Lab2 - Named entity recognition

Jiangnan Huang & You Zuo

February 14, 2021

1 Introduction

In our work, You ZUO analysed the two French corpus and wrote the section 1, then performed the experiment (b) in the section 2.2, Jiangnan HUANG performed the experiments (a)(c)(d) in section 2.2 and wrote section 2 and 3, both You ZUO and Jiangnan HUANG reviewed and checked the full report.

1.1 Dataset

During this lab exercise, we used two datasets with named entity annotations : a small medical corpus (QUAERO_FrenchMed) and a larger news corpus (QUAERO_FrenchPress).

The datasets are provided in conll-like format and contain five columns separated by white spaces. Each word (or token) corresponds to a line and sentences are separated by an empty line. The columns correspond to: the token index within the sentence, the token, two columns that will not be used in this lab and finally, the last column contains the named entity tag.

index	token	var2	var3	entity
1	PRIALT	21	27	B-CHEM
1	EMEA	0	4	O
2	/	5	6	O
3	H	7	8	O
4	/	9	10	O
5	C	11	12	O
6	/	13	14	O
7	551	15	18	O
1	Qu	30	32	O
2	’	33	34	O

Table 1: First lines of EMEA train file

Remarkably, they did not removed the numbers and the punctuations because sometimes they can be the part of entity. For example, in the training set of EMEA, there is one entity of type "Living Beings" which compromises both number and symbol.

index	token	var2	var3	entity
1	Adultes	5155	5162	B-LIVB
2	(	5163	5164	O
3	y	5165	5166	O
4	compris	5167	5174	O
5	sujets	5175	5181	B-LIVB
6	âgés	5182	5186	I-LIVB
7	≥	5187	5188	I-LIVB
8	65	5189	5191	I-LIVB
9	ans	5192	5195	I-LIVB
10	)	5196	5197	O

Table 2: One example of entity containing number and symbol

1.2 Descriptive Statistics

To better understand our data structure, we did the descriptive statistics on the two corpora. To remind, due to name entities' properties (it might include digits and punctuation), we did not remove them during the prepossessing of data.

1.2.1 FrenchPress

First let us look at the Descriptive statistics of three dataset of FrenchPress corpus:

	Train.	Dev.	Test.
Tokens	1156339	95222	95807
Entities	135268	9869	6296
Unique Entities	12778	2652	2227
Sentences	35723	2825	2880

Table 3: Descriptive statistics of FrenchPress

We also looked at the distribution of entities for each splits:

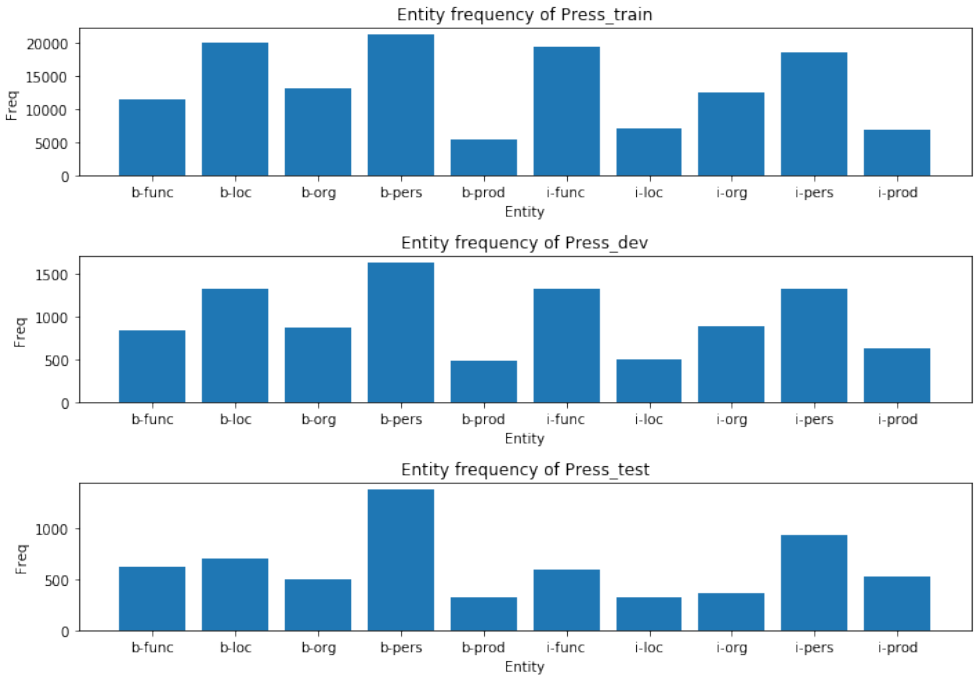

Figure 1: Distribution of entities in FrenchPress

As we can observe from the histograms above, their distributions are very similar. And they have five types of entities: PersonType(func), Location, Organization, Person and product. The entity of type pers is the most frequent, them comes the location and PersonType.

However, the frequency of i-func is higher than that of b-func, which means that the entity of persontype usually has many modifiers, such as *professeur en chirurgie* or *ministre des transports*.

1.2.2 FrenchMed-EMEA

There are two dataset in the directory of FrenchMed: EMEA and Medline. But due to time limitation we will explore only the EMEA, which is a smaller dataset.

	Train.	Dev.	Test.
Tokens	15339	13543	12388
Entities	3198	2704	2604
Unique Entities	914	739	664
Sentences	706	649	578

Table 4: Descriptive statistics of FrenchMed

Compared to the results of the FrenchPress, EMEA have smaller size of data, and the number of sentences in each split do not vary too much. Then we also analyzed the ratio of the tokens and the entities and the ratio of the entities and unique entities:

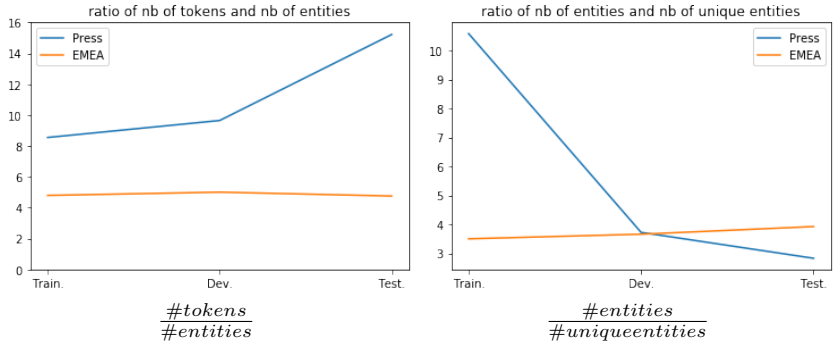

Figure 2: Ratio of results

From the plot on the left, we can observe that the ratio of tokens to entities of EMEA dataset is always lower than the EMEA, which means that EMEA has a higher proportion of entities. It is easy to understand because there are usually more technical terms in the medical context.

After that, the plot on the right shows the ratio of entities to unique entities. We can find that the FrenchPress dataset's entities appear more than ten times in the training set on average, but the entities in the EMEA dataset only appear less than four times on average.

To sum up, the medical corpus has a higher proportion of entities, and it has a higher level of variants since the number of their occurrences is not too concentrated. These two reasons will make the training of the NER model of the medical corpus more complicated than that of the Press corpus.

Finally, we explored the distribution of entities for FrenchMed-EMEA:

