CS626 - Speech, Natural Language Processing, and the Web

Assignment-1b
POS Tagging using CRF

Group Id-

Priyansh Singh ,<22b1856.>,<Electrical engineering.>
Arnav Agarwal,<22b3917.>,< Electrical engineering.>
Hardik Jangir,<22b3901.>,< Electrical engineering >
Kanishk Garg,<210050080.>,<computer science >

Date: 02/10/24

Problem Statement

- **Objective:** Given a sequence of words, produce the POS tag sequence using Conditional Random Field (CRF)
- **Input:** The quick brown fox jumps over the lazy dog
- **Output:** The$_{\text{DET}}$ quick$_{\text{ADJ}}$ brown$_{\text{ADJ}}$ fox$_{\text{NOUN}}$ jumps$_{\text{VERB}}$ over$_{\text{ADP}}$ the$_{\text{DET}}$ lazy$_{\text{ADJ}}$dog$_{\text{NOUN}}$
- **Dataset:** Brown corpus
- Use Universal Tag Set (12 in number)
  - [NUM, PRT, CONJ, DET, ADP, VERB, ADV, PRON, ADJ, X, NOUN, .]
- k-fold cross validation (k=5)

Data Processing Info (Pre-processing)

- loaded the dataset from the NLTK **Brown Corpus**
- converted the data from an iterator to a list.

Did the conversion into lower case, done while using ‘stemmer'

Tokenized the sentence into words in data preparation part

Removed the ‘apostrophe’ like don’t to do not

Features used

```
'word_stem'
'word_suffix
'word_length
'word_position
'prev_word
'next_word
'is_first_word
'is_last_word
'prev_pos_tag
```

Overall performance

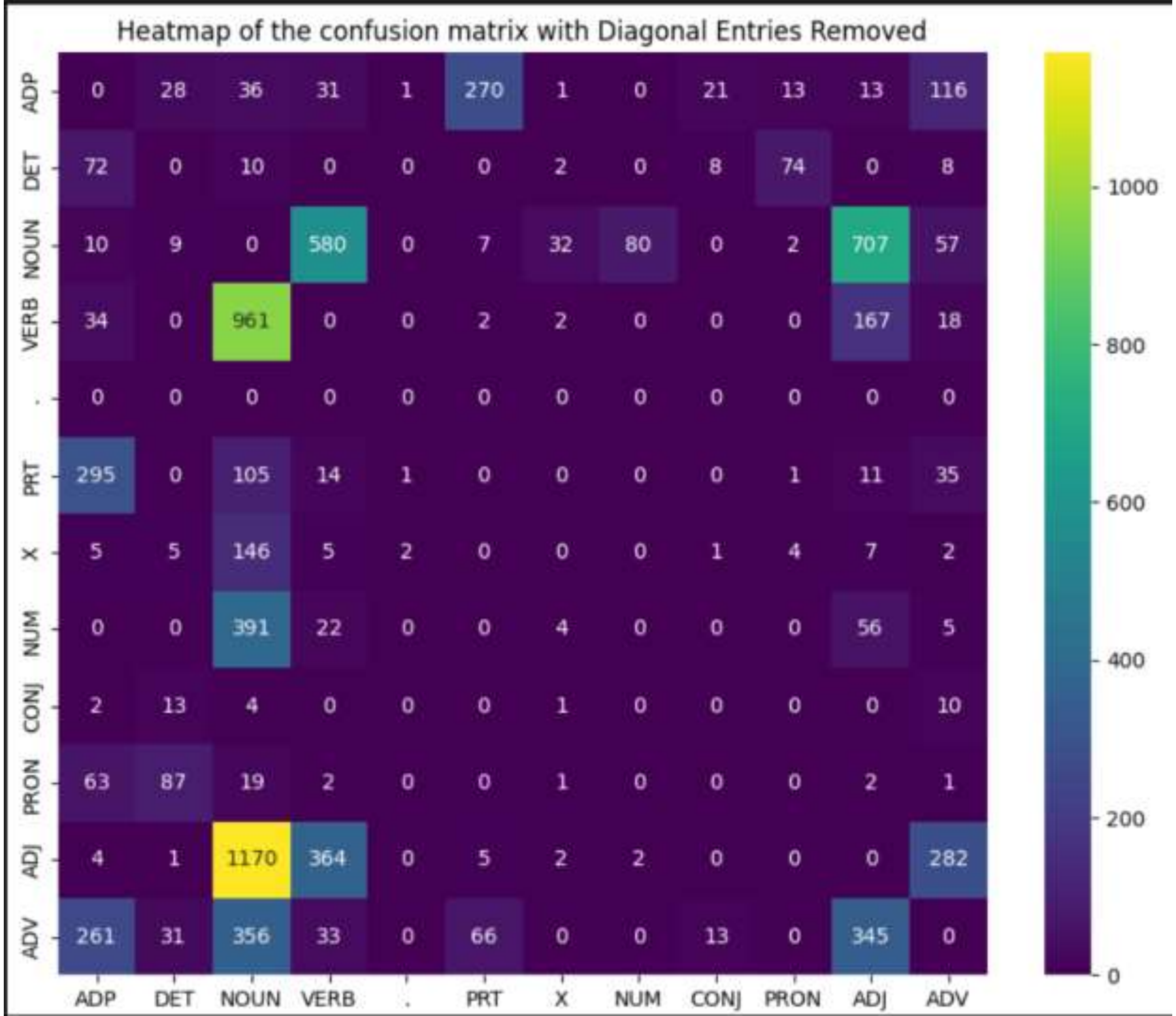
- Precision is 0.967
- Recall is 0.967
- F-score
  - F_1-score is 0.967
  - $F_{0.5}$-score is 0.242
  - F_2-score is 0.605

