université PARIS-SACLAY

Class on
Information Retrieval
AI Master, first year

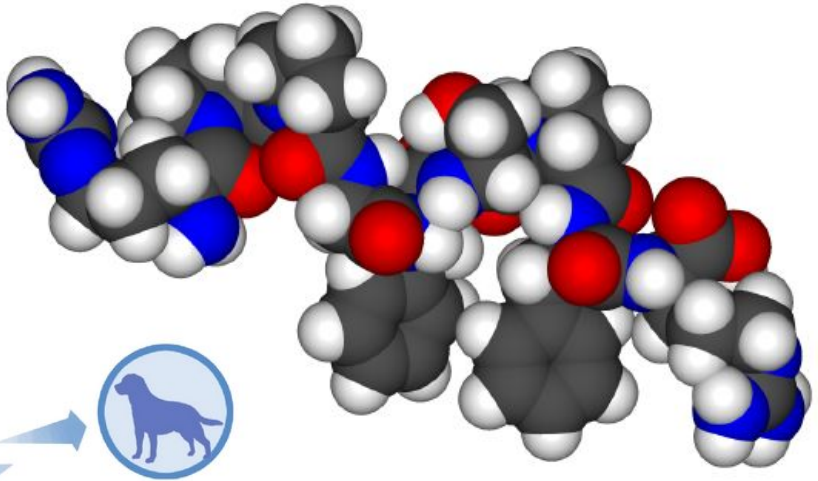
Instructor
Kim Gerdes

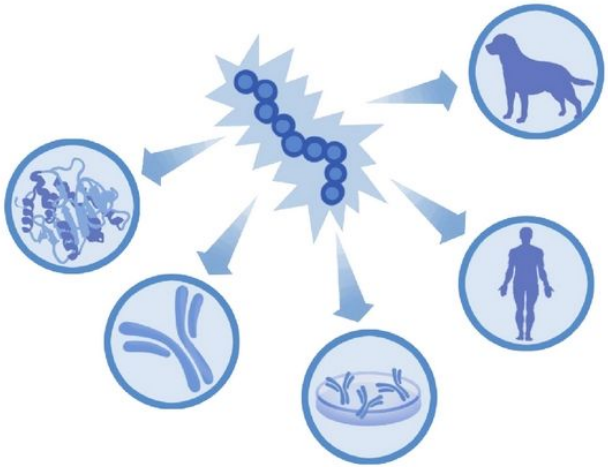
Information Retrieval Project
C07K - Peptents Bert for NER

Khuong Thanh Gia Hieu, Dorin Doncenco

Dataset

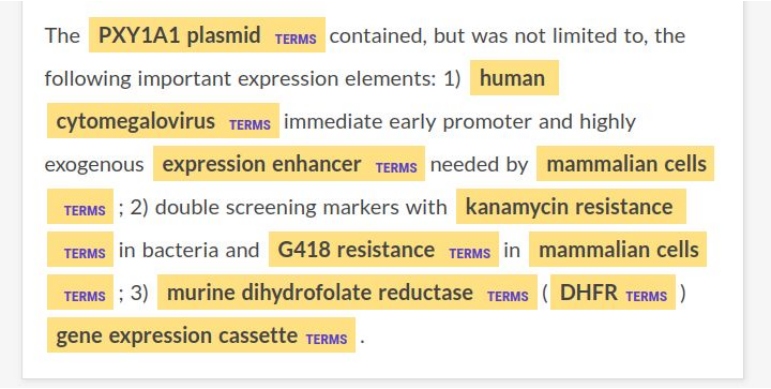
- Biology - Peptides
  - Hormones, enzymes
- 1994 patents (~300 lines/patent)
- Terms:
  - chemicals,
  - processes,
  - diseases,
  - organs,
  - etc.





