CS626 - Speech, Natural Language Processing, and the Web

Assignment-1b
POS Tagging using CRF

Group Id- 37

Amish Sethi, 22B3029, BS. Economics
Ashutosh Agarwal, 22B2187, B.Tech. Mechanical
Jayesh Ahire, 22B2101, B.Tech. Mechanical
Pratham Agarwal, 22B2111, BS. Economics

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_{DET} $\text{quick}_{\text{ADJ}}$ $\text{brown}_{\text{ADJ}}$ fox_{NOUN} $\text{jumps}_{\text{VERB}}$ over_{ADP} the_{DET} lazy_{ADJ} dog_{NOUN}
- **Dataset:** Brown corpus
- Use Universal Tag Set (12 in number) –

  <., ADJ, ADP, ADV, CONJ, DET, NOUN, NUM, PRON, PRT, VERB, X>
- k-fold cross validation (k=5)

Data Processing Info (Pre-processing)

- **Tokenization**
- **Lower Casing**
- **Suffix and Prefix Analysis:** Features for the last two and three characters of each word are extracted (word[-2:], word[-3:]). This helps in identifying common suffixes that might correlate with certain POS tags (e.g., "-ly" for adverbs, "-ing" for verbs).
- **Word Structure:** Identifying words with hyphens or numeric content (has_hyphen, is_numeric)
- **Position-Based Features:**
  - First or last word of the sentence (is_first, is_last).
  - Surrounding words (prev_word, next_word)
- **Capitalization**:
  - Capital letters inside the word (capitals_inside).
  - Word capitalization (is_capitalized)

Overall performance

- Precision : 0.991
- Recall : 0.984
- F-score (3 values)
  - F_1-score: 0.987
  - $F_{0.5}$-score: 0.989
  - F_2-score: 0.985

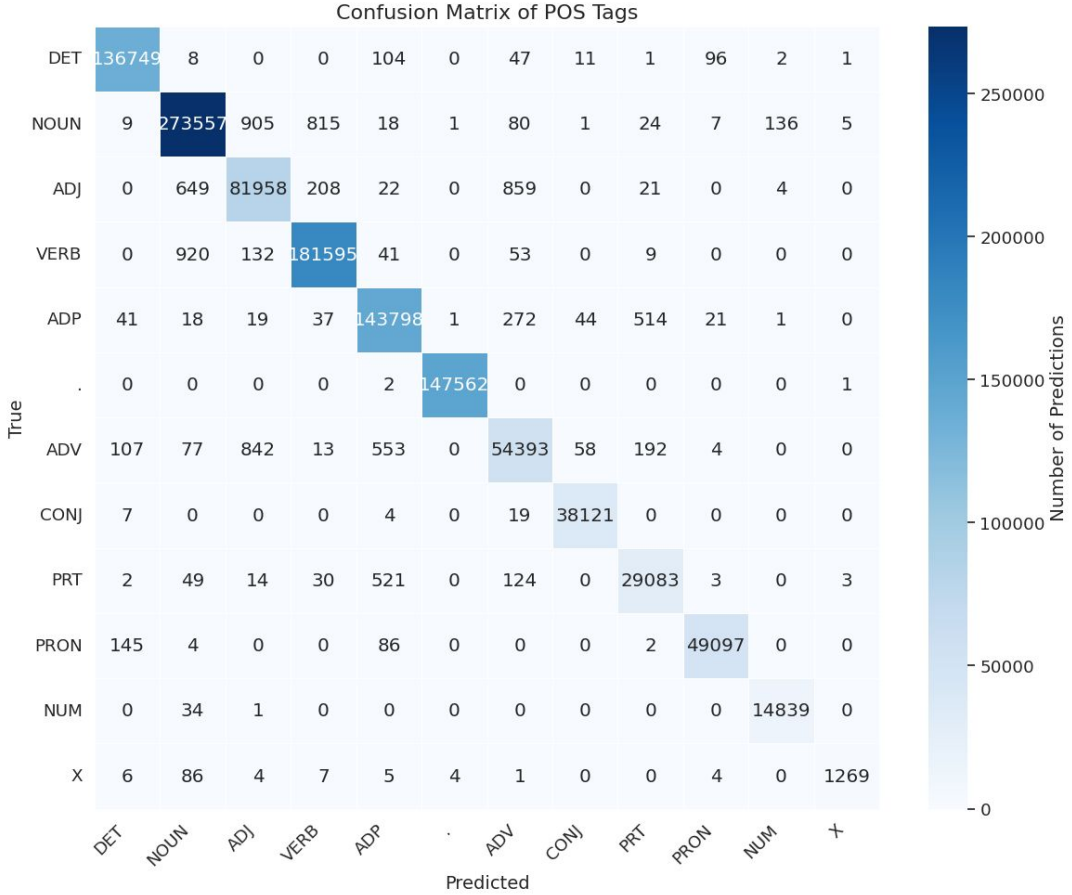
The general formula for non-negative real β is:

