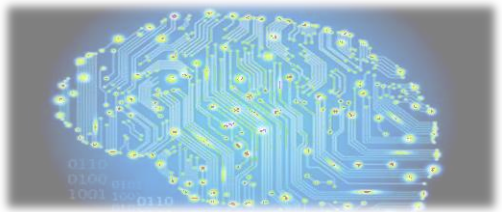
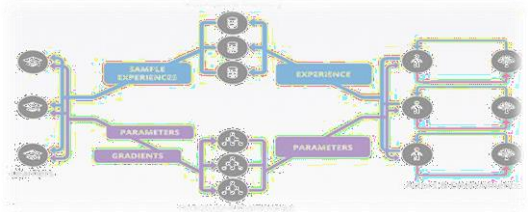
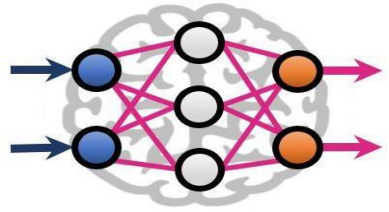
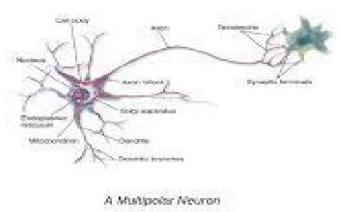
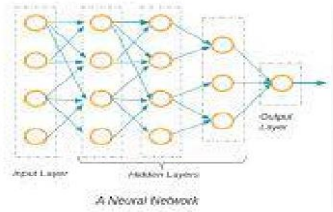





Hypernym discovery : CRIM experiments(3)

CRIM 구현 및 실험

Geonil Yun,
rjsdlf45@gmail.com
2021. 10. 12.

Introduction

- **CRIM – Hybrid Approach**
  - Combination of two approaches

    - Unsupervised : Pattern-based approach
    - Supervised : Projection learning approach

  - Dataset – SemEval2018 (general, medical, music)

    - 2 query types, concept and named entity

Introduction

- **SemEval2018-Task9 : Hypernym discovery**
  - 5 subtasks in SemEval2018-Task9

General-purpose

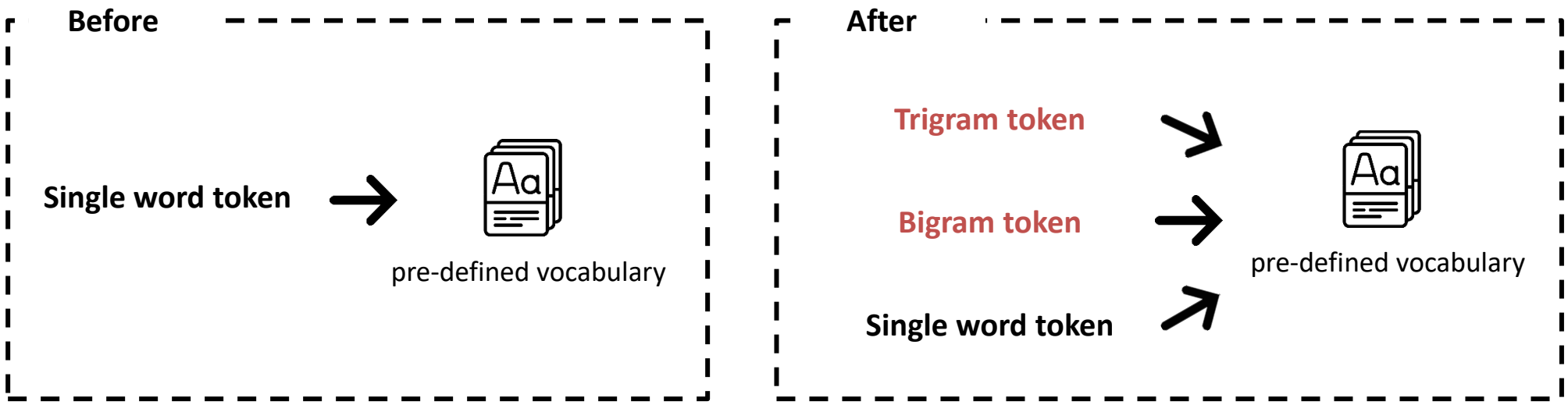
Domain-specific

Task	General			Medical	Music
Subtask	1A	1B	1C	2A	2B
Language	English	Italian	Spanish	English	English
Words	3B	3B	1.8B	130M	100M

    - Find candidate hypernyms in pre-provided corpus.

