Biomedical Named Entity Recognition

ABIDAR Bouchra
BOUSSEBAINE Mustapha
LARBI Abderrahman

Supervised by :
AFFELDT Severine
LABIOD Lazhar

May 29, 2020

UNIVERSITÉ PARIS DESCARTES

Université de Paris

Presentation Outline

1 Experimental study

- Conditional Random Fields
- BLSTM-CNNs-CRF
- Flair : BERT based
- Spacy

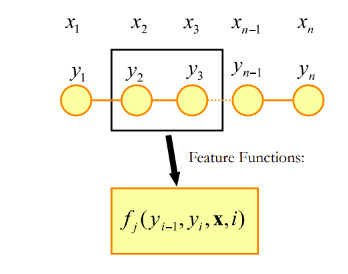
2 Generalized model

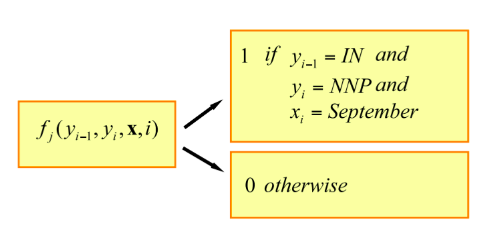
3 Conclusion

4 References

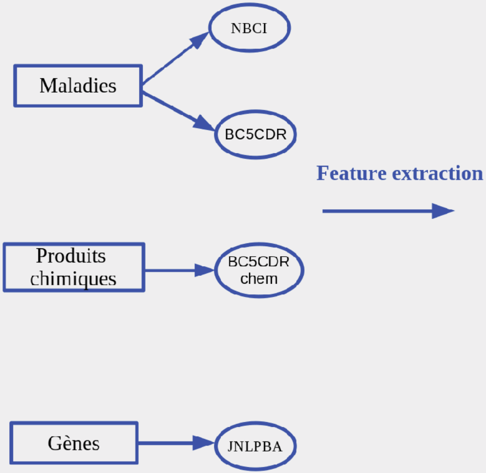
Conditional Random Fields : Technical aspect

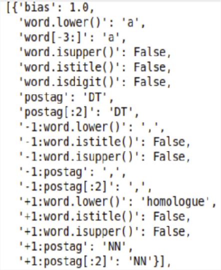
- Discriminant model
- Sequential data prediction
- Features function




$f_j(y_{i-1}, y_i, \mathbf{x}, i)$
1 *if* $y_{i-1} = IN$ *and*
$y_i = NNP$ *and*
$x_i = September$
0 *otherwise*

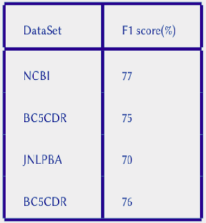
Conditional Random Fields : Application





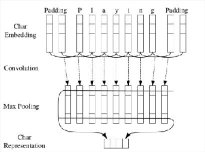
```
[{'bias': 1.0,
  'word.lower()': 'a',
  'word[-3:]': 'a',
  'word.isupper()': False,
  'word.istitle()': False,
  'word.isdigit()': False,
  'postag': 'DT',
  'postag[:2]': 'DT',
  '-1:word.lower()': ',',
  '-1:word.istitle()': False,
  '-1:word.isupper()': False,
  '-1:postag': ',',
  '-1:postag[:2]': ',',
  '+1:word.lower()': 'homologue',
  '+1:word.istitle()': False,
  '+1:word.isupper()': False,
  '+1:postag': 'NN',
  '+1:postag[:2]': 'NN'}],
```

CRF

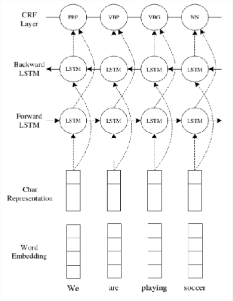


DataSet	F1 score(%)
NCBI	77
BC5CDR	75
JNLPBA	70
BC5CDR	76

BLSTM-CNNs-CRF



La couche CNN



Modéle BiLSTM-CNN-CRF

2

DataSet	F1 Score(%)
NCBI	70
BC5CDR	65
JNLPBA	-
BC5CDR-chem	60

Score de Performance

- CNN model for generating character embedding.
- Bi-directional LSTM for Word-Level Encoding.
- Conditional Random Fields(CRF) for output decoding

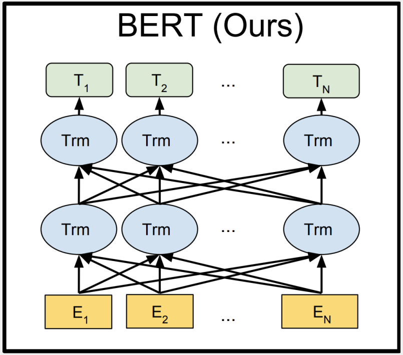
BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

- Pre-training deep bidirectional representation models
- Fine tuned with just a single additional layer to produce state of the art results
- Applied on a wide range of NLP STOA Tasks.

