CS626 - Speech, Natural Language Processing, and the Web

Assignment-1b
POS Tagging using CRF

Group ID – 68
Rana Das – 22B0684 – Civil Dept.
Ankit - 22B0684 – Civil Dept.
Aman Mitra - 22B0684 – Civil Dept.
Swapna Sourav Rout – 22D1623 – KCDH

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)

{"PAD": 0, "DET": 1, "NOUN": 2, "ADJ": 3, "VERB": 4, "ADP": 5, ".": 6, "ADV": 7, "CONJ": 8, "PRT": 9, "PRON": 10, "NUM": 11, "X": 12}

- k-fold cross validation (k=5)

Data Processing Info (Pre-processing)

1. Data Loading: Sentences and tags are loaded and converted into numerical indices using vocabularies (word_to_ix, tag_to_ix).

2. Feature Extraction: Features such as word stems and suffixes are extracted to enhance POS tagging performance, providing richer input for the CRF model.

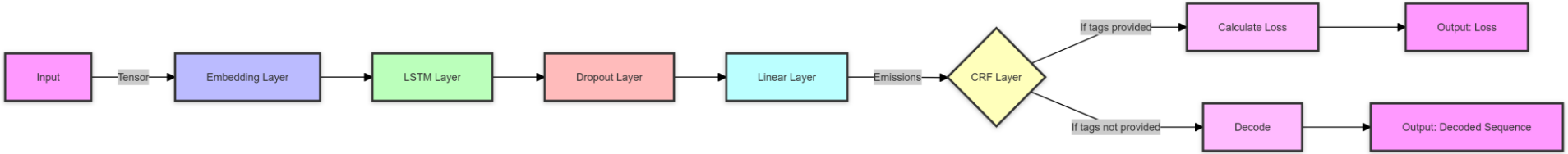
3. Padding & Masking: Sequences are padded for uniform length, and masks are created to identify valid tokens for model processing.

4. Data Splitting: The dataset is split into training (80%) and validation (20%) sets for model evaluation.

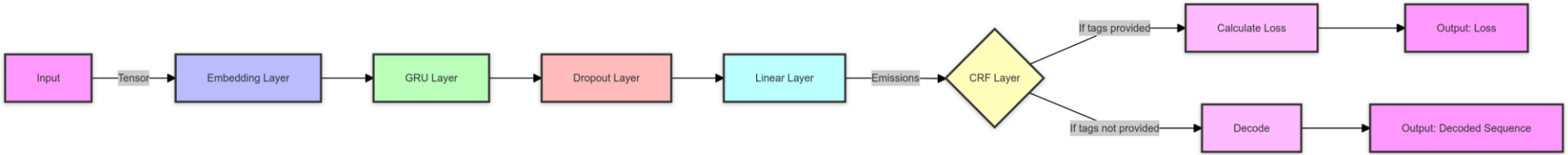
5. CUDA Handling: If available, data tensors (sentences, tags, mask) are moved to GPU for faster training.

Model Architecture

LSTM Model



GRU Model



Overall performance *for LSTM Model*

- Precision: 99.519%
- Recall: 99.520%
- F-score
  - F_1-score: 99.518%
  - $F_{0.5}$-score: 99.519%
  - F_2-score: 99.519%

Overall performance *for GRU Model*

- Precision: 99.485%
- Recall: 99.486%
- F-score
  - F_1-score: 99.485%
  - $F_{0.5}$-score: 99.485%
  - F_2-score: 99.485%

