Exploring neural architectures for NER

Vincent Billaut, Marc Thibault

CS224n, Stanford University of Stanford, 03/21/2018

Motivation

- NER with many classes is a complex task
- NER requires a fine understanding of context information
- We need to incorporate past and future dependencies
- A CRF implements a memory flow on labels

Dataset: CoNLL-2002

Labeled new articles:

“The protest comes on the eve of the annual conference
O O O O O O O O O O

of Britain’s ruling Labor Party in the southern
O B-geo O O B-org I-org O O O

English seaside resort of Brighton. “
B-gpe O O O B-geo

- 40, 000 train sentences,
- 4, 000 test sentences,
- 17 label classes,
- Training takes ∼ 1h on multi-CPU.

Base Models

- Words embedded with GloVe
- Model
  - LSTM w/ or w/o extra top layer(s)
  - Bi-LSTM w/ or w/o extra top layer
  - Stacked LSTMs
- Cross-Entropy loss on labels
- Evaluation with F_1 score

Comparing different architectures

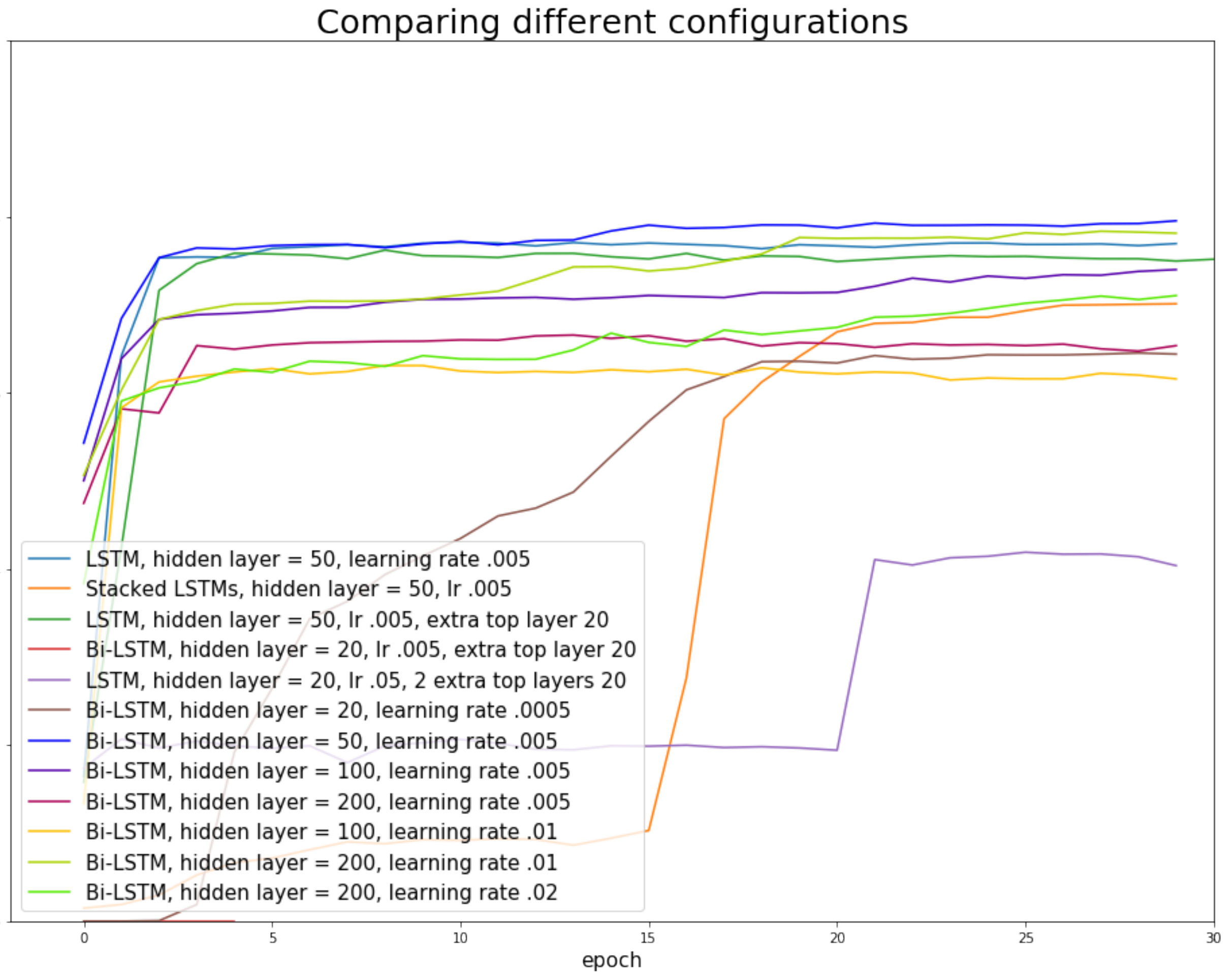

Figure 1: We have tried many different configurations for various types of architectures, and quickly concluded that the best performing architecture for our problem's specificity and scale was going to be a **Bi-directional LSTM** ; amongst them we chose the simplest one to give top results, namely one with a size-50 hidden layer

