Computational Linguistics - 1, Assignment 1

Zubair Abid, 20171076

The assignment works on two language data/sets, English and Bengali. The task is to assign (correctly) appropriate POS Tags for different words in either language based on rules.

1 English POS Tagging

Tagset used: the Brown Corpus tagset (As given in assignment resources)

1.1 Method used for tagging

For the given ruleset, each word was initially tagged with the first given tag, and then corrected manually if needed. Along the way rules were constructed to automatically tag further data. The general procedure followed is Brill's method.

1.2 Development data with tags

The correct tag for each word is highlighted in bold text

1. *She has been absent since last Wednesday.*
She/**PPS** has/**HVC** been/**BEN** absent/JJ/**VB** since/**IN**/CS/RB last/**AP**/VB/NN/RB Wednesday/**NR** ./**.**/-HL

2. *It doesn't matter what excuse he gives me, I can't forgive him.*
It/**PPS**/PPO doesn't/**DOZ*** matter/NN/**VB** what/**WDT** excuse/**NN**/VB he/**PPS** gives/**VBZ** me/**PPO** ,/**-HL** I/**PPSS** can't/**MD*** forgive/**VB** him/**PPO** ./**.**/-HL

3. *I canceled my appointment because of urgent business.*
I/**PPSS** canceled/**VBD**/VBN my/**PP$** appointment/**NN** because/**CS**/RB of/**IN** urgent/**JJ** business/**NN** ./**.**/-HL

4. *What do you do in Japan?*
What/**WDT** do/**DO** you/**PPSS**/PPO do/**DO** in/**IN**/RP Japan/**NP** ?/**.**/-HL/.-TL/.-NC

5. *The Handmaid's Tale is an awesome piece of dystopian fiction.*
The/**AT** Handmaid's/**NP$** Tale/**NN** is/**BEZ** an/**AT**/CC awesome/**JJ** piece/**NN**/VB of/**IN** dystopian/**JJ** fiction/**NN** ./**.**/-HL

6. *OK. Now what?*
OK/JJ/**RB** ./**.**/-HL Now/**RB** what/**WDT** ?/**.**/.-HL/.-TL/.-NC

7. *I was laughed at by everyone.*
I/**PPSS** was/**BEDZ** laughed/VBD/**VBN** at/**IN** by/**IN**/RB everyone/**PN** ./**.**

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.*
There/**EX**/RB were/**BED** people/**NNS**/VB everywhere/**RB**/NN ,/**-HL** covering/**VBG**/NN the/**AT** roads/**NNS** along/**IN**/RB the/**AT** route/**NN** from/**IN** the/**AT** BJP/**NP** headquarters/**NN**/NNS to/**TO**/IN/NPS/QL the/**AT** Smriti/**NP** Sthal/**NP** from/**IN** side/**NN**/JJ to/**TO**/IN/NPS/QL side/**NN**/JJ ,/**-HL** with/**IN**/RB security/**NN** personnel/**NNS** maintaining/**VBG** strict/**JJ** vigil/**NN** to/**TO** ensure/**VB** that/**IN** nothing/**PN** goes/**VBZ** wrong/**JJ**/RB/NN ./**.**/-HL

1.3 Disambiguation Rules

Some disambiguation rules (based entirely on the given development dataset and tags) are:
1. If there is a JJ after HV*/BE*, change to VB
2. NN after P*/N* and VB* –> VB
3. Single word sentence JJ –> RB
4. If a word ends with -ed and has VBN in options and follows BE*/HV*, tag as VBN
5. VB* after J* is NN*
6. Untagged words default to NP, NP$ with apostrophe

1.4 Tagged Test Data

During/IN a/AT visitVB to/TO the/AT Cleveland/NP Indians/NPS ,/-HL Beane/NP meets/VBZ Peter/NP Brand/NP ,/-HL a/AT young/JJ Yale/NP economics/NN graduate/NN with/IN radical/JJ ideas/NNS about/IN how/WRB to/TO assess/VB player/NN value/NN ./. Beane/NP tests/VBZ Brand's/NP$ theory/NN by/RB asking/VBG whether/CS he/PPS would/MD have/HV drafted/VBN Beane/NP out/IN of/IN high/JJ school/NN ./. Though/CS scouts/NNS considered/VBD Beane/NP hugely/JJ promising/VBG ,/-HL his/PP$ career/NN in/IN the/AT Major/NP Leagues/NP was/BEDZ disappointing/JJ ./. Brand/NP admits/VBZ that/CS ,/-HL based/VBN on/IN his/PP$ method/NN of/IN assessing/VBG player/NN value/NN ,/-HL he/PPS would/MD not/* have/HV drafted/VBN him/PPO until/IN the/AT ninth/OD round/NN ./. Impressed/NP ,/-HL Beane/NP hires/VBZ Brand/NP as/CS his/PP$ assistant/NN manager/NN ./.