Per POS performance

- ORDER OF VALUES (P, R , F1)
- 'ADP’: (0.982, 0.974, 0.978)
- 'DET’: (0.993, 0.993, 0.993)
- 'NOUN': 0.973, 0.944, 0.958
- 'VERB': 0.967, 0.971, 0.969
- '.’: 1.0, 0.999, 0.999
- 'PRT': 0.922, 0.940, 0.931
- 'X': 0.294, 0.621, 0.4
- 'NUM': 0.842, 0.968, 2.252
- 'CONJ': 0.996, 0.994, 0.995
- 'PRON': 0.982, 0.99, 0.986
- 'ADJ': 0.890, 0.990, 0.904
- 'ADV': 0.903, 0.951, 0.926

Confusion Matrix (12 X 12) (can give heat map)

	ADP	DET	NOUN	VERB	.	PRT	X	NUM	CONJ	PRON	ADJ	ADV
ADP	28498	28	36	31	1	270	1	0	21	13	13	116
DET	72	27460	10	0	0	0	2	0	8	74	0	8
NOUN	10	9	53977	580	0	7	32	80	0	2	707	57
VERB	34	0	961	35668	0	2	2	0	0	0	167	18
.	0	0	0	0	29543	0	0	0	0	0	0	0
PRT	295	0	105	14	1	5497	0	0	0	1	11	35
X	5	5	146	5	2	0	74	0	1	4	7	2
NUM	0	0	391	22	0	0	4	2548	0	0	56	5
CONJ	2	13	4	0	0	0	1	0	7710	0	0	10
PRON	63	87	19	2	0	0	1	0	0	9613	2	1
ADJ	4	1	1170	364	0	5	2	2	0	0	14944	282
ADV	261	31	356	33	0	66	0	0	13	0	345	10376


Heatmap of the confusion matrix with Diagonal Entries Removed
ADP 0 28 36 31 1 270 1 0 21 13 13 116
DET 72 0 10 0 0 0 2 0 8 74 0 8
NOUN 10 9 0 580 0 7 32 80 0 2 707 57
VERB 34 0 961 0 0 2 2 0 0 0 167 18
. 0 0 0 0 0 0 0 0 0 0 0 0
PRT 295 0 105 14 1 0 0 0 0 1 11 35
X 5 5 146 5 2 0 0 0 1 4 7 2
NUM 0 0 391 22 0 0 4 0 0 0 56 5
CONJ 2 13 4 0 0 0 1 0 0 0 0 10
PRON 63 87 19 2 0 0 1 0 0 0 2 1
ADJ 4 1 1170 364 0 5 2 2 0 0 0 282
ADV 261 31 356 33 0 66 0 0 13 0 345 0
ADP DET NOUN VERB . PRT X NUM CONJ PRON ADJ ADV
1000
800
600
400
200
0

Interpretation of confusion (error analysis)

– **ADP confused with PRT**: 295 instances

Example: "He stood **up** the hill.“ (here up is ADP but can be confused with PRT)
Because : Words like *up, off, out* often serve as both adpositions and particles, depending on their usage in a sentence, like: "He looked **up** the word in the dictionary."

NOUN confused with ADJ: 707 instances

Example: The **building** collapsed. (can be classified as ADJ , due to presence of sentence like "It was a **building** project.“)
Nominalized adjectives and nouns used in attributive roles (modifying nouns) may confuse taggers, especially when a noun is used to describe another noun.

ADV confused with NOUN: 356 instances.

