Damoun Ayman
Github repository
February 14, 2021

Named Entity Recognition assignement

The objective of this lab session is to build and compare different models of named entity recognition using the word embeddings. We will use the **biLSTM** model implemented by Ma and Hovy [2].

1 Introduction and recall

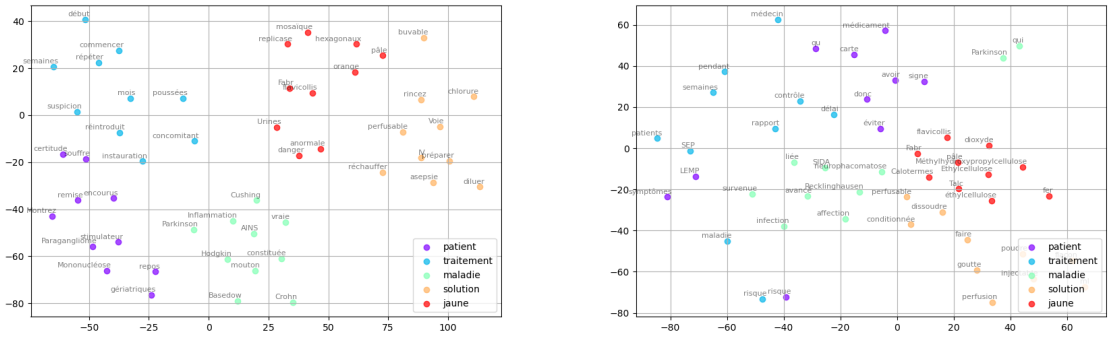
In the previous work we were interested in the Word embedding, so let's recall the principle behind. Word embedding is a learned representation of a word in which each word is represented using a vector in an n-dimensional space. In this approach, Words with the same meaning have the same representation. These representations make it possible to identify synonyms, antonyms and other relationships between words. The idea behind Word embedding can be generalized to develop embeddings for individual sentences, documents, and so on.

word2vec: word2vec is the deep learning Google framework to train word embeddings. This framework use all the words of the corpus to predict the neighboring words. The word2vec algorithm create a vector for all the words present in the data in a way that the context is captured. Word2Vec is an iterative method. Its main idea is as follows [5]:

- take all corpus;
- move through the text in a sliding window, moving one word at a time. At each step, there is a central word and context words ;
- for the central word, compute probabilities of context words;
- adjust the vectors to increase these probabilities.

word2vec is mainly compensated by:

- Skip-Gram
- Continuous Bag of Words (CBOW)



(a) Skipgram

(b) Cbow

Figure 1: Visualizing Word Embeddings (French Medical Corpus)

Question 1
Create descriptive statistics on the corpus: number of words (tokens), sentences, entities of each type.

solution:

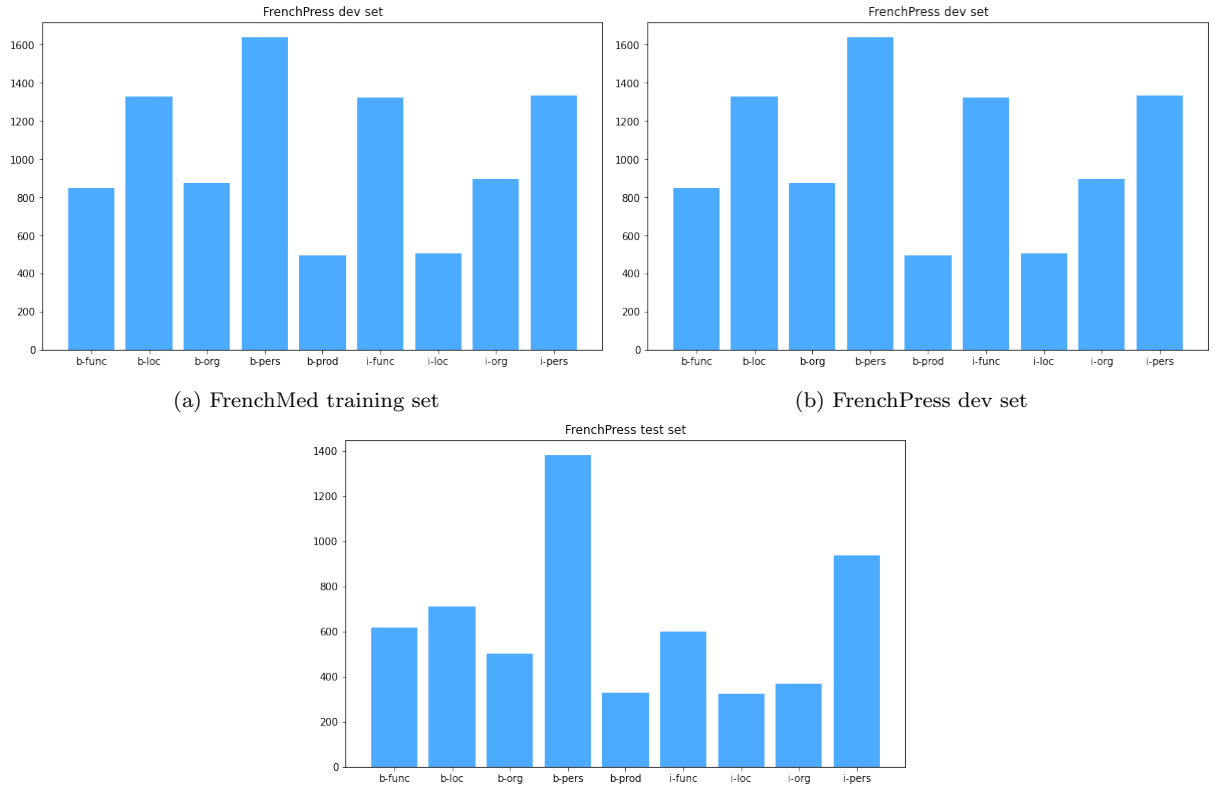
About the dataset

The QUAERO French Medical Corpus: The QUAERO French Medical Corpus has been initially developed as a resource for named entity recognition and normalization [3]. It was then improved with the purpose of creating a gold standard set of normalized entities for French biomedical text, that was used in the CLEF eHealth evaluation lab. It is a complete corpus, tokenized and with one sentence per line.

The QUAERO French Press Corpus: It is a complete corpus, tokenized and with one sentence per line.
The following table and figures summarizes some statistics on the FrenchPress and frenchMed corpus:

	FrenchPress training set	FrenchPress dev set	FrenchPress test set	FrenchMed training set	FrenchMed dev set	FrenchMed test set
Words	1156339	95222	95807	15339	13543	12388
sentences	35723	2825	2880	706	649	578