Gold dataset

- 236 lines annotated, 777 “processed”
  - meds, descriptions, lab values, procedures
- Difficulties
  - prodigy recipe implementation
  - tokenization
  - bert model help
  - conversion <->spacy



Models

- Transformers
  - Maccrobat & manual annotations
  - no numerically referenced terms
- Spacy en_core_web_sm
  - user-friendlier to annotate
- Performance comparison

Starting point: jsylee/scibert_scivocab_uncased-finetuned-ner

As for the refining of LABEL_0 polym LABEL_1 yxin B sulfate LABEL_2 product, it is usually carried out by a spray drying method (patent application CN201210519331.7) or a lyophilization method (patent application CN201510775580.6) due to its difficulty to crystallize. The current patent about polym LABEL_0 yxin B LABEL_2 crystal includes polym LABEL_0 yxin B LABEL_2 1 di LABEL_0 hydrate LABEL_4 crystal disclosed in the patent application CN201210379231.9, which explicitly claims a compound having a molecular formula of C56H98N16O13·2H2O, which is not a sulfate. The polym LABEL_0 yxin B LABEL_2 1 LABEL_0 dihydrate LABEL_4 crystal of the patent is obtained by precipitation using a mixture of acetone and diethyl ether. Diethyl ether is extremely volatile and easily to be oxidized in air and causes an LABEL_0 explosion LABEL_3 , which is not suitable for industrial production. In addition, the products on the market are all mixtures of LABEL_0 polym LABEL_1 yxin B sulfate LABEL_2 . Regarding the preparation of LABEL_0 polym LABEL_1 yxin B1 LABEL_2 sulfate monomer, the patent application filed by the present applicant has been granted (Patent No. ZL201110390624.5), in which the LABEL_0 polym LABEL_1 yxin B LABEL_2 1 sulfate has a purity of 99.5%, and its solid is obtained by spray drying method. However, the LABEL_0 polym LABEL_1 yxin B sulfate LABEL_2 prepared by the current spray drying method is difficult to form a crystal form, and the product is very easy to agglomerate, which brings inconvenience to production and research, and also affects the quality and efficacy of the drug. LABEL_0

Chemical terms: alvaroalon2/biobert_chemical_ner

As for the refining of polymyxin B sulfate CHEMICAL product, it is usually carried out by a spray drying method (patent application CN201210519331.7) or a lyophilization method (patent application CN201510775580.6) due to its difficulty to crystallize. The current patent about polymyxin B CHEMICAL crystal includes polymyxin B1 dihydrate CHEMICAL crystal disclosed in the patent application CN 201210379231 CHEMICAL .9, which explicitly claims a compound having a molecular formula of C56H98N16O13·2H2O CHEMICAL , which is not a sulfate CHEMICAL . The polymyxin B1 CHEMICAL dihydrate CHEMICAL crystal of the patent is obtained by precipitation using a mixture of acetone CHEMICAL and diethyl ether CHEMICAL . Diethyl ether CHEMICAL is extremely volatile and easily to be oxidized in air and causes an explosion, which is not suitable for industrial production. In addition, the products on the market are all mixtures of polymyxin B sulfate CHEMICAL . Regarding the preparation of polymyxin B1 sulfate CHEMICAL monomer, the patent application filed by the present applicant has been granted (Patent No. ZL201110390624.5), in which the polymyxin B1 sulfate CHEMICAL has a purity of 99.5%, and its solid is obtained by spray drying method. However, the polymyxin B sulfate CHEMICAL prepared by the current spray drying method is difficult to form a crystal form, and the product is very easy to agglomerate, which brings inconvenience to production and research, and also affects the quality and efficacy of the drug.

More general terms: d4data/biomedical-ner-all

As for the refining of poly myxin Coreference B sulfate product, it is usually carried out by a spray drying method Therapeutic_procedure (patent application CN201210519331.7) or a l Therapeutic_procedure yo philization method Therapeutic_procedure (patent application CN201510775580.6) due to its difficulty to crystallize. The current patent about poly Coreference myxin Coreference B crystal Coreference includes poly Coreference myxin Coreference B1 dihydra te Detailed_description crystal disclosed in the patent application CN201210379231.9, which explicitly claims a compound having a molecular formula of C56H98N16O13·2H2O, which is not a sulfate. The poly Coreference myxin Coreference B1 di hydrate Detailed_description crystal of the patent is obtained by precipitation Detailed_description using a mixture of ace Medication tone and diethyl ether. Diet Diagnostic_procedure hyl ether is extremely volatile Detailed_description and easily to be oxidized in air Detailed_description and causes an explosion, which is not suitable for industrial production. In addition, the products on the market are all mixtures of polymyxin B sulfate. Regarding the preparation of polymy xin Coreference B1 sulfate monomer, the patent application filed by the present applicant has been granted (Patent No. ZL201110390624.5), in which the polymyxin B1 sulfate has a purity of 99.5%, and its solid is obtained by spray drying method. However, the poly Coreference my Coreference xin Coreference B sulfate prepared by the current spray drying Therapeutic_procedure method is difficult to form a crystal form, and the product is very easy to agglomerate, which brings inconvenience to production and research, and also affects the quality and efficacy of the drug.