Experiment setting

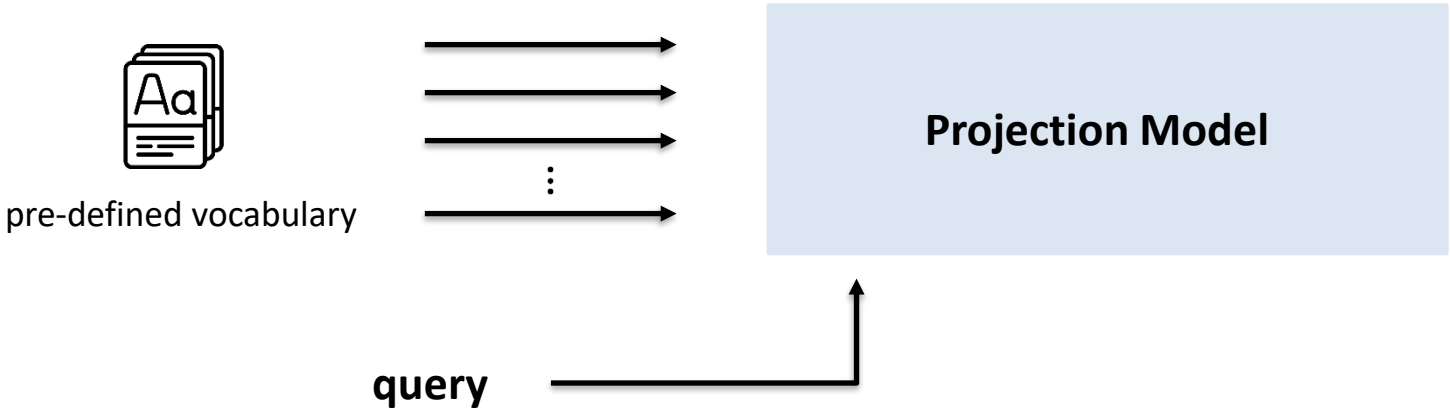
- **Single Word, bigram, trigram**
  - **Word dictionary**



- Start with **trigram**, then search **bigram**, **single word** in vocabulary.
- Build skip gram dataset from corpus with use this tokenization.