Table 1: Statistics on the FrenchPress and frenchMed corpus

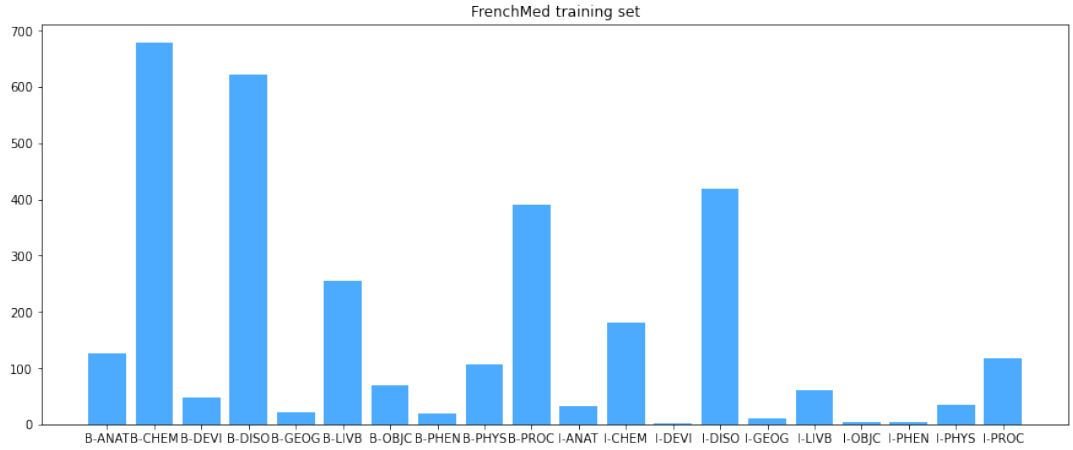


(a) FrenchMed training set

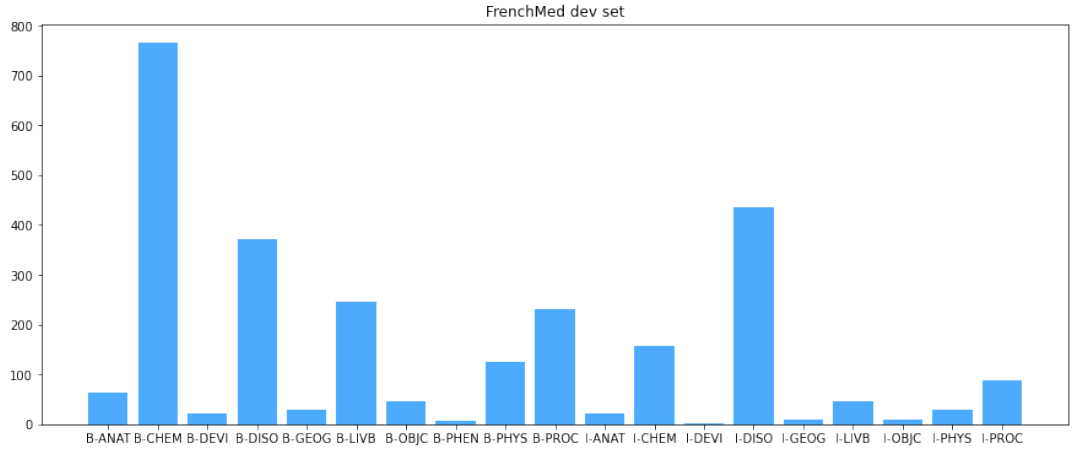
(b) FrenchPress dev set

(c) FrenchPress test set

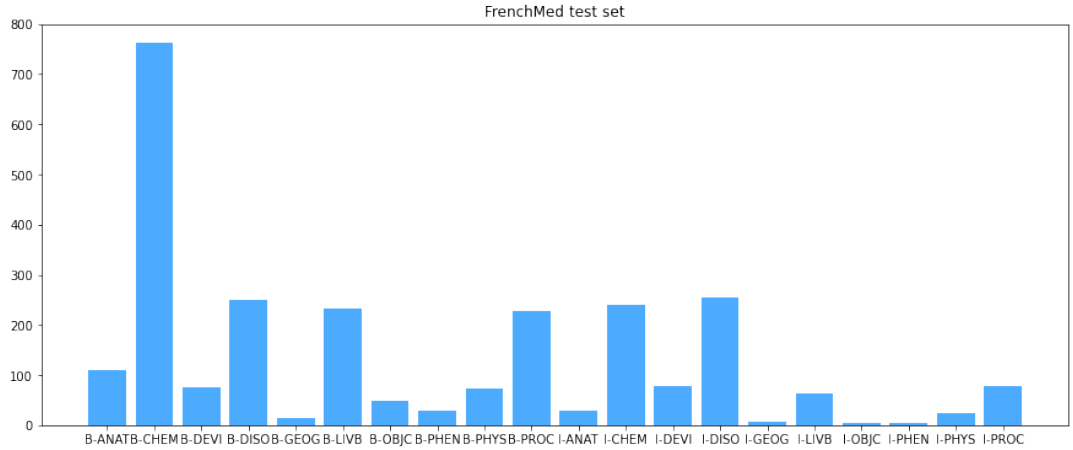
Figure 2: Visualizing entities statistics (French Press Corpus)



(a) FrenchMed training set



(b) FrenchMed dev set



(c) FrenchMed test set

Figure 3: Visualizing entities statistics (French Medical Corpus)

2 Named Entity Recognition (NER)

Named Entity Recognition (NER), also known as entity identification, which is a subtask of Information Extraction, can be defined as the process of determining whether a word or word-group represents a place, organization, or anything else. In other words, is the task of identifying information units in text into predefined units such as location, the name of a person, and so on. This can be broken down into two subtasks: identifying the boundaries of the named entity, and

identifying its type.