More general terms: d4data/biomedical-ner-all

As for the refining of poly myxin Coreference B sulfate product, it is usually carried out by a spray drying method Therapeutic_procedure (patent application CN201210519331.7) or a l Therapeutic_procedure yo philization method Therapeutic_procedure (patent application CN201510775580.6) due to its difficulty to crystallize. The current patent about poly Coreference myxin Coreference B crystal Coreference includes poly Coreference myxin Coreference B1 dihydra te Detailed_description crystal disclosed in the patent application CN201210379231.9, which explicitly claims a compound having a molecular formula of C56H98N16O13·2H2O, which is not a sulfate. The poly Coreference myxin Coreference B1 di hydrate Detailed_description crystal of the patent is obtained by precipitation Detailed_description using a mixture of ace Medication tone and diethyl ether. Diet Diagnostic_procedure hyl ether is extremely volatile Detailed_description and easily to be oxidized in air Detailed_description and causes an explosion, which is not suitable for industrial production. In addition, the products on the market are all mixtures of polymyxin B sulfate. Regarding the preparation of polymy xin Coreference B1 sulfate monomer, the patent application filed by the present applicant has been granted (Patent No. ZL201110390624.5), in which the polymyxin B1 sulfate has a purity of 99.5%, and its solid is obtained by spray drying method. However, the poly Coreference my Coreference xin Coreference B sulfate prepared by the current spray drying Therapeutic_procedure method is difficult to form a crystal form, and the product is very easy to agglomerate, which brings inconvenience to production and research, and also affects the quality and efficacy of the drug.

=> **Dataset (MACCROBAT) is good**, but the model is under-performing

To fine tune, we have to reformat the MACCROBAT dataset

CASE: A 28-year-old previously healthy man presented with a 6-week history of palpitations.

The symptoms occurred during rest, 2–3 times per week, lasted up to 30 minutes at a time and were associated with dyspnea.

Except for a grade 2/6 holosystolic tricuspid regurgitation murmur (best heard at the left sternal border with inspiratory accentuation), physical examination yielded unremarkable findings.

An electrocardiogram (ECG) revealed normal sinus rhythm and a Wolff– Parkinson– White pre-excitation pattern (Fig.1: Top), produced by a right-sided accessory pathway.

T1	Age 8 19	28-year-old
T2	History 20 38	previously healthy
T3	Sex 39 42	man
T4	Clinical_event 43 52	presented
E1	Clinical_event:T4	
T5	Sign_symptom 31 38	healthy
E2	Sign_symptom:T5	
T6	Duration 60 66	6-week
E3	Duration:T6	
T7	Sign_symptom 78 90	palpitations