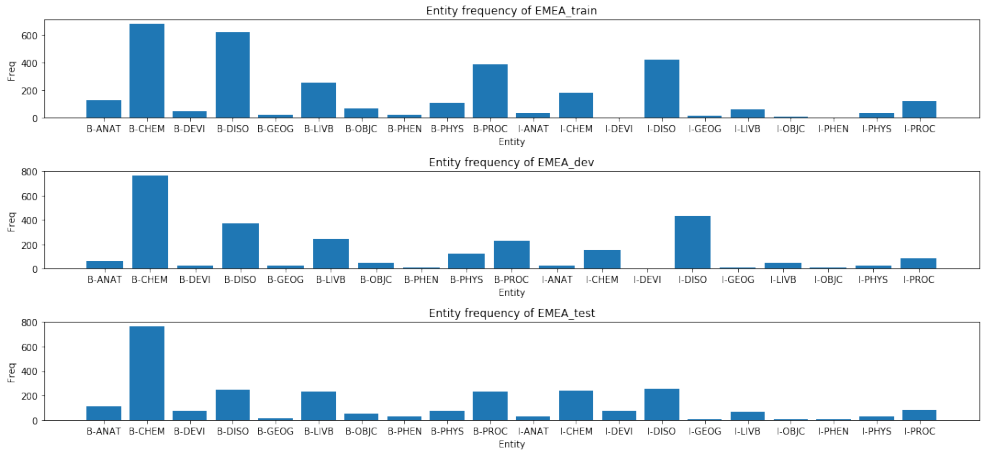

Figure 3: Distribution of entities in FrenchMed-EMEA

For NER of medical corpus, it has ten types of clinical entitie: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures. Because there are many types, the number of certain entities will be especially small such as GEOG, PHEN, DEVI etc.

2 Experiments and Results

2.1 Models and Configuration

The datasets we used are the FrenchPress and FrenchMed-EMEA, and we will use all the train, dev, test splits as given in the directory.

To train the ner models, we used the engine from NeuralNLP2. Each time we trained an NER model, we only need to modify the configurations in NeuralNLP2/ experiments/scripts/xx.sh and then run this shell.

For the word embeddings, we used the skip-gram models and fasttext models trained with medical corpus and non-medical corpus, so totally four embedding models.

To remind, the basic setting of hyper-parameters of embedding models:

- min count (word with frequency lower than this will be ignored) = 1
- length of word embedding = 100
- windom size (to determine the context) = 10

- number of epochs = 20

Finally, since the size of FrenchPress is large while the size of FrenchMed-EMEA is very small, we set the number of epochs for NER of FrenchPress to 10 and number of epochs for NER of FrenchMed-EMEA to 50.

2.2 Evaluation

In this part, we will study the variation in precision, recall, and F-score when the following parameters vary (while keeping the others fixed):

(a) size of the training set:

•use training set subsets of increasing sizes;

The medical corpus EMEA used to train the NER model has 693 sentences. We reduced the size of the corpus to 400 sentences and 200 sentences respectively, then applied the Word2Vec embedding model to train our new NER models for medical dataset EMEA.

		Accuracy	Precision	Recall	F1
FrenchMed	sg693(dev)	**86.53%**	**59.87%**	**39.79%**	**47.81%**
	sg400(dev)	85.60%	58.57%	34.28%	43.25%
	sg200(dev)	83.56%	47.36%	25.41%	33.07%
	sg693(test)	**85.50%**	**58.76%**	**38.47%**	**46.50%**
	sg400(test)	83.90%	52.27%	29.69%	37.87%
	sg200(test)	83.15%	49.08%	26.29%	34.24%

Table 5: NER trained with different size of training set

From the results obtained, we can find that NER model trained with all the 693 sentences achieved the best results on both dev set and test set. Thus for training the NER model, larger size of training dataset will lead to a better performance.

(b) size of the corpus used to train word embeddings:

•use raw text samples of increasing sizes to train word embeddings;

The original medical corpus used to train the embedding model has 3022 sentences. We reduced the size of the corpus to 2000 sentences and 1000 sentences respectively, and trained new skip-gram models. Based on these pre-trained embeddings, we then trained new NER models for medical dataset EMEA :

		Accuracy	Precision	Recall	F1
FrenchMed	sg3022(dev)	86.53%	59.87%	**39.79%**	47.81%
	sg2000(dev)	**86.82%**	**64.86%**	38.95%	**48.67%**
	sg1000(dev)	85.87%	64.05%	33.86%	44.30%
	sg3022(test)	84.77%	**64.61%**	30.76%	41.68%
	sg2000(test)	**85.78%**	64.29%	**38.53%**	**48.18%**
	sg1000(test)	84.79%	61.73%	33.64%	43.55%

Table 6: NER trained with different size of embedding tables

By observing the metrics of different NER models, we can find that skip-gram trained with 2000 sentences achieved the best results on accuracy and F1 both in dev set and test set, it had the best recall on test set and the best presision on dev set. Thus, embeddings trained on a larger corpus does not necessarily bring the best results. On the contrary, due to the small size of EMEA, too large embedding table may cause over-fitting problem.