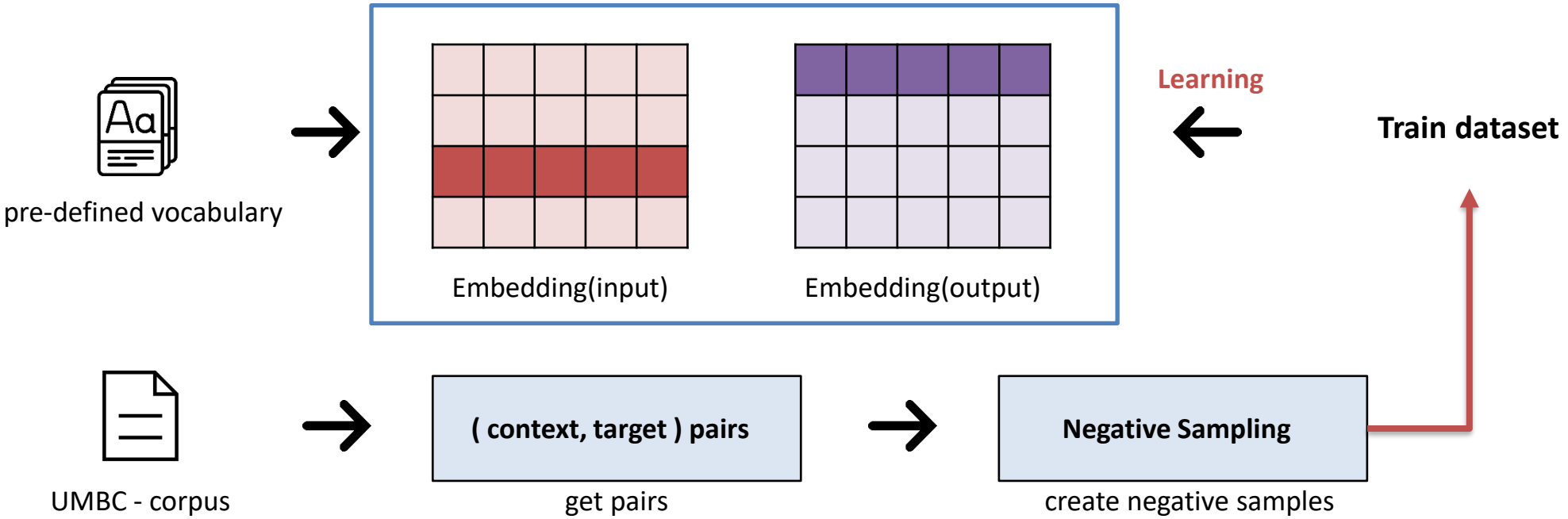
Experiment setting

- **Hypernym candidate**
  - **Use every word in vocabulary as a hypernym candidate**



Experiment setting

- **Word Embeddings**
  - **Skip-gram model structure**



Experiment setting

- **Word Embeddings**

  - **Skip-gram model structure**

hyperparameter	setting
embedding dimension	100
negative samples	15
batch size	1024
learning rate	0.0001
epoch	10
window size	2

  - Embeddings are trained by 50/408 corpus.
  - Compare each embedding trained with 1, 10, 20, 30, 40, 50 corpus.

Experiment setting

- **Pattern-based**
  - Used patterns in CRIM

	Pattern
Co-hyponym	X_1, X_2, .. and X_n

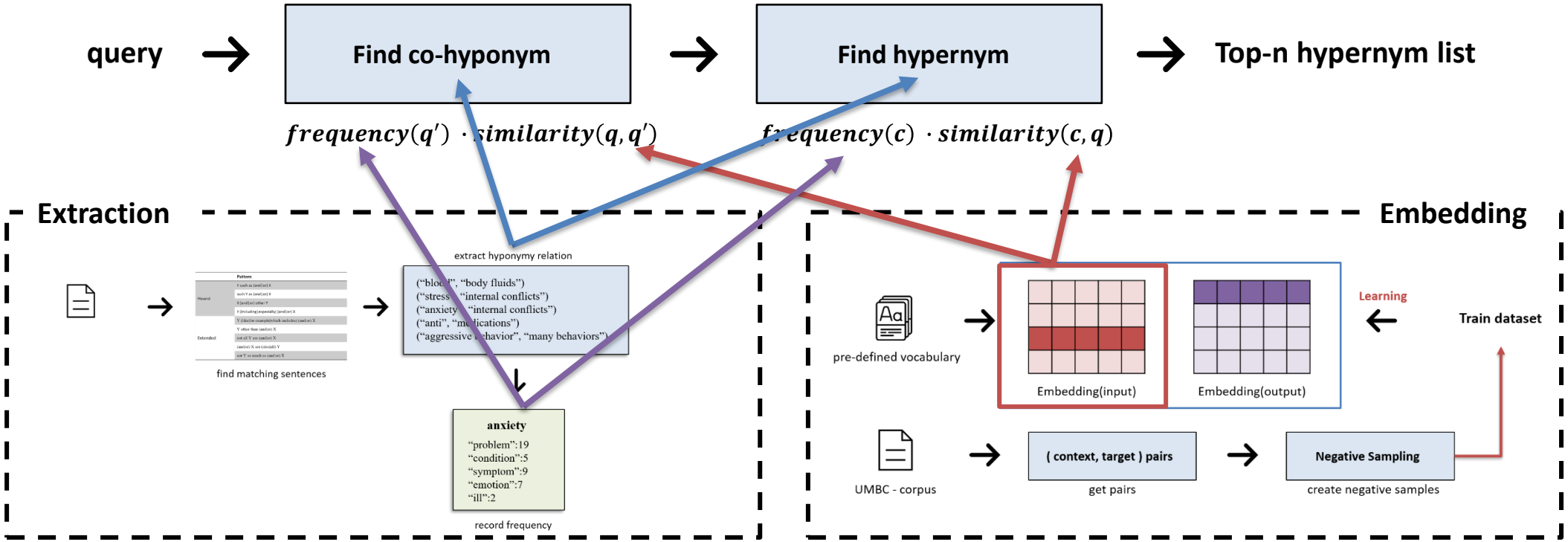
	Pattern
Hearst	Y such as (and|or) X
	such Y as (and|or) X
	X (and|or) other Y
	Y (including|especially) (and|or) X
Extended	Y (like|for example|which includes) (and|or) X
	Y other than (and|or) X
	not all Y are (and|or) X
	(and|or) X are (also|all) Y
	not Y so much as (and|or) X