Example of the annotation **BEFORE**

To fine tune, we have to reformat the MACCROBAT dataset

tokens (sequence)	tags (sequence)
["A", "68", "-", "year", "-", "old", "female", "nonsmoker", ",", "nondrinker", "with", "a",…	["O", "B-Age", "I-Age", "I-Age", "I-Age", "I-Age", "B-Sex", "B-History", "O", "B-History",…
["A", "14", "-", "month", "-", "old", "boy", "was", "referred", "to", "our", "hospital",…	["O", "B-Age", "I-Age", "I-Age", "I-Age", "I-Age", "B-Sex", "O", "B-Clinical_event", "O",…
["Our", "patient", "was", "a", "68", "-", "year", "-", "old", "woman", "with", "chronic"…	["O", "O", "O", "O", "B-Age", "I-Age", "I-Age", "I-Age", "I-Age", "B-Sex", "O", "O", "O"…
["We", "present", "a", "case", "of", "pancreatic", "tumor", "without", "a",…	["O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-…
["A", "41", "-", "year", "-", "old", "Caucasian", "woman", "underwent", "a",…	["O", "B-Age", "I-Age", "I-Age", "I-Age", "I-Age", "B-Personal_background", "B-Sex", "O",…
["A", "54", "year", "-", "old", "diabetic", "woman", "complained", "of", "blurred",…	["O", "B-Age", "I-Age", "I-Age", "I-Age", "B-History", "B-Sex", "O", "O", "B-Sign_symptom",…
["This", "is", "a", "53", "-", "year", "-", "old", "male", "patient", "who", "went", "to",…	["O", "O", "O", "B-Age", "I-Age", "I-Age", "I-Age", "I-Age", "B-Sex", "O", "O", "O", "O",…

Example of the annotation **AFTER**

Finetune *bert-for-patents* on MACCROBAT:

As for the refining of polymyxin B sulfate Medication product, it is usually carried out by a spray drying method (patent application CN201210519331.7) or a lyophilization method (patent application CN201510775580.6) due to its difficulty to crystallize. The current patent about polymyxin Medication B crystal includes polymyxin B1 dihydrate crystal disclosed in the patent application CN201210379231.9, which explicitly claims a compound having a molecular formula of C56H98N16O13·2H2O, which is not a sulfate. The polymyxin B1 dihydrate crystal of the patent is obtained by precipitation using a mixture of acetone and diethyl ether. Diethyl ether is extremely volatile and easily to be oxidized in air and causes an explosion, which is not suitable for industrial production. In addition, the products on the market are all mixtures of polymyxin B sulfate Medication. Regarding the preparation of polymyxin B1 sulfate Medication monomer, the patent application filed by the present applicant has been granted (Patent No. ZL201110390624.5), in which the polymyxin Medication B1 sulfate Medication has a purity of 99.5% Lab_value, and its solid is obtained by spray drying method. However, the polymyxin B sulfate prepared by the current spray drying method is difficult to form a crystal form, and the product is very easy to agglomerate, which brings inconvenience to production and research, and also affects the quality and efficacy of the drug.

Finetune *bert-for-patents* on MACCROBAT:

As for the refining of polymyxin B sulfate Medication product, it is usually carried out by a spray drying method (patent application CN201210519331.7) or a lyophilization method (patent application CN201510775580.6) due to its difficulty to crystallize. The current patent about polymyxin Medication B crystal includes polymyxin B1 dihydrate crystal disclosed in the patent application CN201210379231.9, which explicitly claims a compound having a molecular formula of C56H98N16O13·2H2O, which is not a sulfate. The polymyxin B1 dihydrate crystal of the patent is obtained by precipitation using a mixture of acetone and diethyl ether. Diethyl ether is extremely volatile and easily to be oxidized in air and causes an explosion, which is not suitable for industrial production. In addition, the products on the market are all mixtures of polymyxin B sulfate Medication. Regarding the preparation of polymyxin B1 sulfate Medication monomer, the patent application filed by the present applicant has been granted (Patent No. ZL201110390624.5), in which the polymyxin Medication B1 sulfate Medication has a purity of 99.5% Lab_value, and its solid is obtained by spray drying method. However, the polymyxin B sulfate prepared by the current spray drying method is difficult to form a crystal form, and the product is very easy to agglomerate, which brings inconvenience to production and research, and also affects the quality and efficacy of the drug.

=> We need biology vocabulary for our model

Finetune *RoBERTa-large-PM-M3-Voc* on MACCROBAT:

