**Natural Language Processing**

**Homework 2a**

By Sharad Tanwar

UIN: 673426160

**Question 1**

1.

Sentence 1:

When a guy and a doll get to taking peeks back and forth at each other, why there you are indeed.

…I am sitting in Mindy’s restaurant putting on the gefillte fish, which is a dish I am very fond of,…  
  
**SNOW POS:**

NN … PRP I  VBP am  VBG sitting  IN in  NNP Mindy NN ’ VBZ s  NN restaurant  VBG putting  IN on  DT the  NN gefillte  NN fish , ,  WDT which  VBZ is  DT a  NN dish  PRP I  VBP am  RB very  JJ fond  IN of , , NN …

**STANFORD POS:**

.../: I/PRP am/VBP sitting/VBG in/IN Mindy/NNP 's/POS restaurant/NN putting/VBG on/IN the/DT gefillte/JJ fish/NN ,/, which/WDT is/VBZ a/DT dish/NN I/PRP am/VBP very/RB fond/JJ of/IN ,/, .../.

**Differences**

They do not agree on all tags. Below are the differences:

|  |  |  |
| --- | --- | --- |
| No | SNOW POS | STANFORD POS |
| 1 | NN- … | : - ... |
| 2 | NN- ‘ , VBZ - s | POS – ‘s |
| 3 | NN - gefillte | JJ - gefillte |
| 4 | NN - … | . - … |
|  |  |  |

**Mistakes**

1. Gefillte – It is a common dish and therefore should be tagged as a noun. One of the tagger mistook it for an adjective because fish is a common noun and it can be preceded by an adjective, which describes the noun.
2. … - The punctuation at the end of the statement can be interpreted both as an end of sentence marker or a mid sentence punctuation as we have a comma right at the end of the statement given. This confuses the parser.
3. … - The punctuation in the beginning of the sentence is supposed to be a mid sentence punctuation because it looks like the sentence begins mid sentence as we never have this punctuation when a sentence is starting afresh. This confuses the parser as one tries to interpret it as a new sentence and one thinks of it as a mid sentence.
4. ‘s – The correction tag for this is POS or possessive ending. The SNOW tagger broke down the tag into two different tags and therefore confused it.

2.

Sentence 2:

Wearn a suit tomorrow for da interview ;+) dats how serious da money is

**SNOW POS:**

NNP/Wearn  DT/a  NN/suit  NN/tomorrow  IN/for  NNP/da  NN/interview  : ; NN/+ -RRB- )  NN/dats  WRB/how  JJ/serious  NNP/da  NN/money  VBZ/is

**STANFORD POS :**

Wearn/VB a/DT suit/NN tomorrow/NN for/IN da/NN interview/NN ;/: +/VBN -RRB-/-RRB- dats/VBZ how/WRB serious/JJ da/NN money/NN is/VBZ

**Differences**

They do not agree on all tags. Below are the differences :

|  |  |  |
| --- | --- | --- |
| No | SNOW POS | STANFORD POS |
| 1 | NNP- Wearn | VB – Wearn |
| 2 | NNP – da | NN - da |
| 3 | VBZ – dats | NN - dats |
| 4 | NN - + | VBN - + |
| 5 | -RRB- ) | -/-RRB- |

**Mistakes**

1. da – da is a preposition in Italian i.e. from used in names like ‘Vasco Da Gama’ . But it is also in the Russian language and is an adverb in that. The word was probably meant to be a determiner ‘The’ but the spelling mistake makes it tough to understand.
2. Dats- The word is another spelling mistake and was probably meant to be ‘That’s’ which is a WDT or who-pronoun. The parser cannot make this out and hence gives a wrong output.
3. Wearn – This is another case where the word does not exist in English Lexicon. It probably meant wear and the Stanford POS still gave the correct answer as a verb.
4. + - It was supposed to be tagged as SYM but was given incorrectly by both taggers. This is because of the fact that the user was trying to create an emoticon instead of just a symbol and there was no space in between.
5. The Stanford POS removed the ‘)’ tag as it can sometimes cause issues with parsing. Both taggers tagged it as RRB instead of leaving it as it is.

3.

Sentence 3:

The more episodes you have, the more likely you are to have more episodes, and in general the episodes, over a lifetime, get worse and closer together.

**SNOW POS:**

DT/The JJR/more NNS/episodes PRP/you  VBP/have , /,  DT/the  RBR/more  JJ/ likely  PRP/you  VBP/are  TO/to  VB/have  JJR/more  NNS/episodes ,/,  CC/and IN/in  JJ/general  DT/the  NNS/episodes ,/,  IN/over  DT/a  NN/lifetime ,/,   VB/get  JJR/worse  CC/and  RBR/closer  RB/together . /.

**STANFORD POS:**

The/DT more/RBR episodes/JJ you/PRP have/VBP ,/, the/DT more/RBR likely/JJ you/PRP are/VBP to/TO have/VB more/JJR episodes/NNS ,/, and/CC in/IN general/NN the/DT episodes/NNS ,/, over/IN a/DT

lifetime/NN ,/, get/VBP worse/RBR and/CC closer/RBR together/RB ./.