Recurrent neural network

To understand an LSTM network, we must first understand a recurrent neural network.
A recurrent neural network (RNN) is a type of artificial neural network that uses sequential or time series data. In Recurrent Neural Networks (or RNNs) the information can be propagated in both directions, including from the deep layers to the first layers.
A recurrent neural network can be considered as several copies of the same network, each transmitting a message to a successor (4).

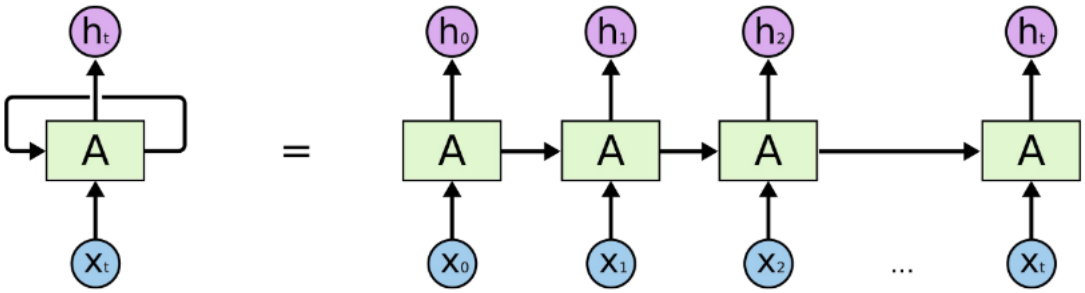


Figure 4: Recurrent Neural Networks (RNNs) [4]

LSTM

The LSTM network was proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997. The idea associated with LSTM is that each computational unit is not only related to a hidden state h but also to a state c of the cell that plays the role of memory. Intuitively, an LSTM can be seen as a neuron having, in addition to the external connections, a self-recurring connection with a constant coefficient equal to 1. This allows to save the successive states of the neuron from one iteration to another.

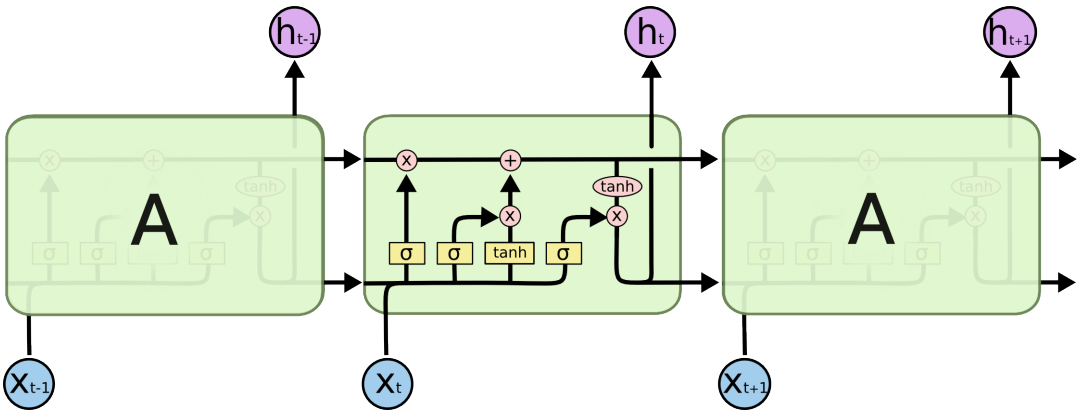
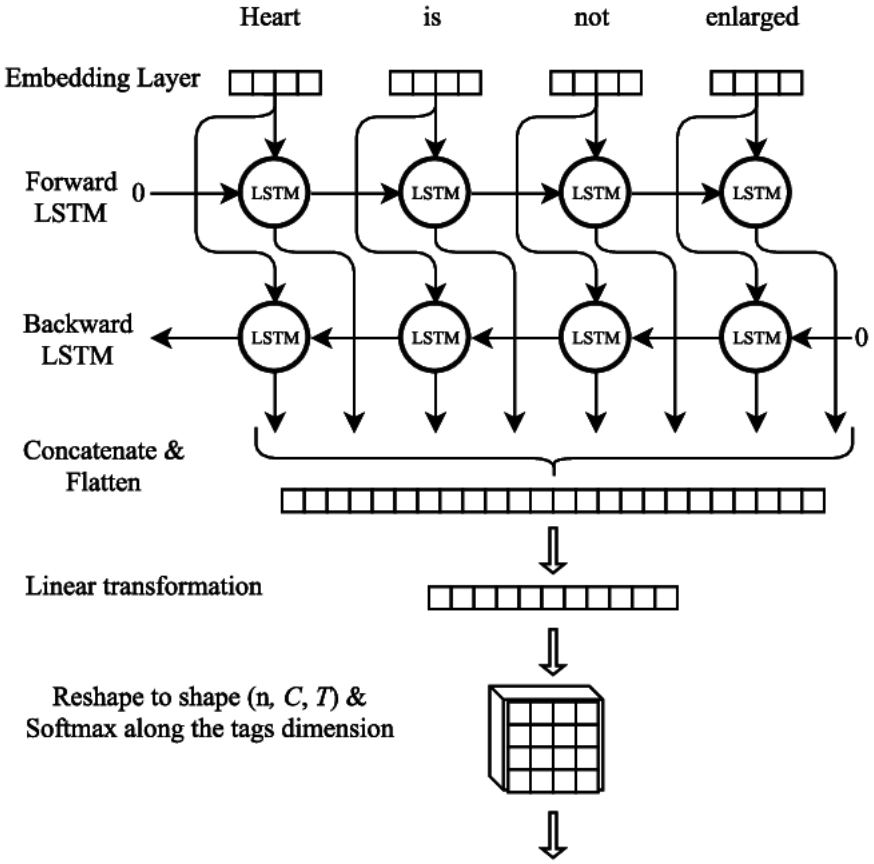


