Assignment 2: Named Entity Recognition (NER)

Dataset

- **Dataset:** CoNLL-2003 (English)
- **Split sizes:**
  - Training: 14,041 sentences
  - Validation: 3,250 sentences
  - Test: 3,453 sentences
- **Entity types:** LOC (location), PER (person), ORG (organization), MISC (miscellaneous)

Models

- **Model 1:** Conditional Random Field (CRF) with handcrafted lexical and POS features
  - Features: word form, capitalization, digit status, POS tag, word prefixes/suffixes, neighboring context
  - Transition-aware: enforces valid BIO tag sequences
- **Model 2:** BiLSTM tagger with word embeddings and a linear classifier
  - Architecture: Embedding → BiLSTM → Dropout → Linear classifier
  - Bidirectional context: processes sequences left-to-right and right-to-left simultaneously

Hyperparameter Tuning

- **CRF:** Grid search over C1 (L1 regularization) and C2 (L2 regularization); 5 configurations tested
- **BiLSTM:** Sweep over embedding dimension (100), hidden size (128/256), dropout (0.2/0.3), learning rate (1e-3/5e-4); 2 configurations evaluated; 3 epochs each

Results

Test Set Performance (Entity-Level Metrics)

Model	Precision	Recall	F1 Score
CRF	0.7977	0.7835	0.7905
BiLSTM	0.6954	0.5837	0.6347

Per-Entity Breakdown (CRF on test set)

- LOC: Precision 0.8275, Recall 0.8195, F1 0.8235
- PER: Precision 0.8453, Recall 0.8479, F1 0.8466
- ORG: Precision 0.7307, Recall 0.6960, F1 0.7129
- MISC: Precision 0.7685, Recall 0.7564, F1 0.7624

Per-Entity Breakdown (BiLSTM on test set)

- LOC: Precision 0.8035, Recall 0.6691, F1 0.7301
- PER: Precision 0.6553, Recall 0.5832, F1 0.6171

- ORG: Precision 0.6242, Recall 0.5400, F1 0.5791
- MISC: Precision 0.7164, Recall 0.4858, F1 0.5789

Error Analysis

Common Error Patterns and Examples

1. **Country/Location Ambiguity (ORG vs LOC)**
   - Example: "United Arab Emirates" tagged as token sequence `[United(B-ORG), Arab(I-ORG), Emirates(I-ORG)]` when ground truth is `[United(B-LOC), Arab(I-LOC), Emirates(I-LOC)]`
   - Reason: Geographic entities that function as political organizations (country names, regional groupings) are inherently ambiguous. The CRF model relies heavily on surface-level features and POS tags, which may struggle to distinguish geopolitical entities.

2. **Person Name Boundary Errors (B-PER vs I-PER)**
   - Example: "Hassan" correctly tagged as B-PER but "Bitar" predicted as B-ORG instead of B-PER; "Cuttitta" sometimes confused with B-LOC
   - Reason: Lack of character-level features makes it difficult to recognize non-English surnames. The model cannot infer that certain word prefixes/suffixes are indicative of personal names. Without pretrained word embeddings from larger datasets, rare names remain underrepresented.

3. **Boundary Detection on Multi-Token Entities (O vs B-*)**
   - Example: "LUCKY" predicted as B-ORG when true label is O; "ROME" predicted as B-ORG instead of B-LOC
   - Reason: The BiLSTM model in particular lacks sufficient training signals to disambiguate word boundaries. With only 3 epochs and basic feature vectors, it cannot reliably capture the contextual cues needed to identify whether a word is the start of an entity or simply a common word. The CRF's explicit feature engineering helps mitigate this, but still shows boundary errors (~10% of misclassifications).

4. **Event/Miscellaneous Entity Detection**
   - Example: "1995" in sports context (FIFA World Cup qualifier) predicted as O instead of B-MISC; "World" predicted as B-MISC instead of I-MISC
   - Reason: MISC (miscellaneous) entities require deeper contextual understanding—they encompass events, products, awards, and named events. The models lack domain-specific awareness and cannot leverage external knowledge about major sporting events or their typical date patterns.

Error Type Frequency (CRF Test Set)

- ORG entity errors: 665 misclassifications
- O (non-entity) errors: 429 misclassifications
- LOC entity errors: 383 misclassifications
- PER entity errors: 302 misclassifications

- MISC entity errors: 232 misclassifications

Performance Analysis: CRF vs BiLSTM

Why CRF Outperforms BiLSTM

The CRF model achieves significantly higher test F1 (0.7905 vs 0.6347, a 24.6% absolute improvement):

1. **Feature Engineering Advantage:** CRF's handcrafted features explicitly encode POS tags and morphological cues that are highly informative for NER. The model directly observes "word is capitalized" and "word is preceded by a proper noun," which are strong signals for entity boundaries.

