Extraction of Drug-Drug Interactions from Biomedical Texts

Carlo Alessi & Doria Saric

May 31, 2018

Outline

Software

- Base language: Python
- XML parser: ElementTree
- Pandas
- Numpy
- Gensim: Word2Vec
- NLTK
  a. Word tokenizer
  b. Part-of-Speech tagger
  c. English stemmer
- Scikit Learn
  a. Data preprocessing
  b. Pipelines
  c. SVM
- Keras: ANN

Method

- Task9.1 – 2 approaches
  a. Embeddings: stemming, POS, puctuation removal, lowercase
  b. Hand-crafted features:
     - Is the word capitalized?
     - Is the part of speech NN?
     - Previous word, following word – word2vec
     - If there are some **trigger** words surrounding this word?
     - Does the word have more consonants than vocals?
- Task9.2
  - Hierarchical classification (binary + multiclass)

Outline

Learners

- Training/Validation split: $(90\%, 10\%) \sim (133k, 14k)$ words
- Metrics:

1. Feed-forward Neural Network
   - Architecture: $[vector\_size, 512, 256, 128, 3]$
   - Activations: ReLU, Softmax
   - Dropout fraction: 0.5
   - Objective: Categorical cross-entropy
   - Solver: Adam
   - Number of parameters: 175363
2. Support Vector Machine
   - Kernel: radial basis function
   - $C = 1$
   - $\gamma = auto$
   - $\alpha = 0.001$

Outline

Word Vectors

- Vector size
- Vector type
- Preprocessing

Vector size	Micro F1	Macro F1
10	0.9481	0.5041
20	**0.9517**	**0.5435**
50	0.9495	0.5215
100	0.9499	0.5232
200	0.9479	0.4855

Table: F1-score for different word-vector sizes.

Type	Micro f1	Macro f1
original	0.9499	0.5098
stem	**0.9599**	**0.6144**
original + PoS	0.9481	0.5105
stem + PoS	0.9517	0.5435

Table: F1-score for different word-vector types.

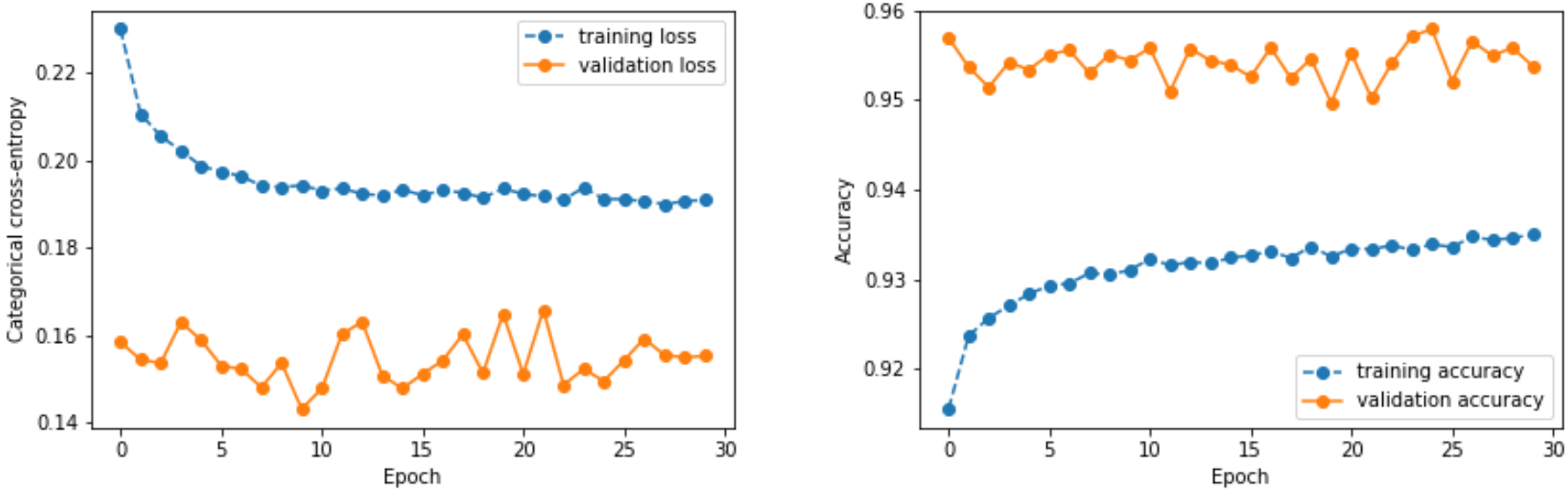
Word Vectors

- Vector size
- Vector type
- Preprocessing

Preprocessing	Micro f1	Macro f1
No	0.9599	0.6144
Minmax(0,1)	0.9515	0.5207
Standardize	0.9581	0.5897

Table: F1-score for different preprocessing.

