Assignment 2: Sequence Labelling

Group 1
Shupei Li (s3430863) Qinshan Sun (s3674320)

1 Introduction

The sequence labelling task consists of four stages: preprocessing, feature engineering, model construction and hyperparameter tuning, training and evaluation. I perform the workflow of sequence labelling on W-NUT data in this assignment.

W-NUT is an open source data set specially for recognizing unusual and previously-unseen entities in noisy user-generated text. Its content was collected from online platforms, i.e. Twitter, Youtube, StackExchange, and Reddit. One row in original data files corresponds to either a word and its label or the end of line character. Table 1 shows the number of sentences and words in data. Besides, Figure 1 reports the count of different entity types in the training, evaluation, and test data set. According to Figure 1, "B-person" is the most common label in data.

Table 1: Statistics of W-NUT data

Data	#sentences	#words
Training set	3394	62370
Evaluation set	1009	15733
Test set	1287	23394

#: The number of items.

The rest of the report is organized as follows. Section 2 describes the details of the workflow. Section 3 is the experiment results. The report is concluded with a brief discussion.

2 Methodology

2.1 Preprocessing

W-NUT doesn't contain the POS tags compared to the data in the tutorial. I use Spacy package with `en_core_web_trf` pipline to add POS tags. `en_core_web_trf` is slower and larger than `en_core_web_sm` pipline, but has a higher accuracy. After preprocessing, one word is represented by a triple `(word, pos, biotag)`.

2.2 Feature Engineering

Baseline model applies the features in the tutorial directly, which contains properties of the word, POS tags, and -1/+1 context. In order to explore possible relevant features, I extend the features to -2/+2 context and -3/+3 context, while keep other features unchanged. Items in -2/+2 context and -3/+3 context are the same as -1/+1 context, except for the involved neighboring words. I also manually select three features, i.e. shape of the word, part of the stopword list, and lemma of the word, to improve the quality of labelling further. The shape and the lemma of a word provide additional semantic information that may be helpful to identify entities. Judging whether a word is a stopword is aiming at improving contextual relevancy and reducing the impact of noise in data. Models with different feature sets are evaluated. Table 2 summarizes feature sets used in experments.

Table 2: Summary of Feature Sets

Feature Set	Content
Baseline	Features from the tutorial
-2/+2 Context	Baseline + -2/+2 context features
-3/+3 Context	-2/+2 context + -3/+3 context features
Selection	-3/+3 context + three selected features

2.3 Hyperparameter Optimization

I use `CRF` API in `sklearn_crfsuite` package to construct models. Randomized search strategy is adopted to tune hyperparameters in the model ($c1$ and $c2$). The parameter space is defined by the exponential distribution $f(x) = \exp(-x)$, $x \geq 0$. During each search iteration, values are sampled from the parameter space and evaluated on the dev set with five-fold cross validation method. The number of maximum iterations is 50. Finally, hyperparameters are set to values that achieve best performance.

2.4 Training and Evaluation

L-BFGS method with the maximal 100 iterations is used to train CRF model. All experiments are deployed on a computer with an Intel(R) Core(TM) i7-10875H CPU. The next section reports the performance of models on three metrics: precision, recall, and F1-score.

3 Results

3.1 Baseline Run

Table 3 and Table 4 report results of the baseline run with and without hyperparameter optimization on the test set, respectively. Best hyperparameter values after tuning are $c1 = 0.015$, $c2 = 0.008$.

After hyperparameter optimization, the model performance enhances on all the metrics.

