Data set description :

Annotated Corpus for Named Entity Recognition using GMB(Groningen Meaning Bank) corpus for entity classification with enhanced and popular features by Natural Language Processing applied to the data set.

Link : <https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus#ner_dataset.csv>

Total Words Count = 1354149 Target Data Column: "tag"

'O': 1146068', geo-nam': 58388, 'org-nam': 48034, 'per-nam': 23790, 'gpe-nam': 20680, 'tim-dat': 12786, 'tim-dow': 11404, 'per-tit': 9800, 'per-fam': 8152, 'tim-yoc': 5290, 'tim-moy': 4262, 'per-giv': 2413, 'tim-clo': 891, 'art-nam': 866, 'eve-nam': 602, 'nat-nam': 300, 'tim-nam': 146, 'eve-ord': 107, 'per-ini': 60, 'org-leg': 60, 'per-ord': 38, 'tim-dom': 10, 'per-mid': 1, 'art-add': 1

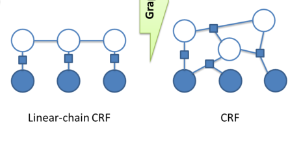
Essential info about entities:

* geo = Geographical Entity
* org = Organization
* per = Person
* gpe = Geopolitical Entity
* tim = Time indicator
* art = Artifact
* eve = Event
* nat = Natural Phenomenon

Machine Learning :

CRF :

CRFs fall into the sequence modeling family. Whereas a discrete [classifier](https://en.wikipedia.org/wiki/Statistical_classification) predicts a label for a single sample without considering "neighboring" samples, a CRF can take context into account; e.g., the linear chain CRF (which is popular in [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing)) predicts sequences of labels for sequences of input samples.



# The best algorithm CRF (Conditional random field)

RESULT :

precision recall f1-score support

B-art 0.00 0.00 0.00 402

B-eve 0.80 0.27 0.40 308

B-geo 0.82 0.90 0.86 37644

B-gpe 0.95 0.92 0.94 15870

B-nat 0.69 0.09 0.16 201

B-org 0.78 0.67 0.72 20143

B-per 0.80 0.76 0.78 16990

B-tim 0.93 0.83 0.88 20333

I-art 0.00 0.00 0.00 297

I-eve 0.64 0.12 0.20 253

I-geo 0.81 0.73 0.77 7414

I-gpe 0.93 0.37 0.53 198

I-nat 0.00 0.00 0.00 51

I-org 0.75 0.76 0.75 16784

I-per 0.80 0.90 0.85 17251

I-tim 0.84 0.67 0.74 6528

O 0.99 0.99 0.99 887908

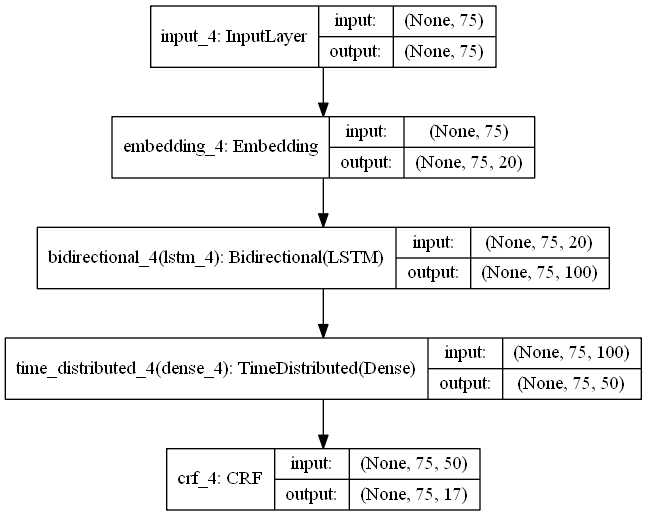
micro avg 0.97 0.97 0.97 1048575

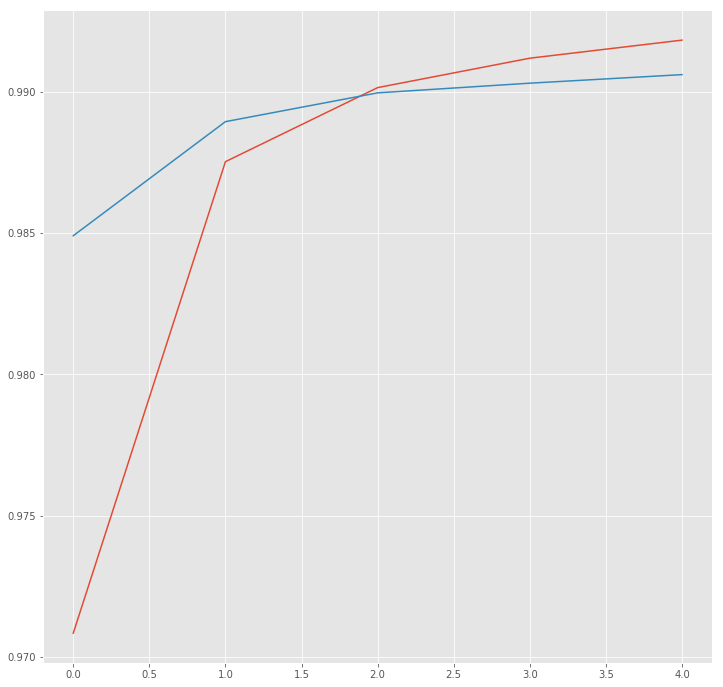
macro avg 0.68 0.53 0.56 1048575

weighted avg 0.96 0.97 0.96 1048575

**DEEP LEARNING : LSTM with CRF:**

An LSTM has a similar control flow as a recurrent neural network. It processes data passing on information as it propagates forward. The differences are the operations within the LSTM's cells. These operations are used to allow the LSTM to keep or forget information





precision recall f1-score support

B-art 0.00 0.00 0.00 41

B-eve 0.44 0.13 0.20 31

B-geo 0.81 0.91 0.86 3772

B-gpe 0.97 0.93 0.95 1694

B-nat 0.11 0.10 0.11 10

B-org 0.81 0.65 0.72 2045

B-per 0.85 0.82 0.83 1713

B-tim 0.91 0.88 0.89 2011

I-art 0.00 0.00 0.00 31

I-eve 0.20 0.07 0.11 27

I-geo 0.72 0.79 0.75 721

I-gpe 1.00 0.63 0.77 19

I-org 0.83 0.73 0.77 1670

I-per 0.84 0.91 0.87 1771

I-tim 0.87 0.73 0.80 665

O 1.00 1.00 1.00 343479

micro avg 0.99 0.99 0.99 359700

macro avg 0.65 0.58 0.60 359700

weighted avg 0.99 0.99 0.99 359700

---------------------------------------------------------------------------------------------------------------------------------

NLTK :

OutPut:

[('writing', 'VERB', 'write'),

('texts', 'NOUN', 'text'),

('Mr.', 'PROPN', 'Mr.'),

('Strzok', 'PROPN', 'Strzok'),

('accused', 'VERB', 'accuse'),

('sending', 'VERB', 'send'),

('highly', 'ADV', 'highly'),

('sensitive', 'ADJ', 'sensitive'),

('search', 'NOUN', 'search'),

('warrant', 'NOUN', 'warrant'),

('personal', 'ADJ', 'personal'),

('email', 'NOUN', 'email'),

('account', 'NOUN', 'account')]