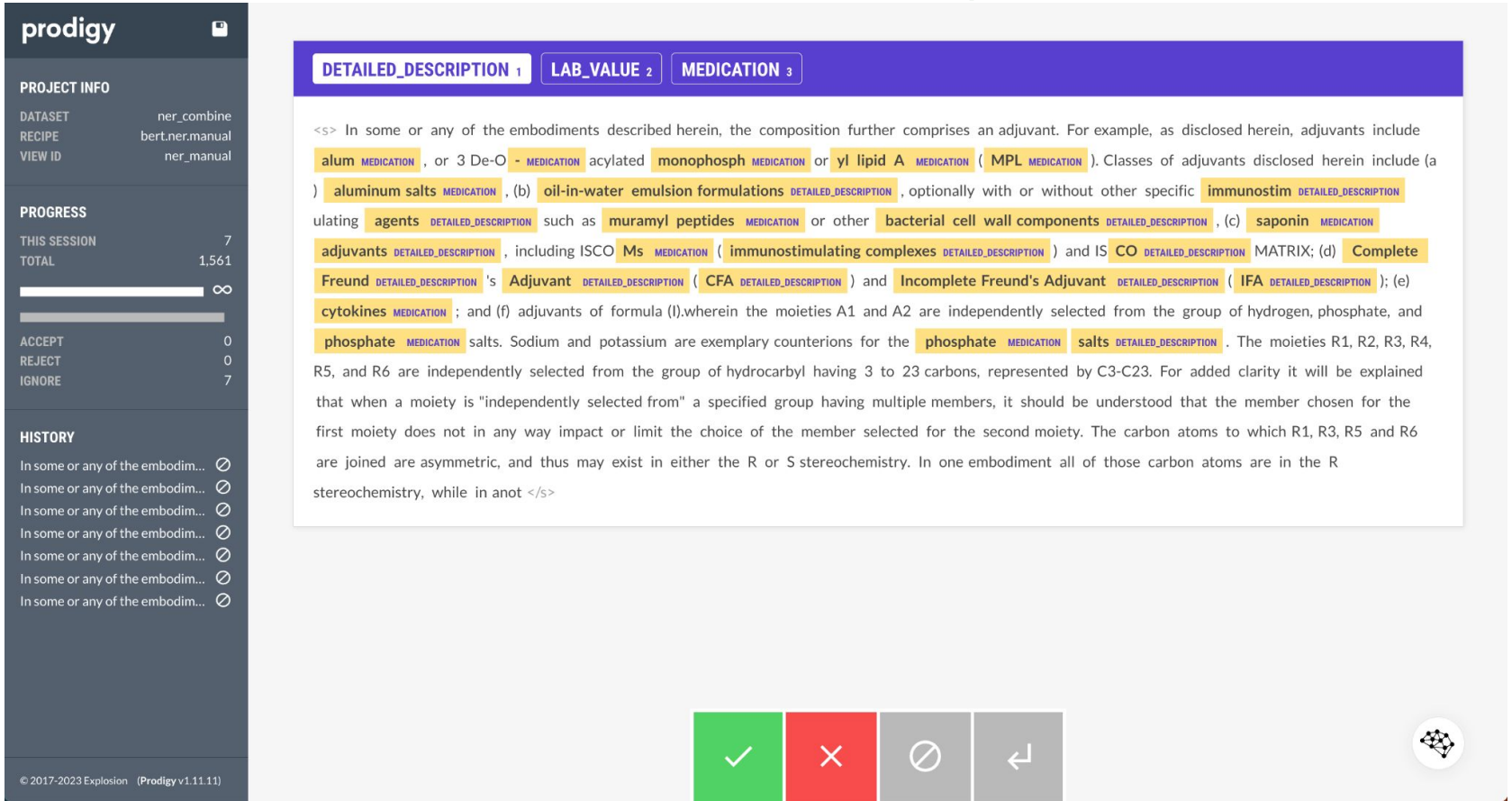
As for the refining of polymyxin B sulfate Chemical product, it is usually carried out by a spray drying method Term (patent application CN201210519331.7) or a lyophilization method Term (patent application CN201510775580.6) due to its difficulty to crystallize Term. The current patent about polymyxin B Chemical crystal includes polymyxin B1 Chemical dihydrate crystal disclosed in the patent application CN201210379231.9, which explicitly claims a compound having a molecular formula of C56H98N16O13·2H2O, which is not a sulfate. The polymyxin B1 Chemical dihydrate crystal of the patent is obtained by precipitation using a mixture of acetone and diethyl ether Chemical. Diethyl ether Chemical is extremely volatile Term and easily to be oxidized Term in air Term and causes an explosion Term, which is not suitable for industrial production. In addition, the products on the market are all mixtures of polymyxin B sulfate Chemical. Regarding the preparation of polymyxin B1 sulfate Chemical monomer, the patent application filed by the present applicant has been granted (Patent No. ZL201110390624.5), in which the polymyxin B1 sulfate Chemical has a purity of 99.5 Lab_value %, and its solid is obtained by spray drying method Term. However, the polymyxin B sulfate Chemical prepared by the current spray drying method Term is difficult to form a crystal form, and the product is very easy to agglomerate, which brings inconvenience to production and research, and also affects the quality and efficacy of the drug.