(c) types of word embedding model:

•use word embeddings created with word2vec skip-gram, fastText skip-gram;

We've tried to use these two different word embedding models for both the medical corpus (QUAERO_FrenchMed/EMEA) and the non-medical corpus (QUAERO_FrenchPress):

		Accuracy	Precision	Recall	F1
FrenchMed	Word2Vec(dev)	**86.53%**	59.87%	**39.79%**	**47.81%**
	fastTest(dev)	84.77%	**64.61%**	30.76%	41.68%
	Word2Vec(test)	**85.50%**	58.76%	**38.47%**	**46.50%**
	fastTest(test)	82.38%	**59.37%**	22.61%	32.75%
FrenchPress	Word2Vec(dev)	**96.37%**	**71.87%**	**68.15%**	**69.96%**
	fastTest(dev)	94.39%	52.27%	52.63%	52.45%
	Word2Vec(test)	**97.23%**	**71.61%**	**61.46%**	**66.15%**
	fastTest(test)	96.04%	49.85%	51.50%	50.66%

Table 7: NER trained with different world embedding models

The results shows that in most of the cases, the Word2Wec skip-gram model works much better than fastText skip-gram on both these two corpus. There is only one exception: For the medical corpus FrenchMed, the fastTest model got better precision on both dev. and test set.

(d) domain of the corpus used for named entity recognition and that of the word embeddings:

During the experiments, we first fixed the domain of the corpus used for named entity recognition to medical domain, then we used the word embeddings trained on both medical and non-medical domain to train the model, then compare the

final results obtained. The word embedding method was fixed to Word2Vec skip-gram.

Corpus(NER)	Corpus(vec)	Accuracy	Precision	Recall	F1
FrenchMed	FrenchMed(dev)	**86.53%**	59.87%	**39.79%**	**47.81%**
	FrenchPress(dev)	83.64%	**64.26%**	24.25%	35.21%
	FrenchMed(test)	**85.50%**	58.76%	**38.47%**	**46.50%**
	FrenchPress(test)	82.77%	**59.94%**	21.68%	31.84%
FrenchPress	FrenchMed(dev)	**96.38%**	71.01%	**70.05%**	**70.53%**
	FrenchPress(dev)	96.37%	**71.87%**	68.15%	69.96%
	FrenchMed(test)	**97.31%**	**71.65%**	**63.73%**	**67.46%**
	FrenchPress(test)	97.23%	71.61%	61.46%	66.15%

Table 8: NER trained with different domains of the corpus

From the results obtained, we can observe that if we use a non-medical embedding table to train our NER model for the medical corpus, the results are worse than the NER model trained with medical embedding. It is logical as there are many specific words in the medical corpus, medical embeddings will provide more information. Besides, as discussed in the first section, the medical corpus has more difficulties when training a NER model. Therefore, the medical NER model's overall metrics are lower than those of the Press NER model.

However, suppose we use the word embedding model trained on the medical corpus to train our NER model for the non-medical corpus. In that case, the results are even better than when the two corpus are fitted(are both non-medical corpus). This result seems weird to explain. The underlying reason may be that the number of embedding vectors trained on the non-medical corpus(38136) is much more than those trained on the medical corpus(8083) and this may make the NER model easier to overfit the training dataset.

3 Comparability with the CLEF eHealth campaign participants

(a)Compare the QUAERO_FrenchMed dataset used in this lab session with the one distributed in the CLEF eHealth 2016 shared task:

The the QUAERO_FrenchMed dataset used in this lab session has already separated the sentences by words and each word has been noted. While in the CLEF eHealth 2016 shared task, the corpus only contains the complete sentences which are not separated.

(b)Can the results obtained in this lab session be directly compared to those of the CLEF eHealth shared task participants?

We don't think that the results obtained in this lab can be compared to those of

the CLEF eHealth shared task participants. As we only reran the given script on both the medical and non-medical corpus and had a global view about the NER model and it's performance, but we didn't go into the details of this NER model and it's algorithm. Also we didn't deeply analysed the corpus as the participants did. Therefore, the results that we obtained can only be used to better understand both the NER and the word embedding model, and has no comparative significance.

References

Névéol, A., Grouin, C., Leixa, J., Rosset, S., and Zweigenbaum, P. (2014). The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30.