NLP HW2 – Experimentation Report

Yash Dagade

February 26, 2025

1. Introduction

In this experiment, I aimed to improve the baseline Conditional Random Field (CRF) by adding additional features in the `Featurizer`. Specifically, I introduced prefix and suffix features of length 1-3 to capture morphological cues (e.g., capitalization and substring features). I also retained the original features such as current/previous/next word, case indicators, and digit indicators.

2. Approach and Motivation

It seems plausible to believe that short prefix and suffix features can help the CRF detect entity boundaries and types more effectively: for instance, named entities often share suffixes (e.g., *"-son"* or *"-ing"*) or prefix patterns (e.g., uppercase first letters in a *"Mc"* prefix for Irish names). By adding these substring features, the model can learn to generalize across words that share such patterns.

3. Variations Tried

I tried several variations, summarized in Table 1. Some of them did not work well, including an attempt to integrate (1) pretrained embeddings as features, and (2) domain-specific features (like a dictionary of known companies or persons). These did not yield improvements in F1 on the validation or hidden test set.

Method	**Validation F1**	**Hidden Test F1**	**Outcome**
Baseline CRF (no extra feats)	0.80–0.81	0.67	Baseline
Add large pretrained embeddings	0.79	0.64	Hurt Performance
Dictionary-based "Company/Person" feats	0.78	0.69	Minimal improvement
Add short prefixes/suffixes	**0.87–0.88**	**0.74**	Best so far

Table 1: Experiment summary and outcomes.

4. Discussion of Results

By adding prefix and suffix features (up to length 3), the CRF was better able to classify entities that share common word patterns. This improved the model's recall on proper names and certain organization abbreviations. The final CRF model reached a validation F1 up to 0.88 and a hidden test F1 of 0.74 :-)