Akın Toksan
2010400186

CMPE 561 – Assignment 2
HMM Part-of-Speech Tagger

https://github.com/akintoksan/cmpe561/tree/master/Assignment2

Training the Tagger

For the training task, we need to matrices: observation likelihood matrix and transition probability matrix. Observation likelihood (OL) matrix keeps the information of how many times a word is labeled with a specific tag. (e.g. the probability of the word "hayır" is used as a NOUN) And transition probability (TP) matrix keeps the information that how many times a specific tag comes after another tag in a sentence. (e.g. the probability of the tag NOUN to come after the tag ADV). After filling these two matrices, the system is considered as trained.

To fill the matrices, I read the ***"turkish_metu_sabanci_train.conll"*** file line by line and kept the information of how many times a word is used as a specific tag and what tag comes after another in the sentences.

First, a traversed the training file once to see which tags are used. And I construct the TP matrix according to that. I also kept the tags in a list. After that I started to read the file line by line. And according to the tagset specification (cpostag or postag) a take the 3rd or 4th column of the file respectively. For the TP a checked for the order of the tags (kept them as bigrams in the matrix) and for the OL I counted the occurrences of each word with each tag.

Finally, I have written these to matrices into files "transition.txt" and "likelihood.txt", so that I can use them for upcoming tasks: tagging a test file and validating the result.

List of Tags

cpostag:

1. Adv
2. Punc
3. Noun
4. Adj
5. Verb
6. Det
7. Pron
8. Postp
9. Conj
10. Ques
11. Num
12. Interj
13. Dup
14. Zero

postag:

1. Adv
2. Punc
3. Noun
4. Adj
5. Zero
6. Verb
7. Prop
8. Det
9. PersP
10. Postp
11. Conj
12. NPastPart
13. APastPart
14. NInf
15. Ques
16. DemonsP
17. Card
18. APresPart
19. QuesP
20. AFutPart
21. Pron
22. ReflexP
23. NFutPart
24. Interj
25. Dup
26. Ord
27. Distrib
28. Real
29. Num
30. Range

Task 1 Details

As mentioned before, I read the training file and filled the TP and OL matrices. Then I wrote this matrices into the files "transition.txt" and "likelihood.txt" to use them later for other tasks. I also create a file named "tag_order.txt" which keeps the information of which tags are assigned to words. So it is basically the full list of the tags used. In the same file, I also keep the most common tag to label the unknown words with. Later while we test the tagger, each time the system encounters a word that is not in the training set, it assigns the word with the most common tag. I also kept the tagset type (cpostag or postag) to understand which tagset is used to train the system.

First, a traversed the training file once to see which tags are used. And I construct the TP matrix according to that. I also added the tag "Start" to understand which tags are used in the beginning of the sentences. I also kept these tags in a list and wrote them into "tag_order.txt". After that, I started to read the file line by line. And according to the tagset specification (cpostag or postag) a take the 3^{rd} or 4^{th} column of the file respectively. For the TP a checked for the order of the tags (kept them as bigrams in the matrix) and for the OL I counted the occurrences of each word with each tag.

Task 2 Details

Second task was to assign tags to the words in a given file. So that, we can test the tagger's accuracy.

First, a read the information of the matrices from the files "transition.txt" and "likelihood.txt".

After filling the matrices, I implemented the Viterbi algorithm. In Viterbi, we pick the maximum probability calculated by multiplying the probability of observation likelihood of the word and the transition probability coming from the previous tag. So for each word we repeat the process of checking the probability of the word having each tag and the probability of this tag is coming after from the previous tag.

Then we write the assigned tags with each word in a file.

Accuracies on the validation set

CPOSTAG

Accuracy for unknown: 0.535612535613
Accuracy for known: 0.818349186867
Overall Accuracy: 0.790858725762

Confusion Matrix

Example: C(Noun, Adj) = 0.18, looking at the intersection of the Noun row and Adjective column indicates number of times (with respect to total number of errors) that a tag noun has been assigned as adjective by mistake.