Figure 5: The repeating module in an LSTM contains four interacting layers [4]

biLSTM

It is interesting to have a memory of the past in order to make good decisions at time t. But in some recognition problems, it is sometimes interesting to look at future observations, when they are available.this is the idea behind Bidirectional LSTM. A Bidirectional LSTM(biLSTM), is a sequence

processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backwards direction. The use of this type of architecture is not possible in the case of applications running in real time.



	Heart	is	not	enlarged
Body Location	S	O	O	O
Medical Device	O	O	O	O
Clinical Finding	B	O	O	E
Descriptor	O	O	O	S
Negation	B	O	O	E

Figure 6: An illustration of the BiLSTM architecture for joint medical entity recognition and negation detection [1]

3 Experiments on NER

Question 2
Compare different named entity recognition models that use word embeddings

Based on the code provided in NeuroNLP2, i create a shell scripts to train the tool in different configurations, for example:

```
{
  "crf": false,
  "bigram": true,
  "embedd_dim": 100,
  "char_dim": 30,
  "rnn_mode": "LSTM",
  "num_layers":1,
  "hidden_size": 128,
  "out_features": 64,
  "dropout": "std",
  "p_in": 0.33,
  "p_out": 0.5,
  "p_rnn": [0.33, 0.5],
  "activation": "elu"
}
```

Listing 1: JSON file example

Command Line

```
#!/usr/bin/env bash
CUDA_VISIBLE_DEVICES=0 OMP_NUM_THREADS=60 python3 ner.py
--config ../data/quaero-100-demi.json --num_epochs 20 --batch_size 16 \
 --loss_type sentence --optim sgd --learning_rate 0.01 --lr_decay 0.99999
 --grad_clip 0.0 --warmup_steps 10 --weight_decay 0.0 --unk_replace 0.0 \
 --embedding sskip --embedding_dict "../data/QUAERO_FrenchPress-w2v.vec.gz"
 --model_path "../data/ner-fra4_ID-w2v-conll17-half" \
 --train "../data/QUAERO_FrenchPress/fra4_ID.train"
 --dev "../data/QUAERO_FrenchPress/fra4_ID.dev"
 --test "../data/QUAERO_FrenchPress/fra4_ID.test"
```

In this shell scripts we can adjust the parameters concerning to the type of embedding, the file contains the vector of embeddings, embedding-dict, the desired output directory, –model path, the corpus –train –dev –test. On the JSON file we can adjust certain parameters in the model configuration. For instance decrease the number of epoch –num.

