Computational Linguistics 1
Assignment 1: Rule Based POS Tagging

Deadline: 11:30 am, 3rd August

The assignment consists of two tasks, with detailed subtasks as mentioned below.

Task 1: Rule Based POS Tagging for English

Development Data (Task 1):

1. *She has been absent since last Wednesday.*
2. *It doesn't matter what excuse he gives me, I can't forgive him.*
3. *I canceled my appointment because of urgent business.*
4. *What do you do in Japan?*
5. *The Handmaid's Tale is an awesome piece of dystopian fiction.*
6. *OK. Now what?*
7. *I was laughed at by everyone.*
8. *There were people everywhere, covering the roads along the route from the BJP headquarters to the Smriti Sthal from side to side, with security personnel maintaining strict vigil to ensure that nothing goes wrong.*

Testing Data (Task 1):

During a visit to the Cleveland Indians, Beane meets Peter Brand, a young Yale economics graduate with radical ideas about how to assess player value. Beane tests Brand's theory by asking whether he would have drafted Beane out of high school. Though scouts considered Beane hugely promising, his career in the Major Leagues was disappointing. Brand admits that, based on his method of assessing player value, he would not have drafted him until the ninth round. Impressed, Beane hires Brand as his assistant manager.

Task 1 consists of 4 subtasks:

1. Annotate (POS-Tag) *development data* using the data provided. In this step you need to list all the possible POS Tags that could be assigned to each word in the data.
2. Disambiguation: Disambiguate the POS tags for each word manually, depending on context, lexical properties etc., and make rules in the process. You can go through the paper on Brill's Tagger[1] (and CLAWS Tagger[2]) to learn about the general format of the rules. You may also write the rules in any other legible form you like.

[1] "A Simple Rule-Based Part of Speech Tagger - Association for ...."
http://www.aclweb.org/anthology/A92-1021.
[2] "The CLAS7S word-tagging - ucrel."
http://ucrel.lancs.ac.uk/papers/ClawsWordTaggingSystemRG87.pdf.

3. Tag the *testing data* as done in step 1 and use these rules for disambiguation.
4. Analyze the annotation of the *testing data*, and in case of discrepancies, analyze and reformalize the rules.

Task 2: Rule Based POS Tagging for an Indian language

Task 2 involves doing Task 1 with an Indian language of your own. You can choose one of these 6 languages -- Marathi, Bengali, Hindi, Telugu, Malayalam, Urdu, Kannada.

Task 2 subtasks::

1. Collect *development* and *testing data* in your language, each consisting of 50-80 words.
2. Annotate (POS-Tag) *development data* using the annotated data provided. In this step you need to list all the possible POS Tags that could be assigned to each word in the data. For words missing in the data, do it manually.
3. Follow steps 2 to 4 from Task 1.

Note:

1. Use Penn Treebank tagset[3] for doing the annotations for Task 1.
2. Use the BIS tagset[4] for doing annotations for Task 2.
3. Please make sure you elaborate on your rules and the logic behind them. If some problems remain unresolved, mention them as well.
4. The assignment can be submitted online or in hand to the faculty or TA, in handwritten or printed form.
5. This assignment is graded. Late submission would lead to an incremental deduction of grade points for this assignment.

[3] "Penn Treebank P.O.S. Tags." https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.
[4] "POS - TDIL-DC." http://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf.