Morphological disambiguation

Improve perceptron tagger

1. First download the code:

   $ git clone https://github.com/ftyers/conllu-perceptron-tagger.git

2. Then download some data, for purpose of this assignment we will be using the corpus for swedish language:

   $ git clone https://github.com/UniversalDependencies/UD_Swedish-LinES.git

3. Then download the CoNLL shared task 2017 official evaluation script and unzip it:

   $ wget http://universaldependencies.org/conll17/eval.zip

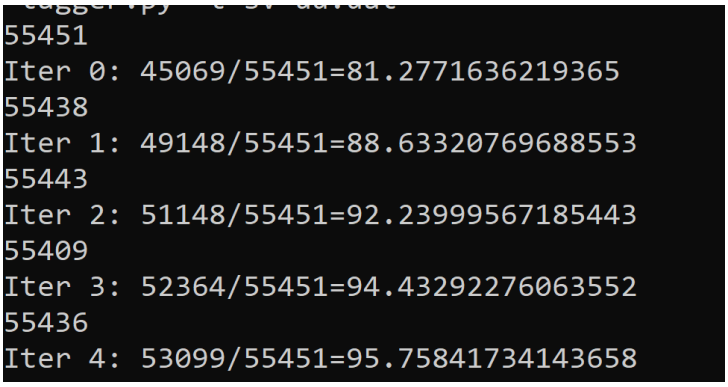
   $ unzip eval.zip

4. Finally enter the directory of the perceptron tagger:

   $ cd conllu-perceptron-tagger

5. You can train the tagger using the following command:

   $ cat ../UD_Swedish-LinES/sv_lines-ud-train.conllu | python3 tagger.py -t sv-ud.dat



6. Now you can run the tagger:
   $ cat ../UD_Swedish-LinES/sv_lines-ud-test.conllu | python3 tagger.py sv-ud.dat > sv-ud-test.out

7. And evaluate(before changing features of tagger.py):

$ python3 ../evaluation_script/conll17_ud_eval.py –verbose ../UD_Swedish-LinES/sv_lines-ud-test.conllu sv-ud-test.out

Prior to making changes to tagger.py, the baseline performance on **sv_lines-ud-train.conllu** was as follows:

Metrics	Precision	Recall	F1 Score	AligndAcc
Tokens	100.00	100.00	100.00	
Sentences	100.00	100.00	100.00	
Words	100.00	100.00	100.00	
UPOS	91.02	91.02	91.02	91.02
XPOS	100.00	100.00	100.00	100.00
Feats	100.00	100.00	100.00	100.00
AllTags	91.02	91.02	91.02	91.02
Lemmas	100.00	100.00	100.00	100.00
UAS	100.00	100.00	100.00	100.00
LAS	100.00	100.00	100.00	100.00

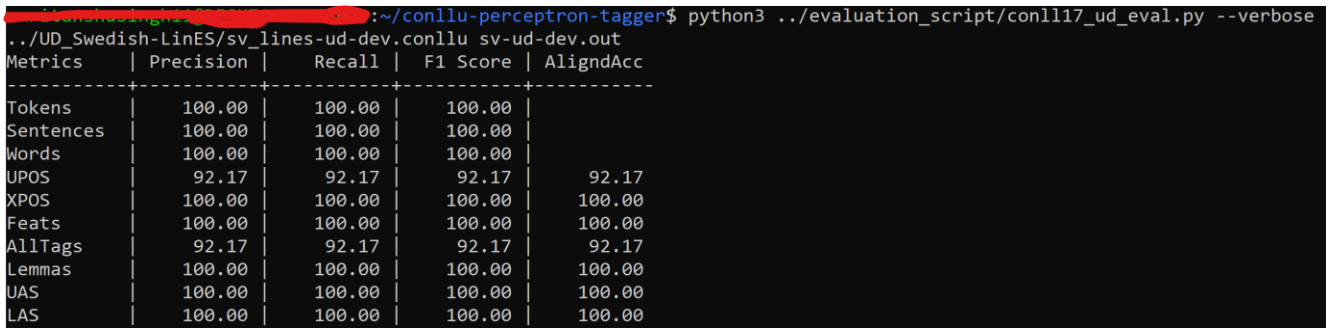
8. After changes were made to tagger.py, which included :
   - Changes made to the way suffixes were handled, which is represented by the following code snippet:

```
add('i suffix', word[-2:])

add('i-1 suffix', context[i-1][-2:])

add('i+1 suffix', context[i+1][-2:])
```

9. Outcome:



```
:~/conllu-perceptron-tagger$ python3 ../evaluation_script/conll17_ud_eval.py --verbose
../UD_Swedish-LinES/sv_lines-ud-dev.conllu sv-ud-dev.out
Metrics     | Precision |    Recall |  F1 Score | AlignedAcc
------------+-----------+-----------+-----------+-----------
Tokens      |    100.00 |    100.00 |    100.00 |
Sentences   |    100.00 |    100.00 |    100.00 |
Words       |    100.00 |    100.00 |    100.00 |
UPOS        |     92.17 |     92.17 |     92.17 |     92.17
XPOS        |    100.00 |    100.00 |    100.00 |    100.00
Feats       |    100.00 |    100.00 |    100.00 |    100.00
AllTags     |     92.17 |     92.17 |     92.17 |     92.17
Lemmas      |    100.00 |    100.00 |    100.00 |    100.00
UAS         |    100.00 |    100.00 |    100.00 |    100.00
LAS         |    100.00 |    100.00 |    100.00 |    100.00
```

After modifying & training tagger.py file and running it on the sv_lines-ud-dev.conllu dev file, I observed there was slight improvement metrics such as Precision, Recall and F1 score of UPOS and AllTags.