	Epoch 10	Epoch 15	Epoch 20
lr	0.008196	0.007669	0.006866
loss	5.4184	5.1913	4.8375
dev acc	96.48%	96.59%	96.67%
dev precision	71.57%	71.58%	72.90%
dev recall	69.25%	71.48%	72.49%
dev F1	70.39%	71.53%	72.70%
test acc	97.27%	97.37%	97.41%
test precision	70.08%	71.77%	72.52%
test recall	62.65%	65.97%	66.62%
test F1	66.16%	68.75%	69.45%

The chunk and named entity labels are in I-TYPE format which indicates that the word is part of a TYPE entity. If two entities of the same type follow each other, the first word of the second

```
25 monsieur dummy dummy b-pers b-pers
26 Europe dummy dummy i-pers i-prod
27 , dummy dummy O O
28 c' dummy dummy O O
29 est dummy dummy O O
30 Alex dummy dummy b-pers b-pers
31 Taylor dummy dummy i-pers i-pers
32 , dummy dummy O O
33 qui dummy dummy O O
34 publie dummy dummy O O
35 Bouche dummy dummy b-prod O
36 <_UNK> dummy dummy i-prod O
37 , dummy dummy i-prod O
38 <_UNK> dummy dummy i-prod O
39 toute dummy dummy i-prod O
40 <_UNK> dummy dummy i-prod O
41 , dummy dummy i-prod O
42 ou dummy dummy i-prod O
43 comment dummy dummy i-prod O
44 tomber dummy dummy i-prod O
45 amoureux dummy dummy i-prod O
46 des dummy dummy i-prod O
47 langues dummy dummy i-prod O
```

Listing 2: Prediction on test set QUAERO FrenchPress

will have a B-TYPE label to indicate that it is a new entity. The label O indicates that the word is not part of an entity.

For the same nember of epochs (20) we get the results in the table 2:

	QUAERO French Press			QUAERO French Medical		
	skipgram	Cbow	FastText	skipgram	Cbow	FastText
lr	0.006569	0.006569	0.006866	0.009908	0.009918	0.009908
loss	4.3756	6.7739	7.6158	8.5539	9.4780	8.5539
dev acc	96.56%	96.17%	95.54%	87.48%	87.02%	87.48%
dev precision	71.70%	66.78%	60.48%	66.51%	61.59%	66.51%
dev recall	71.97%	68.71%	62.79%	43.88%	41.42%	43.88%
dev F1	71.84%	67.73%	61.62%	52.88%	49.53%	52.88%
test acc	97.36%	96.76%	96.83%	85.58%	84.78%	85.58%
test precision	71.76%	59.71%	60.49%	62.12%	58.75%	62.12%
test recall	65.88%	62.71%	57.91%	35.02%	32.05%	35.02%
test F1	68.70%	61.17%	59.17%	44.79%	41.48%	44.79%

Table 2: The results of different named entity recognition models

$$F_1 \text{-score} = \frac{1}{\frac{1}{2}\left(\frac{1}{\text{recall}} + \frac{1}{\text{precision}}\right)}$$

According to F_1-score for QUAERO French Press the skipgram model outperform Cbow and FastText with a score of 68.70%. On the other hand for small datasets like QUAERO French Medical i can't notice a big difference for 20 epochs.

Bibliography

[1] Savelie Cornegruta et al. "Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks". In: Jan. 2016, pp. 17–27. DOI: 10.18653/v1/W16-6103.

[2] Ma and Hovy. *NeuroNLP2*. 2019. URL: https://github.com/XuezheMax/NeuroNLP2.

[3] Aurélie Névéol et al. "The Quaero French Medical Corpus : A Ressource for Medical Entity Recognition and Normalization". In: 2014.

[4] stanford. *Understanding LSTM Networks*. 2015. URL: https://web.stanford.edu/.

[5] Lena Voita. *NLP Course Yandex School*. 2020. URL: https://lena-voita.github.io/nlp_course.html.