1.5 Additions to disambiguation rules

Some of the tags assigned by the previous ruleset turned out to be faulty. Based on the observations, some new rules were added to tag the data better:
7. QL followed by a TO tag is WRB
8. RP followed by a IN* is IN
9. CS at a non initial position is IN

2 Bengali POS Tagging

Tagset used: BIS Tagset

2.1 Method used for tagging

A method similar to the one used for English tagging. Since the tags were unstructured, initially all tags are applied to any word and then disambiguation is done.

2.2 Development data with tags

1. আমার খুব খিদে পেয়েছে
আমার/PR_ PRP খুব/RP_ INTF খিদে/N_ NN পেয়েছে/V_ VM

2. ভেজা মেঝেতে দৌড়োতে হয় না বাবা
ভেজা/V_ VM_ VNG মেঝেতে/N_ NN দৌড়োতে/V_ VM হয়/V_ VM_ VF/V_ AUX না/RP_ NEG/RP_ RPD বাবা/N_ NN

3. পরীক্ষায় ভালো না করলে ঘর থেকে বার করে দেব
পরীক্ষায়/N_ NN ভালো/RB, JJ না/RP_ NEG/RP_ RPD করলে/V_ VM_ VNF ঘর/N_ NN থেকে/PSP বার/N_ NN, RB করে/V_ AUX/PSP/V_ VM/N_ NN/V_ VM_ VNF দেব/V_ VM_ VNF/N_ NNP

4. খাবারে বিরিয়ানি না দিলে বিয়েতে কেউ আসবে না
খাবারে/ বিরিয়ানি/N_ NNP না/RP_ NEG/RP_ RPD দিলে/V_ VM_ VNF/V_ AUX বিয়েতে/N_ NN কেউ/PR_ PRI আসবে/V_ VM_ VNF/V_ AUX না/RP_ NEG/RP_ RPD

2.3 Disambiguation Rules

Some disambiguation rules (based entirely on the given development dataset and tags) are:
1. V_ * without immediately preceding V_ * is V_ VM_ *
2. V_ * after a V_ VM is V_ AUX IF option available
3. In preceding, if no option available flip tags
4. না immediately before or after a V_ is RP_ NEG
5. RB,JJ/N_ * immediately before RP_ NEG/V_ * is RB
6. PSP, * immediately after N_ * is PSP

2.4 Tagged Test Data

1. আমার আজ দুপুর থেকেই বেশ ঘুম পেয়েছে
আমার/PR_ PRP আজ/N_ NN দুপুর/N_ NN থেকেই/PSP/V_ VM_ VINF বেশ/RP_ INTF/QT_ QTF ঘুম/N_ NN পেয়েছে/V_ VM

2. এই কাজ জিনিষটা আমার একদম ভাল লাগে না
এই/DM_ DMD কাজ/N_ NN জিনিষটা/N_ NN আমার/PR_ PRP একদম/RB ভাল/JJ লাগে/V_ VM_ VF না/RP_ NEG

3. পরীক্ষাটা একদম জঘন্ন হতে চলছে
পরীক্ষাটা/N_ NN একদম/RB জঘন্ন/JJ হতে/V_ VM_ VNF/V_ VM/V_ AUX/PSP চলছে/V_ VM_ VF

4. আমার বাড়িতে একদিন আয়, বেশ মজা হবে
আমার/PR_ PRP বাড়িতে/N_ NN একদিন/N_ NN আয়/N_ NN ,/, বেশ/RP_ INTF/QT_ QTF মজা/N_ NN হবে/V_ VM_ VF/V_ AUX

2.5 Additions to disambiguation rules

Not many discrepancies were found with existing rules, but a new rule was added.
1. RP_ INTF, QT_ QTF before a N_ * is RP_ INTF