambiguity (Coordination, Attachment)
(f) Ellipsis
(g) Non-projectivity Ex 1: I saw a man yesterday who was singing Ex 2: A hearing is scheduled on the issue today Ex 3: To his wife, John gave a fantastic gift Ex 4: Which house, John bought?
(h) Particles

3 Tag Statistics

3.1 Tag and Markers

For each tag, indicate the markers used to identify that tag and the number of tokens identified by each marker. Example:

There are a total of 100 k1 tags. k1 comes with the markers $0, -ne$. 0 marker is responsible for 56 of the cases and $-ne$ is responsible for the remaining 44.

3.2 Markers and Tag

For each marker, indicate the types of tags given to it and the number of cases for each tag. Example:

There are a total of 100 tokens with $-ne$ marker. These tokens are marked with $k1, k2$. $k1$ marker is responsible for 32 of the cases and $k2$ is responsible for the remaining 68.

3.3 N-gram of tags

Include statistics about the frequency of n-gram of tags. Take n in the range [2,4].

NOTE: You have to do the above 3 exercises for both UD and AnnCorra. This analysis has to be done on the training set given for both the types of tagging in Assignment 4.

4 Error Analysis for automatic tagging

In this section, describe the errors in the output of your model trained in Assignment 4 on the test data. Give possible solutions, if any, to mitigate these errors in the future.

5 Discussion

5.1 Comparison of UD and AnnCorra

5.2 Need for new tags

5.2.1 Intra-Chunk

propose any new intra-chunk tags you can think of with examples

5.2.2 Inter-Chunk

propose any new inter-chunk tags you can think of with examples.