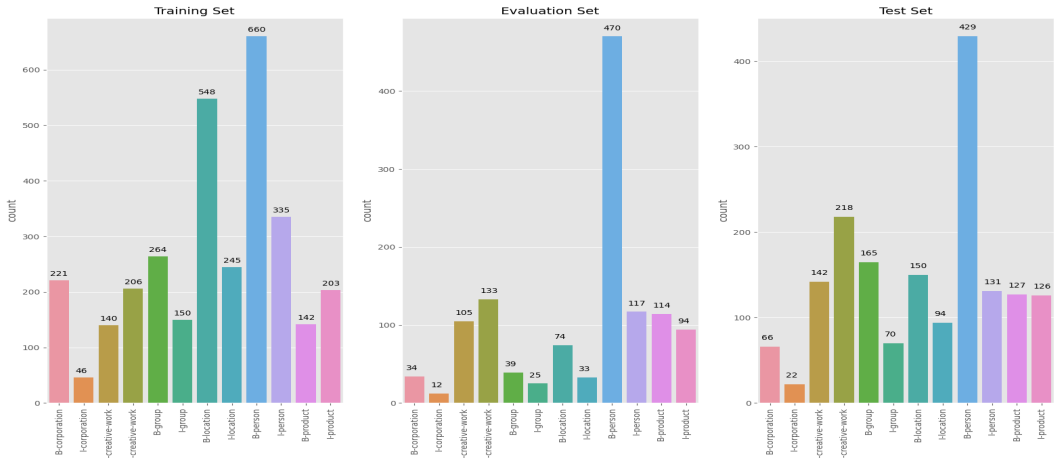

Figure 1: Count of Different Entities

Table 3: Baseline Run without Tuning

	Precision	Recall	F1-score
B-corporation	0.000	0.000	0.000
I-corporation	0.000	0.000	0.000
B-creative-work	0.462	0.042	0.077
I-creative-work	0.611	0.051	0.093
B-group	0.231	0.036	0.063
I-group	0.375	0.086	0.140
B-location	0.411	0.308	0.351
I-location	0.370	0.106	0.165
B-person	0.582	0.208	0.306
I-person	0.524	0.336	0.409
B-product	0.667	0.016	0.031
I-product	0.235	0.032	0.056
Weighted avg	0.455	0.129	0.181

Table 4: Baseline Run with Tuning

	Precision	Recall	F1-score
B-corporation	0.143	0.015	0.027
I-corporation	0.500	0.046	0.083
B-creative-work	0.546	0.042	0.078
I-creative-work	0.563	0.041	0.077
B-group	0.259	0.042	0.073
I-group	0.368	0.100	0.157
B-location	0.385	0.280	0.324
I-location	0.394	0.138	0.205
B-person	0.600	0.217	0.319
I-person	0.575	0.351	0.436
B-product	0.500	0.024	0.045
I-product	0.200	0.032	0.055
Weighted avg	0.463	0.133	0.189

3.2 Different Feature Sets

Hyperparameters in models with different feature sets are optimized on the dev set. Table 5 reports the best values for hyperparameters.

Table 5: Hyperparameter Settings

Feature Set	Hyperparameter	Value
-2/+2 Context	$c1$	0.003
	$c2$	0.004
-3/+3 Context	$c1$	0.100
	$c2$	0.100
Selection	$c1$	0.094
	$c2$	0.043

Table 6, 7, and 8 show the performance of models with settings in Table 5 on the test set.

Table 6: Model with -2/+2 Context Feature Set

	Precision	Recall	F1-score
B-corporation	0.000	0.000	0.000
I-corporation	0.000	0.000	0.000
B-creative-work	0.385	0.035	0.065
I-creative-work	0.450	0.041	0.076
B-group	0.222	0.036	0.063
I-group	0.286	0.057	0.095
B-location	0.467	0.287	0.355
I-location	0.432	0.170	0.244
B-person	0.622	0.226	0.332
I-person	0.631	0.405	0.493
B-product	0.667	0.032	0.060
I-product	0.625	0.040	0.075
Weighted avg	0.479	0.139	0.197

4 Discussion

Model with selection feature set achieves the best performance on all three metrics. Besides, model with -2/+2 context feature set performs better than baseline model, while model with -3/+3 context feature set improves slightly on

Table 7: Model with -3/+3 Context Feature Set

	Precision	**Recall**	**F1-score**
B-corporation	0.000	0.000	0.000
I-corporation	0.000	0.000	0.000
B-creative-work	0.539	0.049	0.090
I-creative-work	0.619	0.060	0.109
B-group	0.286	0.036	0.065
I-group	0.429	0.086	0.143
B-location	0.441	0.273	0.337
I-location	0.385	0.106	0.167
B-person	0.610	0.240	0.345
I-person	0.589	0.405	0.480
B-product	0.333	0.008	0.015
I-product	0.500	0.032	0.060
weighted avg	0.480	0.140	0.197

Table 8: Model with Selection Feature Set

	Precision	**Recall**	**F1-score**
B-corporation	0.000	0.000	0.000
I-corporation	0.000	0.000	0.000
B-creative-work	0.526	0.070	0.124
I-creative-work	0.421	0.073	0.125
B-group	0.333	0.055	0.094
I-group	0.348	0.114	0.172
B-location	0.431	0.313	0.363
I-location	0.333	0.117	0.173
B-person	0.604	0.256	0.360
I-person	0.581	0.382	0.461
B-product	1.000	0.008	0.016
I-product	0.667	0.032	0.061
Weighted avg	**0.511**	**0.153**	**0.211**

precision and recall. Results of experiments supports the effectiveness of context extension and feature selection. However, extending features to a larger context impairs model's performance on identifying corporation entity. The reason may be that most names of corporation have weak dependence on context. Context extension actually introduces noise in corporation identification.