Per POS performance *for LSTM Model*

Format: <Tag>: <P>, <R>, <F1>

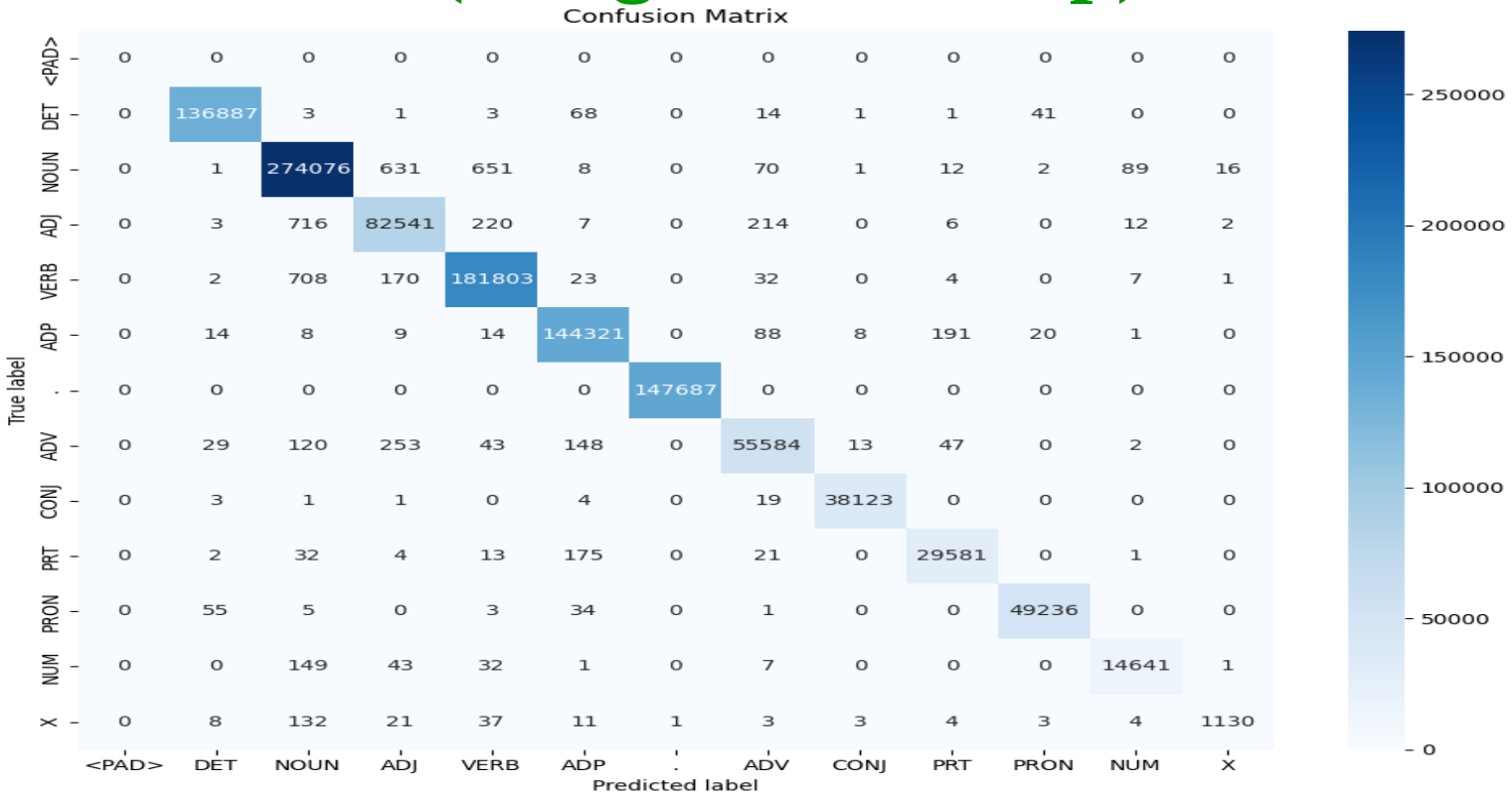
- **<DET>: 0.999, 0.999, 0.999**
- **<NOUN>: 0.993, 0.994, 0.994**
- **<ADJ>: 0.986, 0.986, 0.986**
- **<VERB>: 0.994, 0.995, 0.995**
- **<ADP>: 0.996, 0.997, 0.997**
- **<.>: 0.999, 1.000, 1.000**
- **<ADV>: 0.992, 0.988, 0.990**
- **<CONJ>: 0.999, 0.999, 0.999**
- **<PRT>: 0.991, 0.992, 0.991**
- **<PRON>>: 0.999, 0.998, 0.998**
- **<NUM>: 0.992, 0.984, 0.988**
- **<X>: 0.983, 0.833, 0.901**