Exmple: "We went **home** after work.“ (here home is ADV , but mostly it serve as a noun)
Many words can serve both as nouns and adverbs depending on their syntactic role, leading to confusion between these categories.

VERB vs. NOUN Confusion

Sentence: "They **run** the company.
"Here, **"run"** is a verb (action), but if the context is ambiguous (e.g., no strong neighboring cues), the model might mislabel it as a noun.

CRFs model the probability of a label (POS tag) sequence given the word sequence. They use **neighboring words** and their labels to predict tags, but when a word can be both a verb and a noun (e.g., *run*, *play*, *design*), the model relies heavily on contextual features which we didn’t focus much in this assignment.

ADJ vs. NOUN Confusion

Reason for Confusion: CRF uses **features** such as POS tags of neighboring words, and character-level features (prefixes, suffixes).

In cases where nouns and adjectives overlap, the CRF model might misinterpret the context, especially if the training data has more frequent occurrences of one form over the other.

Comparison with HMM

Comparing accuracies P, R, F-score, confusion matrix, per POS accuracy for HMM and CRF

HMM: accuracy: 96.1%
CRF: accuracy: 96.733%

HMM: (P=0.961 R= 0.962 F1= 0.96)
CRF: (P= 0.967 R=0.967 F1=0.967)

From here CRF SEEMS to perform slightly better

HMM

Confusion Matrix:

	NOUN	ADJ	VERB	.	ADV	PRON	CONJ	DET	PRT	X	ADP	NUM
NOUN	269008	2793	2858	1	132	16	1	115	40	18	33	543
ADJ	1839	78935	327	0	2303	0	0	0	230	3	80	4
VERB	5300	477	176692	0	99	0	0	0	15	5	162	0
.	0	0	0	147565	0	0	0	0	0	0	0	0
ADV	100	3212	22	0	48879	0	124	373	860	1	2668	0
PRON	0	0	0	0	3	46414	0	1139	2	0	1776	0
CONJ	2	0	0	0	55	0	37965	126	0	0	3	0
DET	1	0	0	0	2	244	37	135566	1	5	1161	2
PRT	91	292	28	0	105	12	0	2	23618	0	5681	0
X	109	7	27	24	4	5	4	17	2	1158	24	5
ADP	48	95	97	20	1119	0	148	201	7681	5	135351	1
NUM	204	0	0	0	0	0	0	0	0	0	0	14670

CRF

	ADP	DET	NOUN	VERB	.	PRT	X	NUM	CONJ	PRON	ADJ	ADV
ADP	28498	28	36	31	1	270	1	0	21	13	13	116
DET	72	27460	10	0	0	0	2	0	8	74	0	8
NOUN	10	9	53977	580	0	7	32	80	0	2	707	57
VERB	34	0	961	35668	0	2	2	0	0	0	167	18
.	0	0	0	0	29543	0	0	0	0	0	0	0
PRT	295	0	105	14	1	5497	0	0	0	1	11	35
X	5	5	146	5	2	0	74	0	1	4	7	2
NUM	0	0	391	22	0	0	4	2548	0	0	56	5
CONJ	2	13	4	0	0	0	1	0	7710	0	0	10
PRON	63	87	19	2	0	0	1	0	0	9613	2	1
ADJ	4	1	1170	364	0	5	2	2	0	0	14944	282
ADV	261	31	356	33	0	66	0	0	13	0	345	10376

HMM

Precision, Recall, F1-Score, F0.5-Score, F2-Score:

	Precision	Recall	F1-Score	F0.5-Score	F2-Score
NOUN	0.972194	0.976230	0.974208	0.972998	0.975420
ADJ	0.919870	0.942834	0.931211	0.924372	0.938150
VERB	0.981344	0.966851	0.974044	0.978410	0.969715
.	0.999695	1.000000	0.999848	0.999755	0.999939
ADV	0.927478	0.869130	0.897356	0.915189	0.880205
PRON	0.994067	0.940812	0.966707	0.982938	0.951001
CONJ	0.991797	0.995125	0.993458	0.992460	0.994457
DET	0.985655	0.989396	0.987522	0.986400	0.988645
PRT	0.727850	0.791780	0.758470	0.739796	0.778111
X	0.969038	0.835498	0.897327	0.939020	0.859178
ADP	0.921137	0.934964	0.927999	0.923869	0.932165
NUM	0.963547	0.986285	0.974783	0.968009	0.981652

CRF

```
{'ADP': 0.9781026908292147,
 'DET': 0.9937034088441775,
 'NOUN': 0.9584324727440606,
 'VERB': 0.9696211822593143,
 '.': 0.9999323066508715,
 'PRT': 0.9312214128409283,
 'X': 0.4,
 'NUM': 0.900990099009901,
 'CONJ': 0.99528819466856,
 'PRON': 0.9862015901513209,
 'ADJ': 0.9049839520377884,
 'ADV': 0.9268009468089857}
```

Judging by f1 scores we see that HMM and CRF approximately equally accurate but overall CRF slightly performs better HMM performed Slightly better for (NOUN,ADJ, VERB, NUM) but for the rest of the tags CRF performed quite better than HMM. ‘X’ tag is the exception.

