Resume

Text Mining and Chatbots

Ayman Damoun

Text Mining and Chatbots

28/01/2021

1. Summary of the projects

1. Summary of the projects
   - lab1: Word embeddings training
   - lab2: Named entity recognition

1. Summary of the projects
   - lab1: Word embeddings training
   - lab2: Named entity recognition

Word embeddings

Word embeddings

Word embedding is a learned representation of a word in which each word is represented using a vector in an n-dimensional space.

Word embeddings

Word embeddings

Word embedding is a learned representation of a word in which each word is represented using a vector in an n-dimensional space.

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.

Skip-Gram

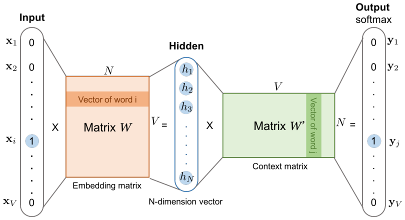


Figure 1: The skip-gram model [1]

skip-gram model [2]

predict a context word when a target word is taken as input.

Continuous Bag of Words

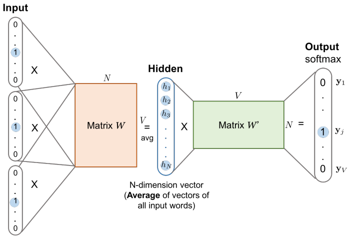


Figure 2: The CBOW model [1]

CBOW model [2]

predict the target word using the context words as input.

Cosine similarity

In order to get semantic similarity between vectors, we need to compare the Word embeddings. **Cosine similarity** is the most used method to compare two vectors.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

About the dataset

- The QUAERO French Medical Corpus
- The QUAERO French Press Corpus

Both datasets are complete corpus, tokenized and with one sentence per line

Experiments

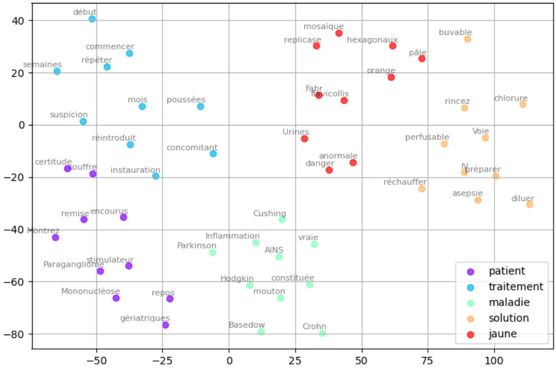


Figure 3: Visualizing Word Embeddings for skipgram (French Medical Corpus)

Experiments

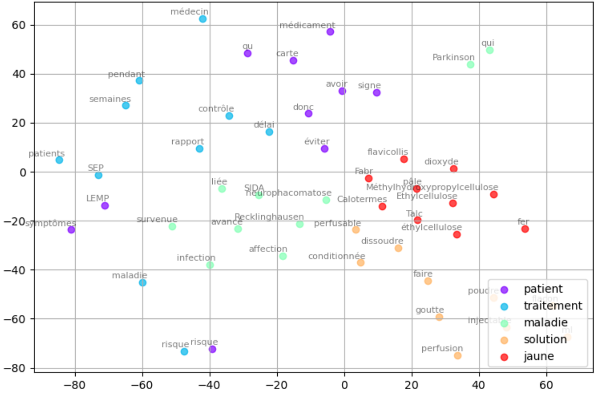


Figure 4: Visualizing Word Embeddings for cbow (French Medical Corpus)

Experiments

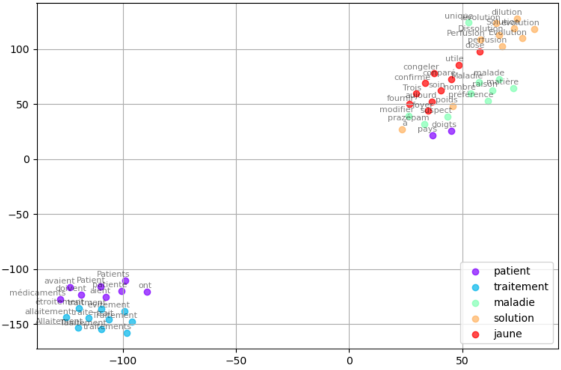


Figure 5: Visualizing Word Embeddings for fastText (French Medical Corpus)