Per POS performance *for GRU Model*

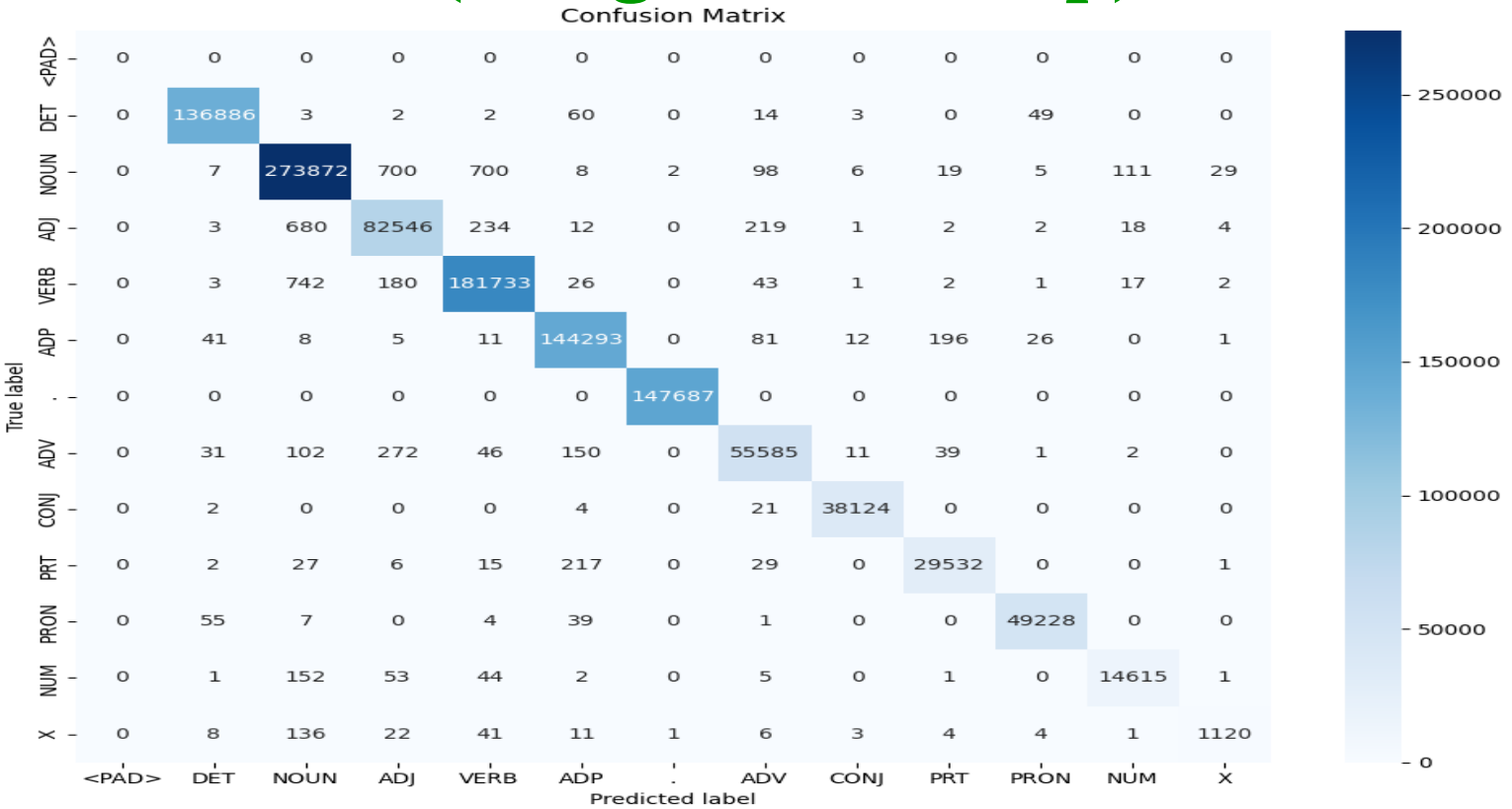
Format: <Tag>: <P>, <R>, <F1>

- **<DET>: 0.999, 0.999, 0.999**
- **<NOUN>: 0.993, 0.994, 0.994**
- **<ADJ>: 0.985, 0.986, 0.986**
- **<VERB>: 0.994, 0.994, 0.994**
- **<ADP>: 0.996, 0.997, 0.997**
- **<.>: 1.000, 1.000, 1.000**
- **<ADV>: 0.991, 0.988, 0.990**
- **<CONJ>: 0.999, 0.999, 0.999**
- **<PRT>: 0.991, 0.990, 0.991**
- **<PRON>: 0.998, 0.998, 0.998**
- **<NUM>: 0.990, 0.983, 0.986**
- **<X>: 0.967, 0.825, 0.891**