- (X, Y) pair is (hyponym, hypernym) “is-a” relation.
- Use these patterns to find relation from corpus.

Experiment setting

- **Pattern-based**

  - **Pipeline**

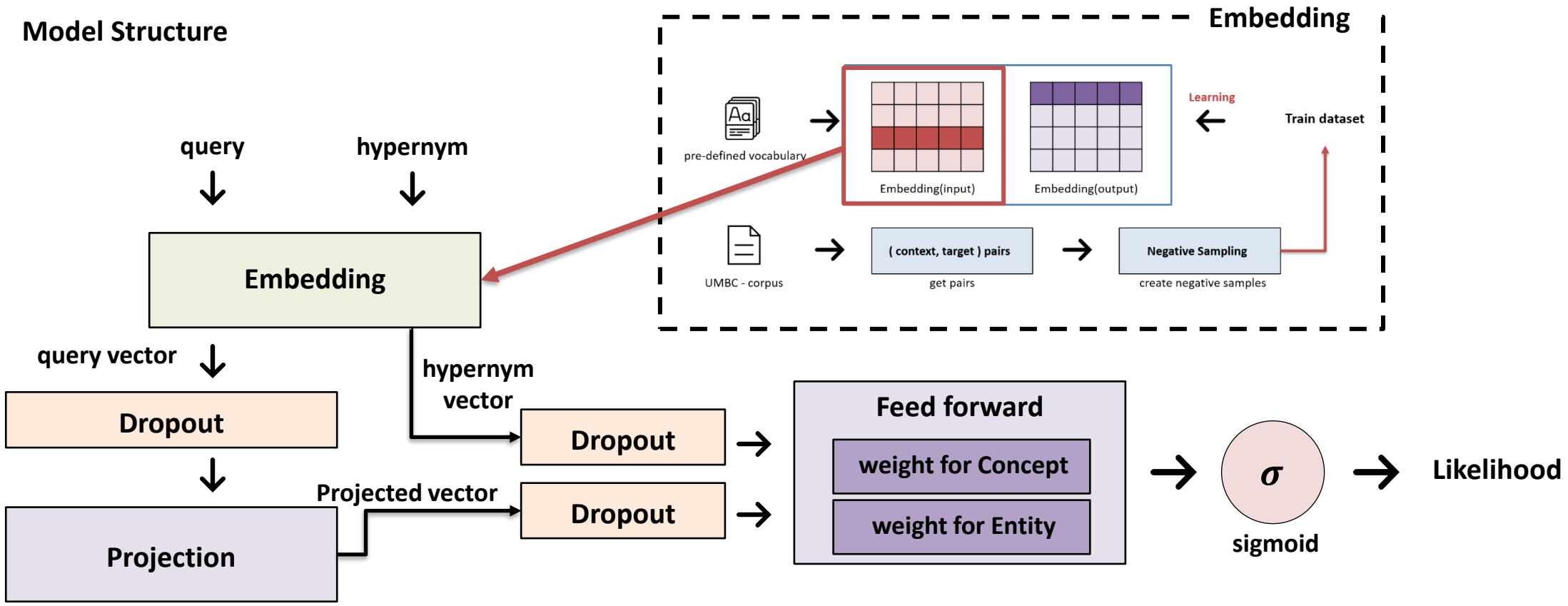


- From input query, find co-hyponyms of query and create co-hyponym set.
- Find hypernyms of co-hyponyms and calculate score based on frequency, embeddings.
- Extract 5 co-hyponyms first, get top 15 hypernyms by ranking hypernym of co-hyponym set.