2. **Sequence Constraints:** CRF enforces valid BIO transitions, preventing the model from predicting impossible tag sequences (e.g., I-ORG followed by B-ORG without an O in between). BiLSTM predicts each tag independently, sometimes generating nonsensical sequences.

3. **Limited BiLSTM Training:** Only 3 epochs of training—the BiLSTM validation F1 jumped from 0.44 (epoch 1) to 0.63 (epoch 2) to 0.69 (epoch 3), indicating the model was still learning but undertrained compared to the CRF (which has converged features).

4. **Word Embedding Quality:** BiLSTM embeddings are learned from scratch on only 14K training sentences (67K unique words), yielding poor representations for rare and out-of-vocabulary words. CRF features degrade gracefully when encountering unknown words.

BiLSTM Advantages (Despite Lower Score)

- **Recall on LOC:** BiLSTM achieves 0.6691 recall on LOC vs CRF's 0.8195—this is lower, but suggests BiLSTM could model longer-range patterns with more training.
- **Potential:** BiLSTM's bidirectional context is conceptually superior for capturing linguistic patterns; with better hyperparameter tuning and more epochs, it could close the gap.

Limitations

1. **CRF Feature Engineering:**
    - Features are manually designed and language-specific. Creating good features requires domain expertise and linguistic knowledge.
    - No character-level features: The model cannot learn subword patterns, making it harder to recognize rare names, acronyms, or transliterated entities.
    - Window-based context: Only immediate local context (±1 tokens) is captured. Long-range dependencies, such as full names split across multiple tokens or entities spanning multiple sentences, are missed.

2. **BiLSTM Architecture and Training:**
    - Only 3 epochs: The model shows clear learning progress (validation F1: 0.44 → 0.63 → 0.69), suggesting it needs more epochs to converge. A longer training schedule could improve performance.
    - No pretrained embeddings: Word embeddings are trained from scratch on CoNLL-2003. Pretrained embeddings (GloVe, FastText) or contextualized models (BERT) would provide much richer linguistic signals.

- No sequence constraint enforcement: BiLSTM predicts tags independently per token. Adding a CRF layer on top (BiLSTM-CRF) would enforce valid transitions and likely boost test F1.
- Hyperparameter sweep is small: Only 2 configs tested. Expanded search over hidden dimensions, dropout rates, and learning rates could yield better results.

3. **Dataset and Task Limitations:**

   - Domain specificity: CoNLL-2003 is largely news text. Models trained on it may struggle on other domains (social media, scientific papers, medical records).
   - Boundary ambiguity: Some entities are genuinely ambiguous. "United States" is both a location and a political organization, and ground-truth labels may be inconsistent across annotators.
   - Class imbalance: ORG and LOC entities are ~3x more frequent than MISC, leading to higher error rates on minority classes.
   - Limited entity types: CoNLL-2003 only covers 4 entity types. Fine-grained NER (distinguishing athlete, politician, scientist) requires additional data and richer models.

4. **Evaluation Metric:**

   - Entity-level F1 is strict: A single token boundary error breaks an entire multi-token entity (e.g., "New York" becomes "New" + "York" if the B-LOC/I-LOC boundary is wrong). This penalizes models that are close but not perfect.
   - No partial credit: Predicting O instead of B-LOC for "London" gives zero credit for identifying the entity at all.

5. **Model Scope:**

   - No external knowledge: Neither model uses gazetteers (lists of known locations, organizations, people). Incorporating such resources (e.g., list of world capitals) would improve recall on common entities.
   - No ensemble methods: Single models are evaluated in isolation. Ensembles or voting strategies could improve robustness.

Key Findings

1. **CRF remains competitive:** On CoNLL-2003, traditional CRF with careful feature engineering outperforms a basic BiLSTM 5-fold more often, demonstrating that feature selection and sequence-aware modeling are critical for NER.

2. **BiLSTM shows potential with more training:** The rapid improvement across epochs (44% → 63% → 69% F1) suggests that with 5-10 epochs and better initialization, BiLSTM could close the performance gap. However, the gap cannot be overcome without addressing data quality and embedding richness.

3. **Entity-specific errors reveal model weaknesses:**

   - ORG errors (665 total) dominate, primarily due to overlap with location names and organizational roles.
   - Both models struggle with rare/unseen names and domain-specific entities (MISC, sports references).
   - CRF excels at frequent entity types (PER, LOC) where POS signals are reliable.

4. **No silver bullet:** Neither model perfectly solves NER. Practical systems often combine multiple approaches: rule-based pre/post-processing, ensemble voting, domain-specific gazetteers, and pretrained contextual models.