**Differences**

They do not agree on all tags. Below are the differences:

|  |  |  |
| --- | --- | --- |
| No | SNOW POS | STANFORD POS |
| 1 | JJR- more | RBR – more |
| 2 | NNS- episodes | JJ – episodes |
| 3 | VB – get | VBP - get |
| 4 | JJR – worse | RBR - worse |
| 5 | JJ- general | NN-general |

**Mistakes**

1. The Stanford tagger tags more as an adverb comparative which is a valid form for ‘more’ but in this case it has to be adjective comparative as we have a higher degree of episodes being described by the word ‘more’.
2. The word episodes is given as ‘JJ’ by the Stanford parser which is incorrect and is correctly tagged as NNS or Noun Plural. The word episodes is the noun in this case and more describes it.
3. SNOW POS gets it correct for the word ‘get’ and tags it as a verb.
4. The word ‘worse’ is tagged as RBR in Stanford Parser which is correct as it is associated with ‘get’ which is a verb. JJR is wrong in this case.
5. The word ‘general’ is again tagged correctly by Stanford POS as a noun as in this case it cannot be an adjective.

**Question 2:**

**Training Corpus: (HMM.txt)**

1. We /PRP arrived /VBD in /IN Calais /NNP France /NNP around /IN one /CD in /IN the /DT afternoon /NN

2. Timing /NN and /CC mental /JJ exhaustion /NN required /VBD an /DT interlude /NN

3. We /PRP needed /VBD to /TO give /VB Odin /JJ time /NN to /TO find /VB Manannan /NNP Mac /NNP Lir /NNP

4. We /PRP had /VBD a /DT comfortable /JJ lead /NN on /IN the /DT huntresses /NNS

5. We /PRP snuck /VBP into /IN a /DT clothier /NN to /TO grab /VB some /DT duds /NNS

6. We /PRP also /RB lifted /VBD six /CD leather /NN belts /NNS for /IN later /JJ use /NN

7. I /PRP took /VBD note /NN of /IN the /DT name /NN to /TO make /VB sure /JJ the /DT establishment /NN got /VBD paid /VBN later /RB for /IN what /WP we /PRP took /VBD

8. Not /RB trusting /VBG ourselves /PRP to /TO nap /VB briefly /RB we /PRP chose /VBD to /TO remain /VB awake /RB and /CC explore /VB the /DT city /NN for /IN a /DT few /JJ hours /NNS

9. I /PRP kept /VBD my /PRP$ eyes /NNS peeled /VBN for /IN possible /JJ enemies /NNS but /CC tried /VBN to /TO conceal /VB my /PRP$ paranoia /NN

10. We /PRP all /DT studiously /RB avoided /VBD talking /VBG of /IN the /DT immediate /JJ past /NN or /CC the /DT future /NN

11. We /PRP were /VBD both /DT desperate /JJ I /PRP think /VBP for /IN a /DT thin /JJ slice /NN of /IN normalcy /NN

12. I /PRP taught /VBD Granuaile /VB a /DT few /JJ French /JJ words /NNS

13. We /PRP pulled /VBD off /RP another /DT meat /NN heist /NN in /IN a /DT café /NN

14. The /DT food /NN was /VBD rather /RB pedestrian /JJ in /IN Oberon /NNP 's /POS view /NN compared /VBN to /TO what /WP he /PRP 'd /VBD had /VBN in /IN Poland /NNP

15. It /PRP took /VBD the /DT edge /NN off /IN our /PRP$ hunger /NN until /IN we /PRP could /MD enjoy /VB something /NN later /RB however /RB

**Question 2a**

1. Tag Bigram Probability in **tagBigram.txt**
2. Lexical Likelihood Probability in **lexicalProbability.txt**

**Question 2b**

New Sentence:

**We needed to give Odin time to nap briefly**

The most likely tag sequence:

We <PRP> needed <VBD> to <TO> give <VB> Odin <JJ> time <NN> to <TO> nap <VB>

briefly <RB>

**Probability:**

Tag Bigram Probability:

P(PRP|/start) \* P(VBD|PRP) \* P(TO|VBD) \* P(VB|TO) \* P(JJ|VB) \* P(NN|JJ) \* P(TO|NN) \* P(VB|TO) \* P(RB|VB) \* P(/end|RB)

Lexical Likelihood Probability:

P(We|PRP) \* P(needed|VBD) \* P(to|TO) \* P(give|VB) \* P(Odin|JJ) \* P(time|NN) \* P(to|To) \* P(nap|VB) \* P(briefly|RB)

Total Probability = 0.8 \* 0.66 \* 0.117 \* 0.875 \* 0.2 \* 0.461 \* 0.115 \* 0.875 \* 0.2 \* 0.111

0.611 \* 0.058 \* 1.0 \* 0.1 \* 0.076 \* 0.038 \* 1 \* 0.1 \* 0.111

= 0.00000000000013