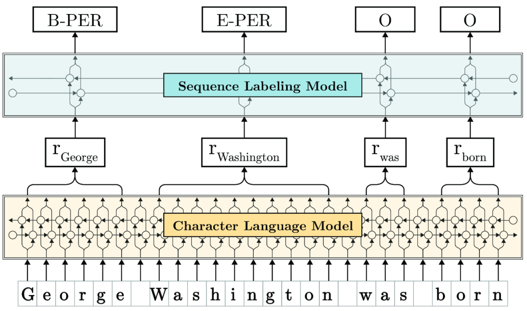
BERT : How it works

- Based on Transformers
- Uses sub word tokenization
  "Here is the sentence I want embeddings for."
  ['[CLS]', 'here', 'is', 'the', 'sentence', 'i', 'want', 'em', '##bed', '##ding', '##s', 'for', '.', '[SEP]']
- Masked LM (MLM)
- Nb : In Named Entity Recognition : The output of the transformer bloc passes by a classification last layer to assign the tag.

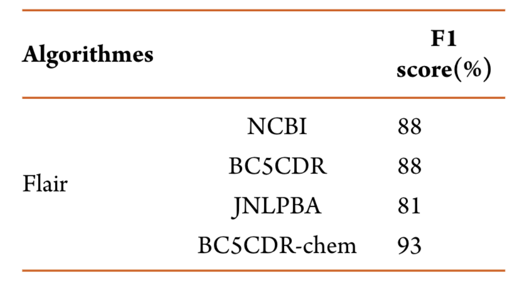
BERT : Architecture



Flair : Architecture & Results



Flair Architecture



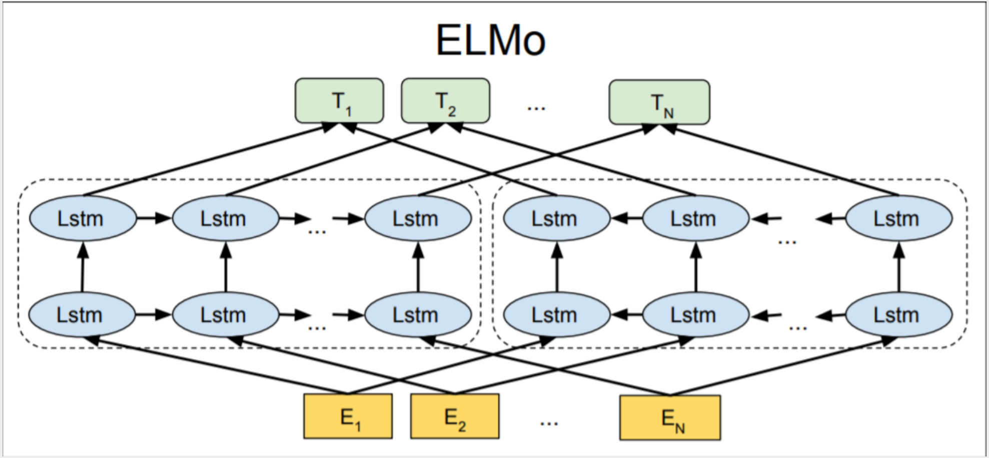
Algorithmes		F1 score(%)
Flair	NCBI	88
	BC5CDR	88
	JNLPBA	81
	BC5CDR-chem	93

Flair's results

Spacy

- spaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python.
- spaCy is designed specifically for production use
- spaCy is not research software.
- spaCy can be used to build information extraction or natural language understanding, or to pre-process text for deep learning.
- spaCy has many features and capabilities.
- spaCy is based on ELMo architecture

ELMo : how does it work



Training spaCy NER & results

- in order to train spaCy's NER model we must transform the training data to JSON format. for exemple : ("Who is keyser soze?", "entities": [(7, 17, "PERSON")])