Dropout rate selection

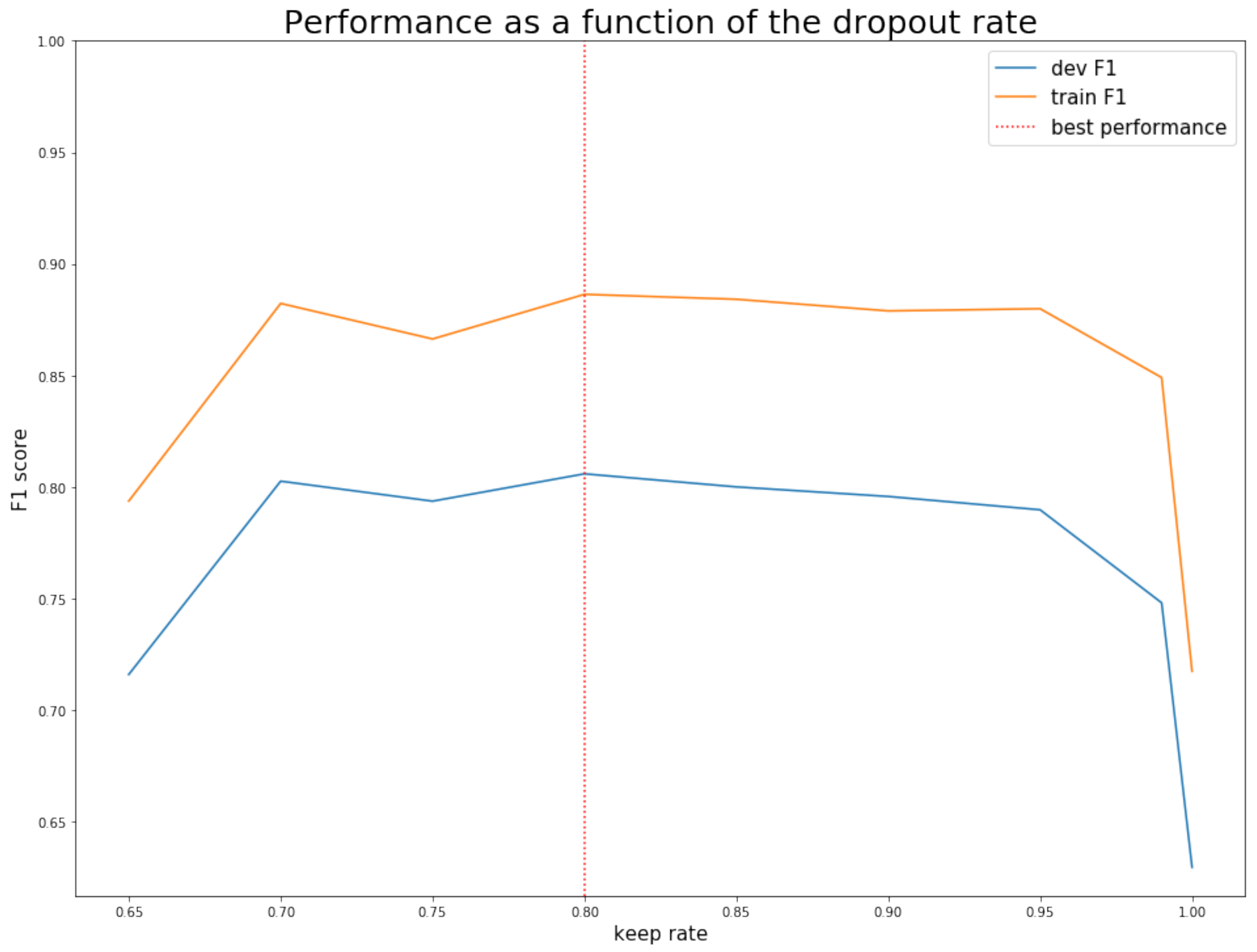

Figure 2: We see that the dropout rate has a significant effect on the model's performance, and we therefore kept a value of **0.2** moving forward.