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}$$

Per POS performance

	precision	recall	f1-score	support
DET	0.997687	0.998029	0.997858	137019.000000
NOUN	0.993301	0.992738	0.993019	275558.000000
ADJ	0.977145	0.978942	0.978042	83721.000000
VERB	0.993925	0.993680	0.993802	182750.000000
ADP	0.990658	0.993313	0.991984	144766.000000
.	0.999959	0.999980	0.999970	147565.000000
ADV	0.973947	0.967176	0.970550	56239.000000
CONJ	0.997018	0.999214	0.998115	38151.000000
PRT	0.974435	0.974991	0.974713	29829.000000
PRON	0.997258	0.995196	0.996226	49334.000000
NUM	0.990455	0.997647	0.994038	14874.000000
X	0.992181	0.915584	0.952345	1386.000000

Confusion Matrix (12 X 12)



Interpretation of confusion(error analysis)

- By trying out many examples, we see that the tag **Noun** and **Verb** are getting confused the most by our model.
- There are some words like Running, fish which are getting confused
- Reasons:

  1. The model is not getting the contextual understanding
  2. Polysemy: Many words have multiple meanings and can function as different parts of speech depending on the context
  3. There is training data imbalance.

Comparison with HMM

	precision	recall	f1-score	support
DET	0.997687	0.998029	0.997858	137019.000000
NOUN	0.993301	0.992738	0.993019	275558.000000
ADJ	0.977145	0.978942	0.978042	83721.000000
VERB	0.993925	0.993680	0.993802	182750.000000
ADP	0.990658	0.993313	0.991984	144766.000000
.	0.999959	0.999980	0.999970	147565.000000
ADV	0.973947	0.967176	0.970550	56239.000000
CONJ	0.997018	0.999214	0.998115	38151.000000
PRT	0.974435	0.974991	0.974713	29829.000000
PRON	0.997258	0.995196	0.996226	49334.000000
NUM	0.990455	0.997647	0.994038	14874.000000
X	0.992181	0.915584	0.952345	1386.000000

	precision	recall	f1-score	support
.	0.98	1.00	0.99	29490
ADJ	0.87	0.89	0.88	16572
ADP	0.92	0.97	0.94	29106
ADV	0.90	0.86	0.88	11214
CONJ	0.99	0.99	0.99	7631
DET	0.92	0.99	0.95	27536
NOUN	0.95	0.92	0.93	55064
NUM	0.99	0.80	0.88	2958
PRON	0.93	0.96	0.94	9933
PRT	0.91	0.85	0.88	5916
VERB	0.97	0.91	0.94	36732
X	0.15	0.31	0.21	279

Sentences	HMM better	CRF Better	Both Equal
India won 2nd world cup in 2011	No	Yes	No
the quick brown fox jumps over a lazy dog	Yes	No	No
I need to call my mom to check on her.	No	Yes	No
The playing field is open for discussion.	No	No	Yes
I can fish in the river	No	Yes	No

Comparison with HMM

Sentences	HMM better	CRF Better	Both Equal
India won 2nd world cup in 2011	No	Yes	No
the quick brown fox jumps over a lazy dog	Yes	No	No
I need to call my mom to check on her.	No	Yes	No
The playing field is open for discussion.	No	No	Yes
I can fish in the river	No	Yes	No

Reasons Explaining the Observations:

1) In the first example…2011 is classified as numeric by CRF but not by HMM since we have added a feature IS_NUMERIC in our CRF Model

2) Similarly, in the sentence, " I can fish in the river ", fish can be classified as both verb and a noun, since HMM just considers the transition probabilities and preceding label of the word, while CRF models entire sentence as a whole, thus better capturing the entire sentence and correctly identifying fish as a verb.

3) CRF's ability to utilize more contextual information helps it distinguish between "to" as a particle (PRT) in "to call" and as a preposition (ADP) in "to check on her," while the HMM struggles with this distinction.

Challenges faced

- **Feature Extraction**: For unknown words, additional features like prefixes, suffixes, and capitalization patterns were implemented to provide context in the absence of word embeddings.
- **Memory Overload**: The runtime failed due to excessive RAM usage when storing and processing embeddings for known and unknown words during feature extraction.
- **Challenge**: Handling unknown words dynamically led to high memory consumption, especially when computing embeddings or looking for similar words in large datasets.

References

1. https://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf
2. https://aclanthology.org/N03-1028
3. https://towardsdatascience.com/pos-tagging-using-crfs-ea430c5fb78b
4. https://www.geeksforgeeks.org/conditional-random-fields-crfs-for-pos-tagging-in-nlp/