Experiment setting

- **Projection Learning**

  - **Model Structure**



  - **Dropout** is applied to the embedding **query** and **hypernym** and **projected vector**.
  - Train **separate** logistic regression classifier(for **Concept**, for **Entity**).

Experiment setting

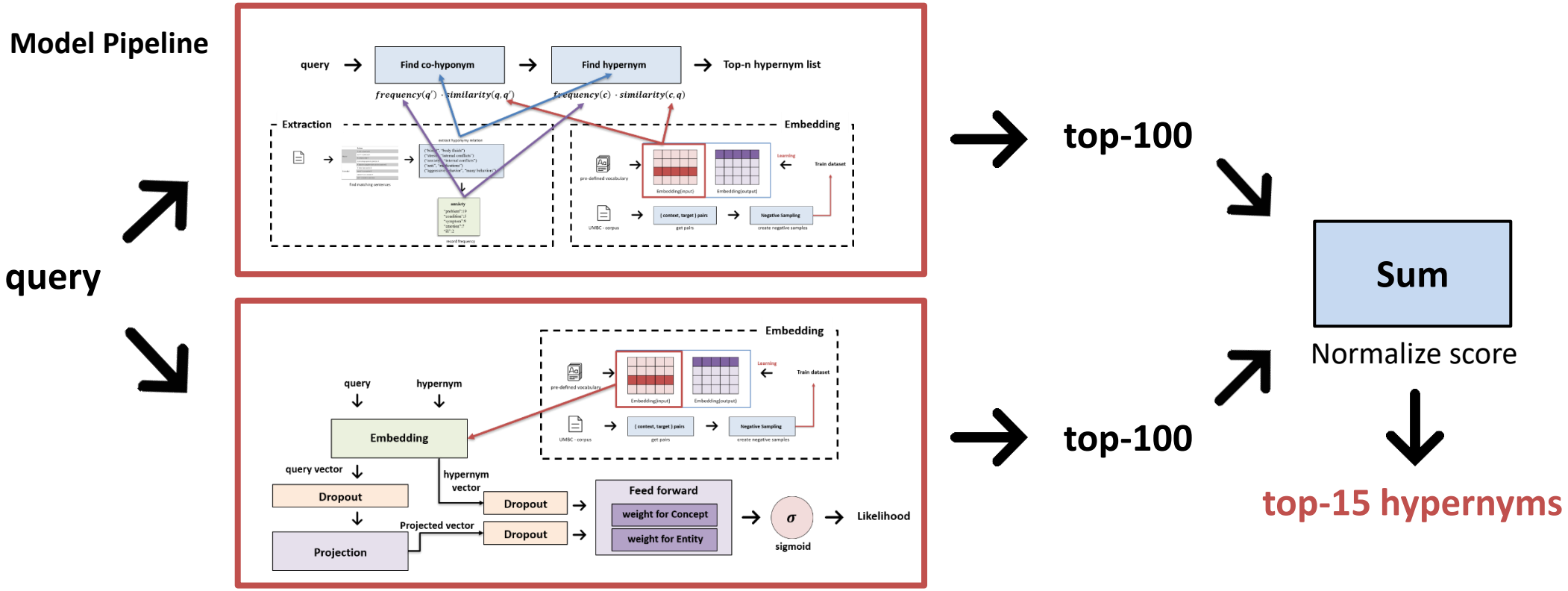
- **Projection Learning**
  - **Model settings**

hyperparameter	setting
embedding dimension	100
negative samples	15
batch size	32
learning rate	0.0001
epoch	20

- Embeddings are trained by 50/408 corpus.
- Compare each embedding trained with 1, 10, 20, 30, 40, 50 corpus.
- Use **gradient clipping** during training.