Finetune *RoBERTa-large-PM-M3-Voc* on MACCROBAT:

As for the refining of polymyxin B sulfate Chemical product, it is usually carried out by a spray drying method Term (patent application CN201210519331.7) or a lyophilization method Term (patent application CN201510775580.6) due to its difficulty to crystallize Term. The current patent about polymyxin B Chemical crystal includes polymyxin B1 Chemical dihydrate crystal disclosed in the patent application CN201210379231.9, which explicitly claims a compound having a molecular formula of C56H98N16O13·2H2O, which is not a sulfate. The polymyxin B1 Chemical dihydrate crystal of the patent is obtained by precipitation using a mixture of acetone and diethyl ether Chemical. Diethyl ether Chemical is extremely volatile Term and easily to be oxidized Term in air Term and causes an explosion Term, which is not suitable for industrial production. In addition, the products on the market are all mixtures of polymyxin B sulfate Chemical. Regarding the preparation of polymyxin B1 sulfate Chemical monomer, the patent application filed by the present applicant has been granted (Patent No. ZL201110390624.5), in which the polymyxin B1 sulfate Chemical has a purity of 99.5 Lab_value %, and its solid is obtained by spray drying method Term. However, the polymyxin B sulfate Chemical prepared by the current spray drying method Term is difficult to form a crystal form, and the product is very easy to agglomerate, which brings inconvenience to production and research, and also affects the quality and efficacy of the drug.

=> Good starting point for Prodigy 🦄

Prodigy with Huggingface model 🦄:



NER Progress

Step 1: Finetune RoBERTa-large-PM-M3-Voc

Step 2: Correct the predicted terms with Prodigy unicorn

- We modify prodigy ner.manual to load the Huggingface model in order to help with annotations.

Step 3.1: Train SpaCy with the corrected terms

Step 3.2: Train Huggingface model with the corrected terms

Spacy VS RoBERTa

Spacy ->
0.62 F1
0.59 precision

```
============================ Training pipeline ============================
Components: ner
Merging training and evaluation data for 1 components
  - [ner] Training: 232 | Evaluation: 57 (20% split)
Training: 192 | Evaluation: 44
Labels: ner (3)
ℹ Pipeline: ['tok2vec', 'tagger', 'parser', 'attribute_ruler',
'lemmatizer', 'ner']
ℹ Frozen components: ['tagger', 'parser', 'attribute_ruler',
'lemmatizer']
ℹ Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SPEED   SCORE
---  ------  ------------  --------  ------  ------  ------  ------  ------
  0       0          0.00     41.88    0.00    0.00    0.00  3969.25    0.00
  7    1000          0.00   8190.20   62.63   58.28   67.69  4131.13    0.63
 19    2000          0.00   4013.96   62.27   59.44   65.38  3862.46    0.62
✔ Saved pipeline to output directory
NER_model_filtered/model-last
```