Loss/Accuracy



(a) loss

(b) accuracy

Figure: Learning curves

Micro F1/Precision/Recall score

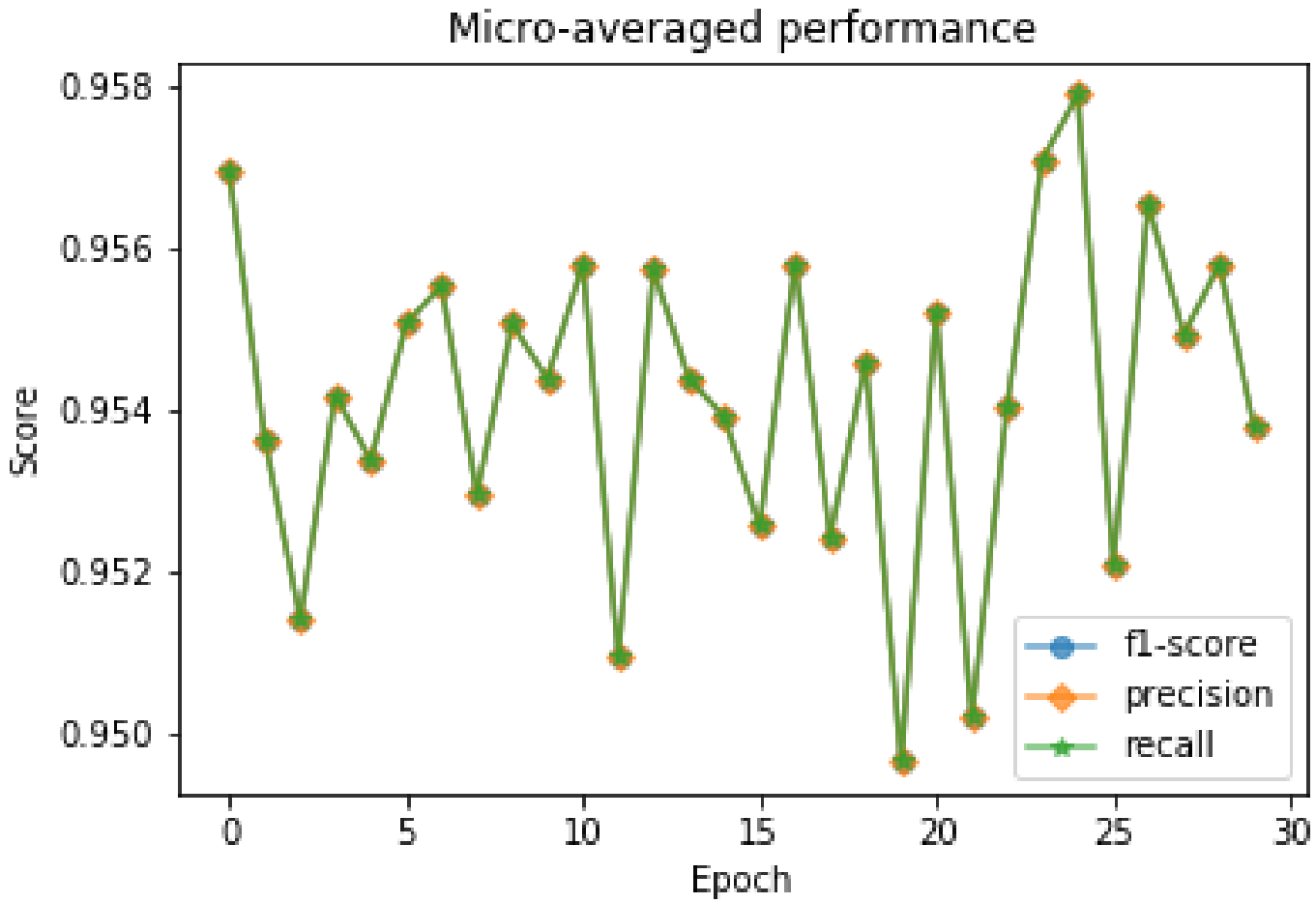
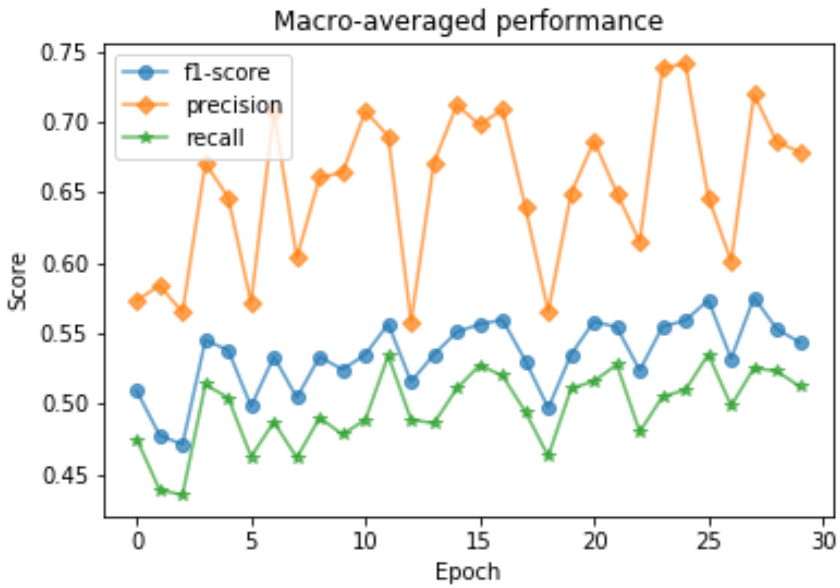
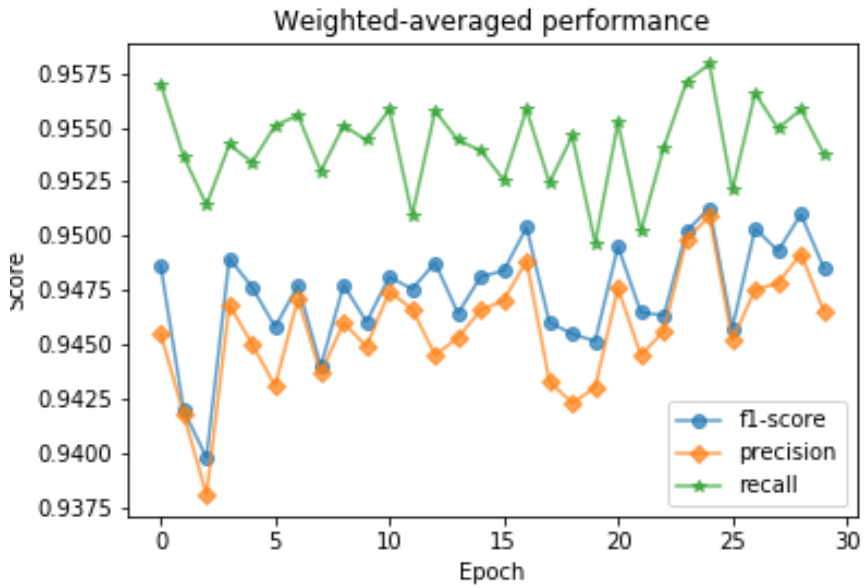


Figure: Micro scores.

Macro/Weighted F1/Precision/Recall score



(a) loss



(b) accuracy

Figure: Macro and Weighted scores.

Outline

Results

	Exact			Partial		
	Precision	Recall	F1	Precision	Recall	F1
DrugBank	0.61	0.43	0.5	0.61	0.5	0.55
MedLine	0.51	0.29	0.37	0.51	0.35	0.41
Both	0.56	0.35	0.43	0.56	0.41	0.48

Table: Results Task1 on gold test dataset.

Outline

Conclusions

- Word vector size (too small, too big)
- stemming and PoS improved performance used individually, but not in conjuction
- Poor results (especially the recall)
  - Poor features?
  - Limited model capacity/expressivess?