1. Summary of the projects
   - lab1: Word embeddings training
   - lab2: Named entity recognition

Named Entity Recognition

Named Entity Recognition (NER)

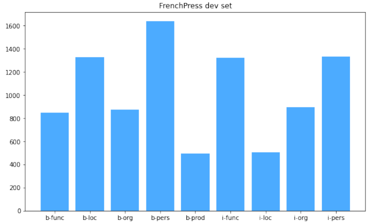
Can be defined as the process of determining whether a word or word-group represents a place, organization, or anything else. This can be broken down into two sub tasks: identifying the boundaries of the named entity, and identifying its type

Data

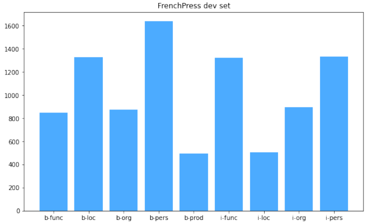
	FrenchPress training set	FrenchPress dev set	FrenchPress test set	FrenchMed training set	FrenchMed dev set	FrenchMed test set
Words	1156339	95222	95807	15339	13543	12388
sentences	35723	2825	2880	706	649	578

Table 1: Statistics on the FrenchPress and frenchMed EMEA corpus

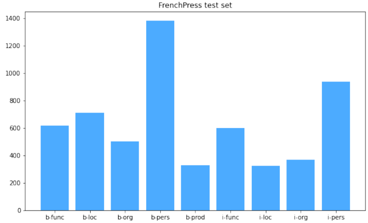
Data: Visualizing entities statistics (French Press Corpus)



(a) FrenchMed training set

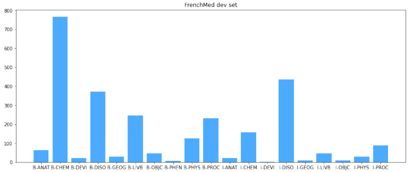


(b) FrenchPress dev set

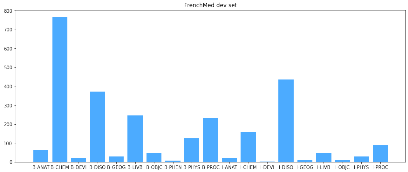


(c) FrenchPress test set

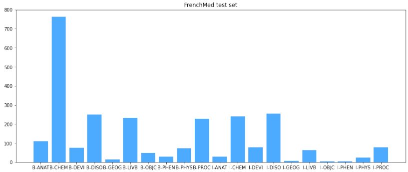
Visualizing entities statistics (French Medical EMEA Corpus)



(a) FrenchMed training set



(b) FrenchMed dev set



(c) FrenchMed test set

biLSTM

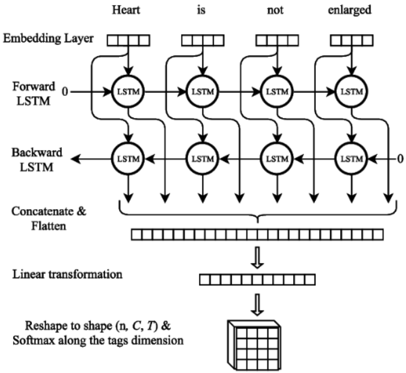


Figure 8: An illustration of the BiLSTM architecture [3]

biLSTM

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.

example of the prediction

Prediction on test set QUAERO FrenchPress

```
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
```

Experiments

	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

Conclusion

- According to F1-score for QUAERO French Press the skipgram model outperform Cbow and FastText Cbow with a score of 68.70% in test dataset. On the other hand for small datasets like QUAERO FrenchMedical i can ’ t notice a big difference for 20 epochs
- We can conclude that the larger the training set the better the result, and skipgram world embeddings get better results.

References I

[1] Lilian Weng.
Learning word embedding, 2017.

[2] Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean.
Distributed representations of words and phrases and their compositionality.
CoRR, abs/1310.4546, 2013.

[3] Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana.
Modelling radiological language with bidirectional long short-term memory networks.
pages 17–27, 01 2016.

Thanks!

And now, we welcome your questions and comments.