warnings.warn(

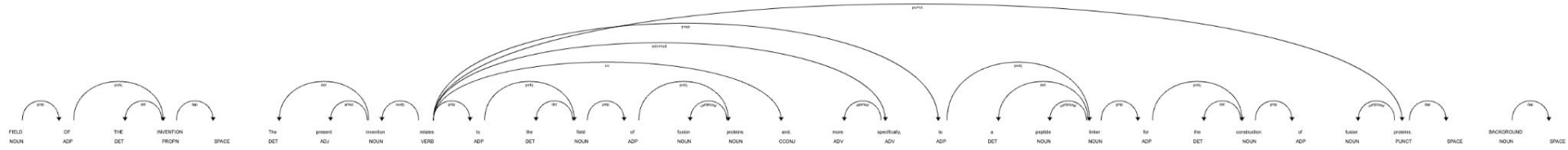
[110/110 01:15, Epoch 10/10]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
1	No log	0.777274	0.000000	0.000000	0.000000	0.722838
2	No log	0.491469	0.456989	0.361702	0.403800	0.829268
3	No log	0.424630	0.409091	0.344681	0.374134	0.840355
4	No log	0.372463	0.489899	0.412766	0.448037	0.865854
5	No log	0.362829	0.621212	0.523404	0.568129	0.879157
6	No log	0.349211	0.645833	0.527660	0.580796	0.882483
7	No log	0.345299	0.633663	0.544681	0.585812	0.887472
8	No log	0.354643	0.628272	0.510638	0.563380	0.884146
9	No log	0.349130	0.608911	0.523404	0.562929	0.885809
10	No log	0.351061	0.627551	0.523404	0.570766	0.885809