Experiment setting

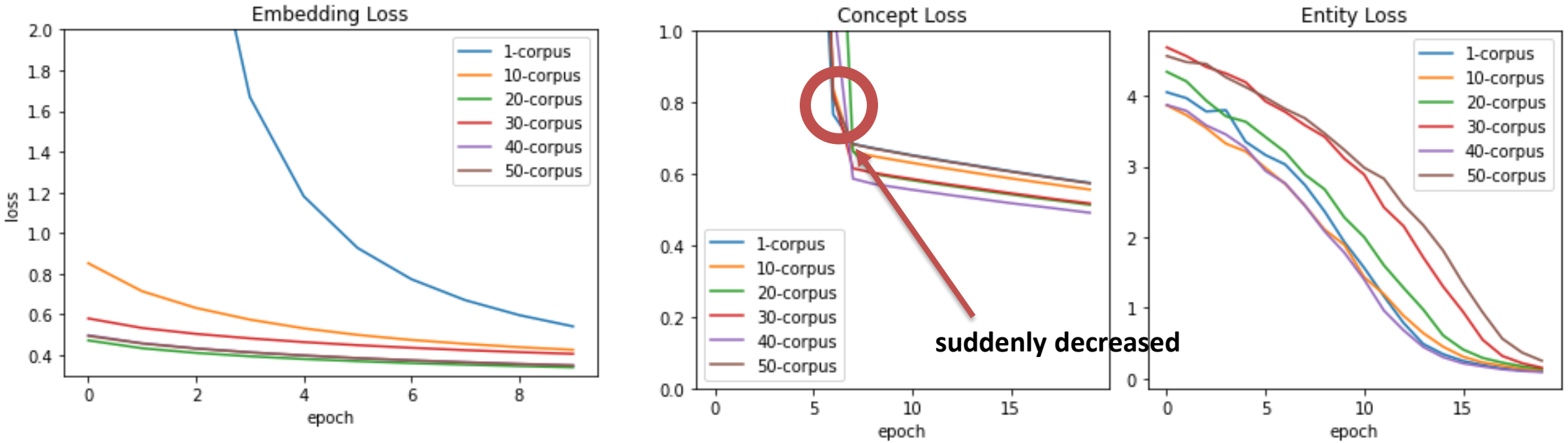
- **Hybrid Model**
  - ➢ **Model Pipeline**



- Embeddings are trained by 50/408 corpus.
- Compare each embedding trained with 1, 10, 20, 30, 40, 50 corpus.
- Use **gradient clipping** during training.

Experiment result

- **Training graph**



  - Embedding loss decreased as more corpus were trained.
  - Check overfitting at Concept data training.

Experiment result

- **Pattern-based**

Model	MAP	MRR	P@1
CRIM	***7.36***	***15.44***	***10.73***
Unsupervised-1	2.87	6.78	4.80
Unsupervised-10	3.68	7.91	5.13
Unsupervised-20	3.88	8.51	5.33
Unsupervised-30	4.15	9.19	6.20
Unsupervised-40	4.37	9.66	6.53
Unsupervised-50	4.44	9.82	6.53

Experiment result

- **Projection Learning**

Model	MAP	MRR	P@1
CRIM	***7.36***	***15.44***	***10.73***
Supervised-1			
Supervised-10			
Supervised-20			
Supervised-30			
Supervised-40			
Supervised-50			

Experiment result

- **Hybrid**

Model	MAP	MRR	P@1
CRIM	***7.36***	***15.44***	***10.73***
Hybrid-1			
Hybrid-10			
Hybrid-20			
Hybrid-30			
Hybrid-40			
Hybrid-50			

Issue & Future

- **last week**
  - **Processing single word, bigram, trigram.**
  - **Add dropout, gradient clipping and several specifics.**
  - Training with whole dataset.
  - Evaluation hybrid model.
  - **Separate Concept model and Entity model.**

- **Future works**
  - Training with whole dataset.
  - Evaluation hybrid model.
  - **Check overfitting.**

Thank you