Conditional Random Field

- Encodes a Markov Chain: probability of transitioning from a label to another:

$$P[i, j] = Pr(y_{t+1} = y_j | y_t = y_i) \text{ } in \text{ } training \text{ } set$$

P stochastic matrix of size (k, k).

- Adds information to the prediction:

$$z_0 = argmax_j(\hat{y}_0[j])$$

and

$$z_t = argmax_j(\hat{y}_t[j] + \alpha.P[z_{t-1}, j]) \quad \forall t \in [1, T]$$

α hyper-parameter which has to be fitted.

L2 regularization study

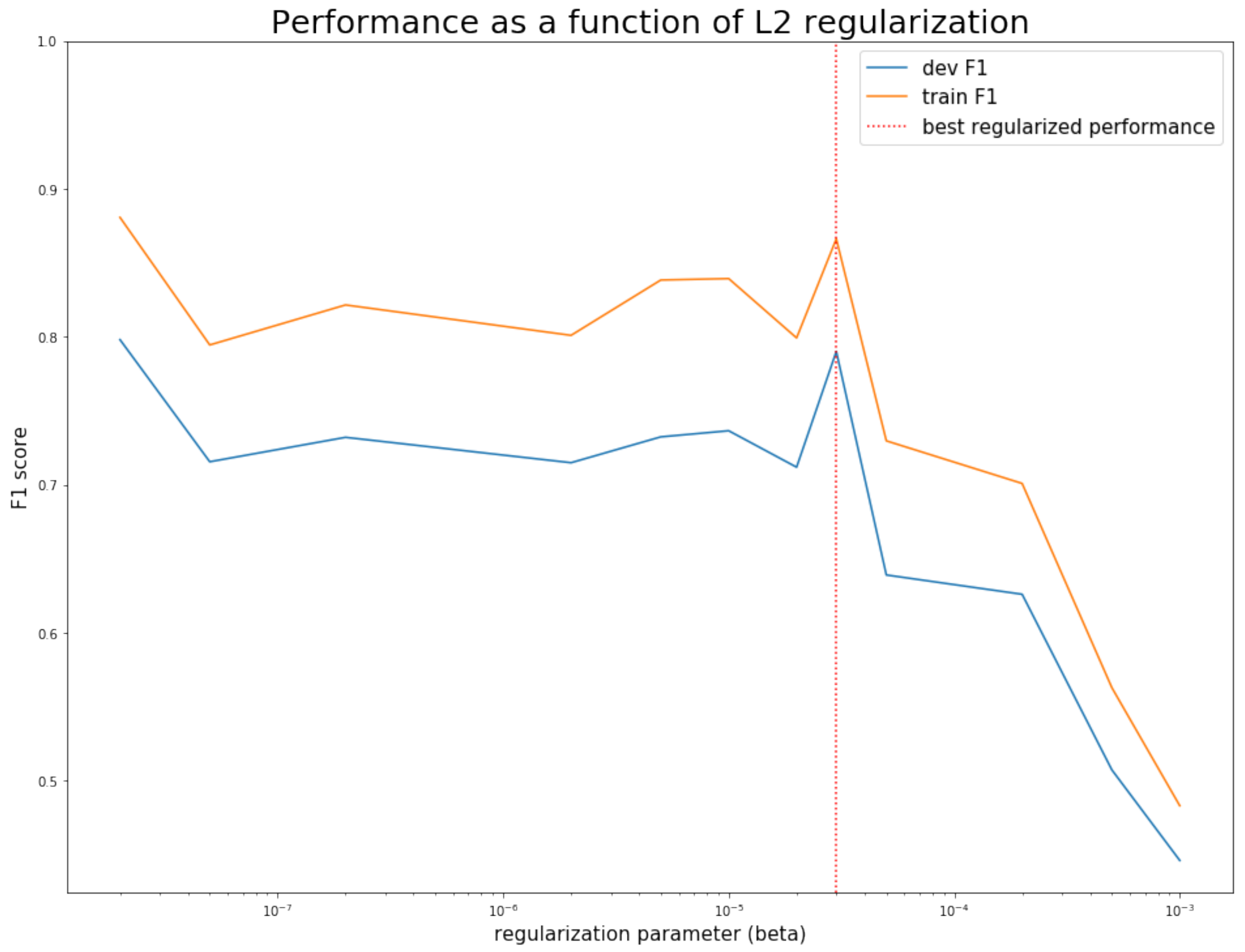

Figure 3: Surprisingly, no regularized model significantly beats the baseline in terms of F1 score on the dev set. Still, we consider that introducing some regularization is safer for robustness, and therefore keep an L2 regularization parameter $\beta = 3\ 10^{-5}$.