Classification tag	Hmm better	CRF better	equal	Example and reason
NOUN	YES	NO	NO	Sentence: “The quick brown fox jumps over the lazy dog. ”HMM advantage: The noun "fox" can be identified correctly by HMM because the words "quick" and "brown" provide a clear sequential pattern typical in noun phrases. HMM excels in learning these patterns because it explicitly models the dependencies between adjacent words.CRF may struggle slightly more because it doesn't rely as heavily on sequential relationships like HMM does.
ADJ	YES	NO	no	Sentence: “She wore a red dress.”HMM advantage: "red" is an adjective that describes the noun "dress." HMM performs well here as it models the adjacency between adjectives and nouns, identifying "red" as likely being an adjective because of its proximity to "dress."CRF may not capture this sequential reliance as strongly since it focuses on feature interactions but does not explicitly model sequences.
Verb	yes	no	no	Sentence: “He is running to the store.” HMM advantage: Verbs like "running" can be better identified by HMM due to the dependency on "is," forming a verb phrase. HMM’s strength in modeling sequences helps identify this verb, especially in context with auxiliary verbs.CRF, which focuses on conditional independence between features, might not capture the temporal aspect of verb phrases as effectively as HMM
DET	no	YES	no	Sentence: “The cat sat on a mat.” CRF advantage: CRF can capture the complex interactions between words that indicate determiner usage. For instance, "the" and "a" function as determiners and are linked to their respective nouns "cat" and "mat." CRF’s ability to model detailed feature interactions allows it to identify the determiners correctly even without strong reliance on sequences.HMM might misclassify determiners in cases where the surrounding context doesn't strictly follow typical sequential patterns.
ADP	no	yes	no	Sentence: “She sat on the chair under the tree. ”CRF advantage: Prepositions like "on" and "under" can be better identified by CRF because it models the context-sensitive nature of prepositions. CRF’s ability to take in a variety of features (such as syntactic and semantic cues) without strong sequential assumptions gives it an edge in handling complex relationships like spatial orientation.HMM may not handle this as well, particularly in cases where prepositions don't always follow a clear sequential pattern with nearby nouns or verbs.PRON Classification (Pronouns):
PRON	no	YES	no	Sentence: “She loves her dog. ”CRF advantage: CRF can better identify "She" and "her" as pronouns by using the contextual information and feature interactions without assuming the sequence must follow a strict order. Pronouns like "her" can sometimes appear in contexts that don't strictly adhere to common word-order patterns, which CRF handles more flexibly.HMM might misclassify some pronouns because it expects certain sequential patterns that pronouns may not always follow.

And for the rest of the cases the performance of both CRF and HMM is almost same but slightly better performance by crf due to its flexibility in modeling

Challenges faced

1. Selecting appropriate features to maximize accuracy , and especially to achieve better output for the input "foxes foxes fox fox foxes" which gave "Noun" tag for all words.

2. using pos tag of the previous word as a feature, for that we stacked 2 crf models, first one was a simpler one to Give pos tag for a word using simple features and the later gave pos tag using the earlier features along with the founded pos tag for pos word.

3. The stacking approach took a lot of time to train the model and applying 5 fold validation on this was itself A big challenge for us.

4. Handling unknown word was also little tricky. Handeled by doing smoothing.

References

1. https://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf
2. https://aclanthology.org/N03-1028
3. https://www.analyticsvidhya.com/blog/2018/08/nlp-guide-conditional-random-fields-text-classification/
4. Lecture slides for basic understanding and basic features used

Marking Scheme (50)

1. Demo working- 10/10 (if not working or no GUI - 0)
2. Implemented CRF and Clarity on CRF- 5/5
3. Forward and Backward vector clearly described- 5/5
4. Confusion matrix drawn and error analysed- 5/5
5. **Overall F_1-score**
    a. **> 90** - 10/10
    b. **>80 & <=90** - 8/10
    c. **>70 & <=80** - 7/10
    d. **so on**.
6. Unknown word handling- done (5/5; else 0)
7. Comparison with HMM (10)

Note: Must have GUI, otherwise no mark will be given for demo.