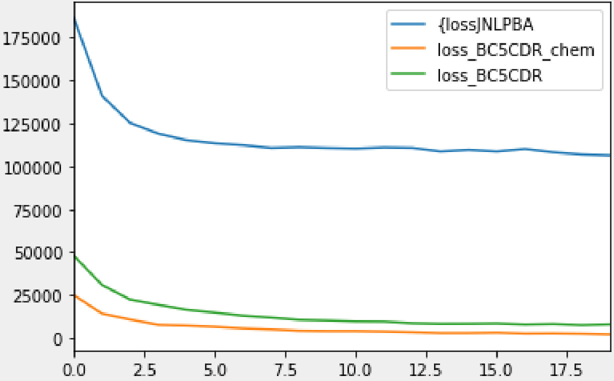


Figure 1: spaCy Loss

Presentation Outline

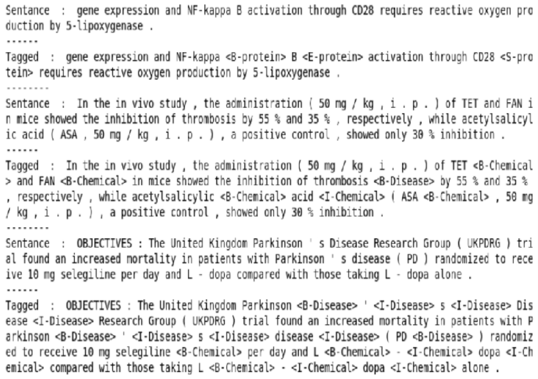
1 Experimental study

2 Generalized model

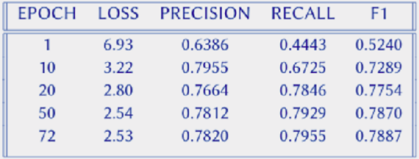
3 Conclusion

4 References

Generalized model



```
Sentance  :  gene expression and NF-kappa B activation through CD28 requires reactive oxygen pro
duction by 5-lipoxygenase .
------
Tagged  :  gene expression and NF-kappa <B-protein> B <E-protein> activation through CD28 <S-pro
tein> requires reactive oxygen production by 5-lipoxygenase .
--------
Sentance  :  In the in vivo study , the administration ( 50 mg / kg , i . p . ) of TET and FAN i
n mice showed the inhibition of thrombosis by 55 % and 35 % , respectively , while acetylsalicyl
ic acid ( ASA , 50 mg / kg , i . p . ) , a positive control , showed only 30 % inhibition .
------
Tagged  :  In the in vivo study , the administration ( 50 mg / kg , i . p . ) of TET <B-Chemical
> and FAN <B-Chemical> in mice showed the inhibition of thrombosis <B-Disease> by 55 % and 35 %
, respectively , while acetylsalicylic <B-Chemical> acid <I-Chemical> ( ASA <B-Chemical> , 50 mg
/ kg , i . p . ) , a positive control , showed only 30 % inhibition .
--------
Sentance  :  OBJECTIVES : The United Kingdom Parkinson ' s Disease Research Group ( UKPDRG ) tri
al found an increased mortality in patients with Parkinson ' s disease ( PD ) randomized to rece
ive 10 mg selegiline per day and L - dopa compared with those taking L - dopa alone .
------
Tagged  :  OBJECTIVES : The United Kingdom Parkinson <B-Disease> ' <I-Disease> s <I-Disease> Dis
ease <I-Disease> Research Group ( UKPDRG ) trial found an increased mortality in patients with P
arkinson <B-Disease> ' <I-Disease> s <I-Disease> disease <I-Disease> ( PD <B-Disease> ) randomiz
ed to receive 10 mg selegiline <B-Chemical> per day and L <B-Chemical> - <I-Chemical> dopa <I-Ch
emical> compared with those taking L <B-Chemical> - <I-Chemical> dopa <I-Chemical> alone .
```



EPOCH	LOSS	PRECISION	RECALL	F1
1	6.93	0.6386	0.4443	0.5240
10	3.22	0.7955	0.6725	0.7289
20	2.80	0.7664	0.7846	0.7754
50	2.54	0.7812	0.7929	0.7870
72	2.53	0.7820	0.7955	0.7887

Presentation Outline

1 Experimental study

2 Generalized model

3 Conclusion

4 References

Conclusion : Improvement perspectives

1- continuous learning (CL)

- The model (CL) takes a tagger as input
- The data entry passes by the three models CRF BERT and SPACY
- Each model yields its results
- The CL model selects the best model on that entry
- The CL starts the retraining with the output dag for the models that yielded the wrong tag.

Presentation Outline

1 Experimental study

2 Generalized model

3 Conclusion

4 References

Reference

- John Lafferty, Andrew McCallum, Fernando Pereira *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001), 2001
- Xuezhe Ma and Eduard Hovy *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF*,2016
- Eliyahu Kiperwasser,Yoav Goldberg *Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations*,2016
- DEVLIN, Jacob, CHANG, Ming-Wei, LEE, Kenton, et al. *Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805* , 2018.

Thank You!

Questions?