CRF performance

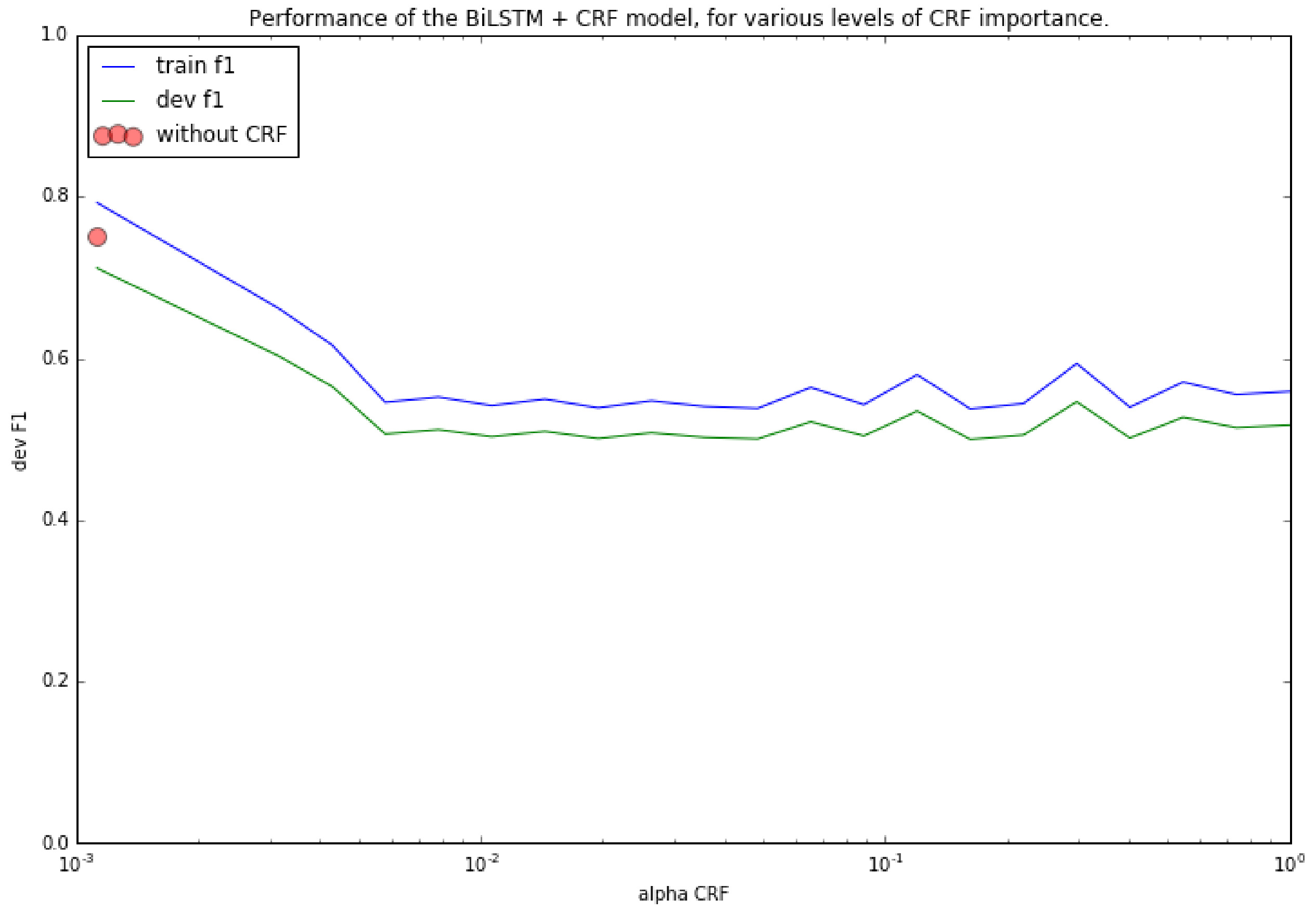

Figure 4: Train and Dev set performance of a BiLSTM model, incorporating a CRF on the prediction step, versus the importance α given to the CRF. No positive level of importance yields superior performance. The best decision is not to use the extra information from the CRF.

Results summary

In the simple way we incorporated CRF in the neural architecture, we could not get it to help a model perform significantly better than our well-tuned bi-LSTM.

Model	Dev F_1
Naive word-by-word	0.691
LSTM	0.597
LSTM + CRF	0.475
LSTM + extra layer	0.770
LSTM + extra layer + CRF	0.751
Bi-LSTM	0.796