Confusion Matrix *for LSTM Model* (12 X 12) (can give heat map)



Confusion Matrix *for GRU Model* (12 X 12)
(can give heat map)



Interpretation of confusion *for LSTM Model* (error analysis)

ADJ ↔ NOUN (716 confusions):

Examples: "Daily," "poor," "rich" can be used both as adjectives and nouns (e.g., "the **rich** get richer" vs. "the **rich** man").

Reason: Just like with the GRU model, adjectives and nouns can often be difficult to distinguish, especially with words that naturally serve dual roles.

VERB ↔ NOUN (708 confusions):

Examples: "Watch," "run," "walk" have similar confusions as described for the GRU model.

Reason: Both models face difficulty with words that have multiple roles depending on context.

NOUN ↔ VERB (651 confusions):

Examples: Similar to the VERB↔NOUN issue above.

Reason: Ambiguous syntactic positions lead to incorrect predictions. When a sentence lacks clear indicators (e.g., an auxiliary verb for verbs), the model may mislabel.

NOUN ↔ ADJ (631 confusions):

Examples: Again, words like "daily" and "paper" are challenging.

Reason: Without enough surrounding context, nouns modifying other nouns (as adjectives) are easy to confuse.

ADV ↔ ADJ (253 confusions):

Examples: "Fast" can be an adverb ("She runs **fast**") or an adjective ("a **fast** car").

Reason: The role of modifiers can be ambiguous when the relationship to the verb isn't clear, leading to frequent errors.

Interpretation of confusion *for GRU Model* (error analysis)

VERB ↔ NOUN (742 confusions):

Examples: Words like "run," "play," or "dance" can be used both as nouns and verbs depending on the context. For instance, "He **runs** fast" (VERB) vs. "The **run** was long" (NOUN).

Reason: Many words in English can serve as both verbs and nouns (a phenomenon known as zero derivation). This confusion is likely exacerbated in sentences where context doesn't strongly indicate the function of the word.

NOUN ↔ VERB (700 confusions):

Examples: "Plan," "attack," "use" can all be both nouns and verbs. Example: "He made a **plan**" (NOUN) vs. "He will **plan** the event" (VERB).

Reason: Similar to the VERB↔NOUN confusion, ambiguous sentence structures or insufficient context can mislead the model into making incorrect predictions.

NOUN ↔ ADJ (700 confusions):

Examples: "Daily" can be both an adjective (as in "a **daily** routine") or a noun ("She reads the **daily**").

Reason: Some nouns function adjectivally, especially when modifying other nouns (e.g., "a **paper** bag"). This type of confusion occurs when the model misinterprets whether a word describes another word or acts as the subject/object itself.

ADJ ↔ NOUN (680 confusions):

Examples: "Cold" can be both an adjective ("a **cold** day") and a noun ("He caught a **cold**").

Reason: When a noun and adjective have the same form, especially in brief or isolated contexts, it becomes difficult for the model to distinguish between them.

ADV ↔ ADJ (272 confusions):

Examples: Words like "quickly" (ADV) and "quick" (ADJ) are easily confused, especially when context is not clear.

Reason: The model might confuse adverbs and adjectives in sentences where the modifier's relationship with the verb is ambiguous.

Comparison with HMM

Overall Metrics:

Metric	HMM	CRF (LSTM)	CRF (GRU)
Precision	0.7465	0.9952	0.9949
Recall	0.7995	0.9952	0.9949
F1-Score	0.7546	0.9952	0.9948

Comparison with HMM

Per POS Tag Performance:

POS Tag	HMM Precision	HMM Recall	HMM F1	CRF (LSTM) Precision	CRF (LSTM) Recall	CRF (LSTM) F1	CRF (GRU) Precision	CRF (GRU) Recall	CRF (GRU) F1
DET	0.8333	0.7056	0.7642	0.9991	0.9990	0.9991	0.9989	0.9990	0.9990
NOUN	0.8333	0.6757	0.7463	0.9932	0.9946	0.9939	0.9933	0.9939	0.9936
ADJ	0.8333	0.6527	0.7321	0.9865	0.9859	0.9862	0.9852	0.9860	0.9856
VERB	0.8333	0.6977	0.7595	0.9944	0.9948	0.9946	0.9940	0.9944	0.9942
ADP	0.8333	0.6915	0.7558	0.9967	0.9976	0.9971	0.9963	0.9974	0.9969
.	0.8333	0.7251	0.7755	0.9999	1.0000	1.0000	0.9999	1.0000	1.0000
ADV	0.8333	0.6519	0.7316	0.9916	0.9884	0.9900	0.9908	0.9884	0.9896
CONJ	0.8333	0.7010	0.7615	0.9993	0.9993	0.9993	0.9990	0.9993	0.9992
PRT	0.8333	0.5561	0.6670	0.9911	0.9917	0.9914	0.9912	0.9900	0.9906
PRON	0.8333	0.7105	0.7670	0.9987	0.9980	0.9983	0.9982	0.9979	0.9980
NUM	0.8333	0.6242	0.7138	0.9921	0.9843	0.9882	0.9899	0.9826	0.9862
X	0.8970	0.6520	0.7535	0.9826	0.8327	0.9015	0.9672	0.8254	0.8907

Comparison with HMM

Confusion Matrix Analysis:

Confusion (Tag 1 -> Tag 2)	HMM Confusion	CRF (LSTM) Confusion	CRF (GRU) Confusion
NOUN -> DET	10,983	N/A	N/A
VERB -> NOUN	10,698	708	742
. -> NOUN	10,161	N/A	N/A
NOUN -> ADJ	9,447	700	700
ADP -> NOUN	8,758	N/A	N/A
ADJ -> NOUN	9447	716	680
NOUN -> VERB	651	700	700
ADV -> ADJ	253	272	272

Comparison with HMM

Specific Examples: (a) HMM is better, (b) CRF (LSTM/GRU) is better, (c) both equal

Example Table:

Word	True Tag	HMM Prediction	CRF (LSTM) Prediction	CRF (GRU) Prediction	Better Model	Reason
the	DET	DET	DET	DET	Both Equal	Determiner (DET) is easily identified by all models.
run	NOUN	VERB	NOUN	NOUN	CRF Better	CRF models (LSTM and GRU) correctly identify ambiguous words as nouns.
quick	ADJ	NOUN	ADJ	ADJ	CRF Better	CRF (both LSTM and GRU) disambiguate adjectives better.
and	CONJ	CONJ	CONJ	CONJ	Both Equal	Simple conjunctions are identified correctly by all models.
5	NUM	NOUN	NUM	NUM	CRF Better	CRF (both LSTM and GRU) better distinguish numbers from nouns.
fast	ADV	ADJ	ADV	ADV	CRF Better	CRF models handle adverb/adjective confusion better than HMM.
is	VERB	NOUN	VERB	VERB	CRF Better	CRF (both LSTM and GRU) correctly tag auxiliary verbs.
.	PUNCT	NOUN	.	.	CRF Better	CRF has near-perfect precision for punctuation, while HMM confuses it.

Challenges faced

•Handling out-of-vocabulary words: One challenge is dealing with words not seen during training. The code addresses this by using an <UNK> token, but deciding how to represent and handle unknown words can significantly impact model performance.

•Balancing model complexity and performance: The code uses both GRU and LSTM models with CRF layers. Determining the optimal architecture, hyperparameters, and training approach to balance computational efficiency and tagging accuracy was challenging.

•Evaluating model quality: While the code tracks accuracy, determining the true quality of a POS tagger often requires more nuanced evaluation, including analysis of specific tag confusions and performance on different types of text. Interpreting these metrics to guide further improvements can be complex.

References

1. https://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf
2. https://aclanthology.org/N03-1028
3. https://pytorch-crf.readthedocs.io/en/stable/
4. https://docs.streamlit.io/
5. https://www.nltk.org/_modules/nltk.html
6. https://medium.com/the-modern-scientist/conditional-random-fields-for-part-of-speech-tagging-in-natural-language-processing-0f444133d455

Marking Scheme (50)

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

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