	Adv	Punc	Noun	Adj	Verb	Det	Pron	Postp	Conj	Ques	Num	Interj	Dup	Zero
Adv	_	0.00	0.18	0.11	0.58	0.03	0.03	0.07	0.00	0.00	0.00	0.00	0.00	0.00
Noun	0.00	0.00	_	0.18	0.79	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Det	0.04	0.00	0.00	0.19	0.00	_	0.78	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zero	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	_
Postp	0.40	0.00	0.40	0.20	0.00	0.00	0.00	_	0.00	0.00	0.00	0.00	0.00	0.00
Pron	0.00	0.00	0.00	0.10	0.00	0.90	_	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ques	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	_	0.00	0.00	0.00	0.00
Verb	0.02	0.00	0.82	0.13	_	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Interj	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	_	0.00	0.00
Dup	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	_	0.00
Punc	0.00	_	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Conj	0.00	0.00	0.00	0.00	0.86	0.00	0.14	0.00	_	0.00	0.00	0.00	0.00	0.00
Num	0.00	0.00	0.69	0.00	0.00	0.31	0.00	0.00	0.00	0.00	_	0.00	0.00	0.00
Adj	0.01	0.00	0.49	_	0.46	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00

POSTAG

Accuracy for unknown: 0.435897435897
Accuracy for known: 0.764344891071
Overall Accuracy: 0.732409972299

Confusion Matrix

Example: C(Adv, Adj) = 0.11, looking at the intersection of the Adverb row and Adjective column indicates the number of times (with respect to total number of errors) that a tag adverb has been assigned as adjective by mistake.

	Adv	Punc	Noun	Adj	Zero	Verb	Prop	Det	PersP	Postp	Conj	NPastPart	APastPart	NInf	Ques	DemonsP	Card	APresPart	QuesP	AFutPart	Pron	ReflexP	NFutPart	Interj	Dup	Ord	Distrib	Real	Num	Range
Adv	-	0.00	0.18	0.11	0.00	0.58	0.00	0.03	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PastPart	0.00	0.00	0.21	0.00	0.00	0.79	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Prop	0.00	0.00	0.98	0.02	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pron	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ques	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QuesP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dup	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00
ReflexP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Adj	0.02	0.00	0.89	-	0.00	0.03	0.00	0.02	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Real	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00
PastPart	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.04	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Verb	0.00	0.00	0.96	0.04	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FutPart	0.00	0.00	0.23	0.00	0.00	0.77	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FutPart	0.00	0.00	0.20	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ord	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.50	0.00
Card	0.00	0.00	0.32	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.52	0.00
Noun	0.00	0.00	-	0.50	0.00	0.48	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Det	0.04	0.00	0.00	0.19	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.78	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DemonsP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Range	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-
Postp	0.40	0.00	0.40	0.20	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Punc	0.00	-	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PersP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zero	0.02	0.00	0.33	0.21	-	0.35	0.00	0.04	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00
Num	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00
Interj	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00
NInf	0.00	0.00	0.30	0.01	0.00	0.68	0.00	0.00	0.00	0.00	0.01	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Distrib	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00
Conj	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PresPart	0.00	0.00	0.08	0.01	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Handling the unknown words

I assigned the most common tag (NOUN, for both tagsets) in the training data set to the unknown words. In the training task, I counted the occurrences of each tag and wrote the maximum one into the file "tag_order.txt"s second line.

Performance of the tagger

Considering that even it is nearly impossible for a human to tag the word 100% correctly, accuracy higher than 70% is not that bad. And accuracy can be higher due to how we handle the unknown words in the files. Also, if we used trigrams or n-grams where n is higher than three, instead of bigrams, for the Viterbi algorithm results would be better.

Ps. Text files "transition.txt", "likelihood.txt", "tag_order.txt", "unknown_words.txt" are created to keep the information from the previous tasks. The program creates them if they are not already in the directory.