<- RoBERTa
0.57 F1
0.62 precision

université PARIS-SACLAY

Relation graph



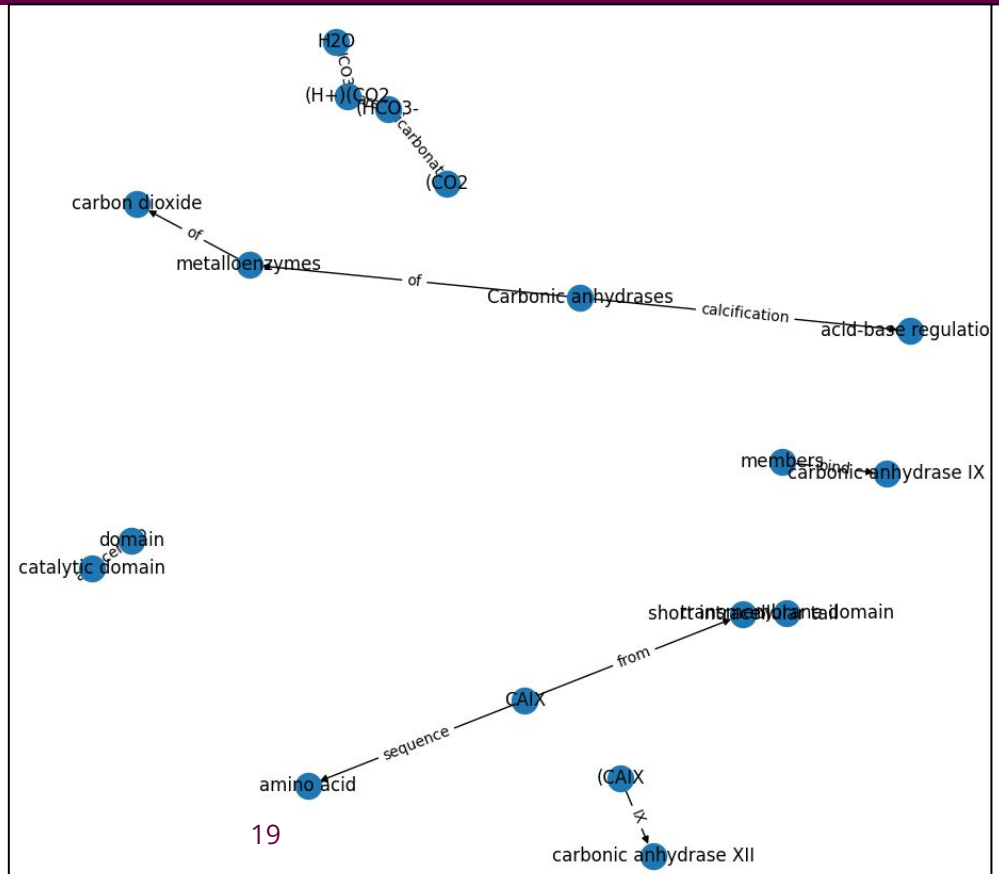
FIELD OF THE INVENTION

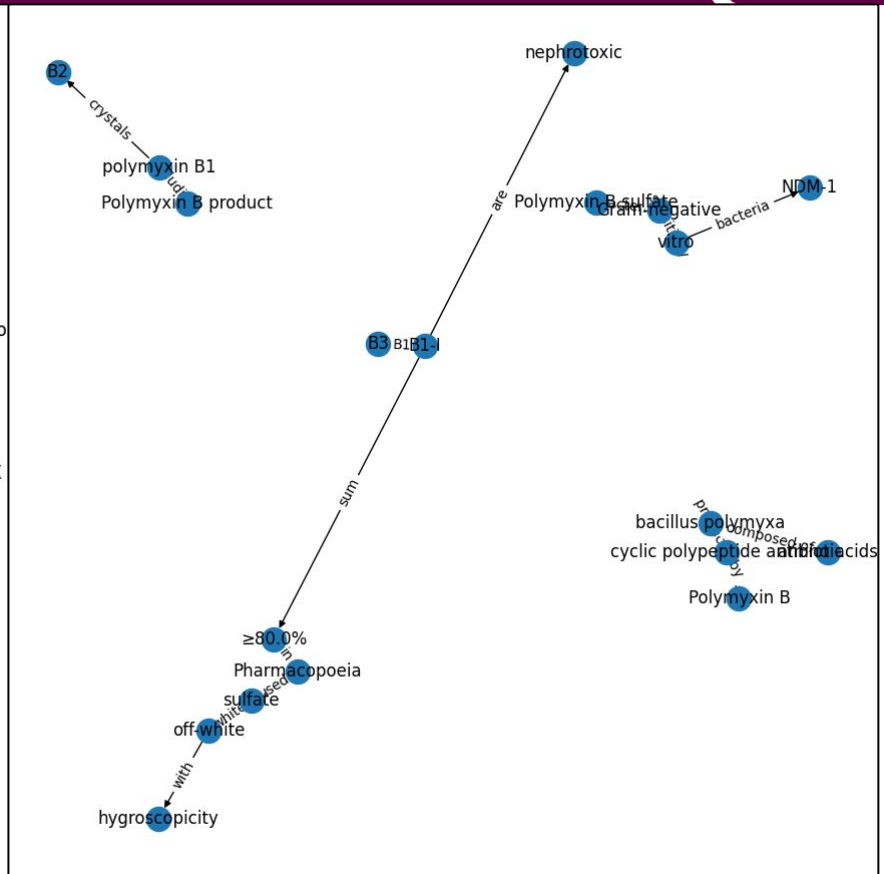
The present invention relates to the field of fusion proteins Term and, more specifically, to PRED a peptide linker Term for the constructio PRED n of fusion proteins Term .

BACKGROUND

In recent two decades, protein fusion technology Term has been widely used in the construction of PRED bifunctional antibodies Term , bifunctional enzymes Term , and bifunctional proteins Term . However, a variety of problems have been encountered in the construction of fusion proteins Term . For example, proteins that fold correctly during expression alone do not fold properly in the fusion protein Term ; the active site Term is blocked PRED after fusion due to the short distance between PRED the two fused proteins Term ; the fusion protein Term molecule PRED is easily degraded by PRED proteases Term when it cannot fold properly or when its conformation has changed; the protein catalytic domain Term with certain flexibility loses PRED its original function after fusion; and so on. The emergence of these problems often leads to reduction or even complete loss of the activity of the fusion proteins Term . It is generally believed that the activity of the original protein Term molecule will decrease to a certain extent after the protein molecule is constructed in PRED the fusion protein Term . A favorable fusion protein Term is the one that keeps more than 50% activity of PRED the original protein molecule(s Term). In order to solve the above problems, researchers conducted many studies and explorations on the design and construction of fusion proteins Term to improve the activity of PRED fusion proteins Term . such as changing the linking order of the fused proteins Term , changing PRED different fusion sites Term , using PRED different fusion partners Term , or using PRED a peptide linker Term , etc.

Relation graph





Summary and Future work

- Focused on NER w/ Bert models
- Knowledge graphs:
  - Implement graph search
  - Generalize relationships (with synonyms, etc)
  - Improve Coreference
  - Improve terms (e.g. clustering)
- Search engine?
- Other data (scientific paper, other patents, etc) ?
- Final project: https://github.com/Dorin-D/IR-peptents

THE END

Thank you!