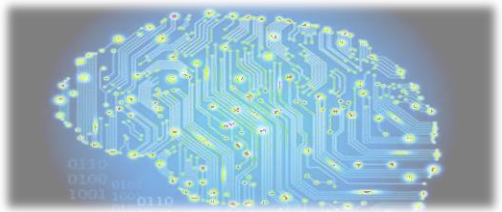
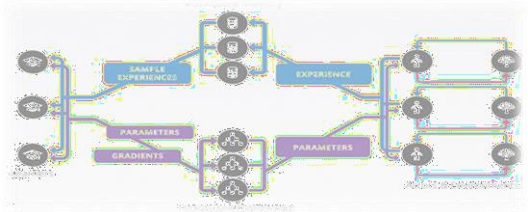
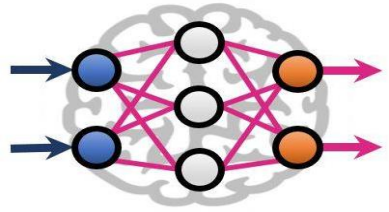
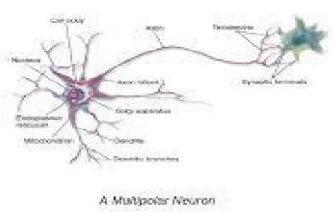
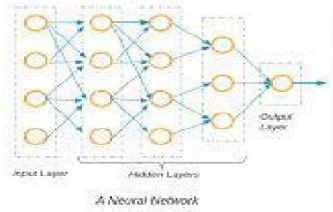





Hypernym discovery : CRIM experiments(2)

CRIM 구현 및 실험

Geonil Yun,
rjsdlf45@gmail.com
2021. 10. 05.

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

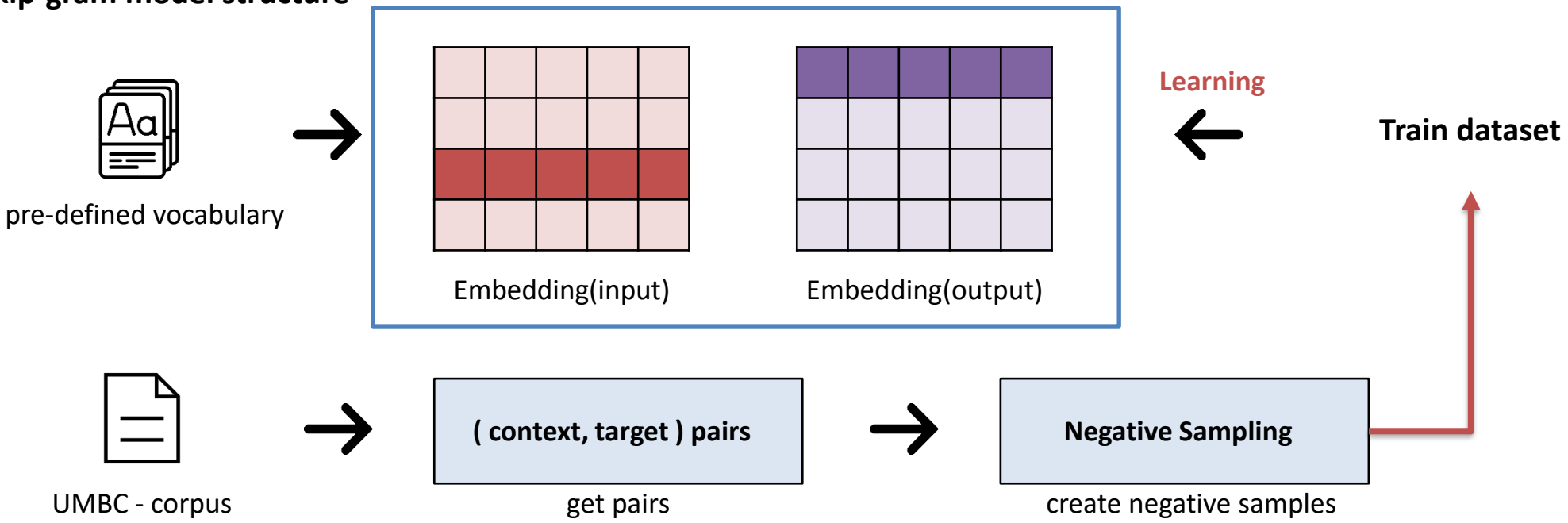
Domain-specific

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

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



- vocab size : 162,827
- Embedding dimension : 30
- Window size : 2
- Negative samples : 5 for each sample

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**

  - Extraction hyponymy

UMBC - corpus

	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

find matching sentences

extract hyponymy relation

("blood", "body fluids")
("stress", "internal conflicts")
("anxiety", "internal conflicts")
("anti", "medications")
("aggressive behavior", "many behaviors")

anxiety

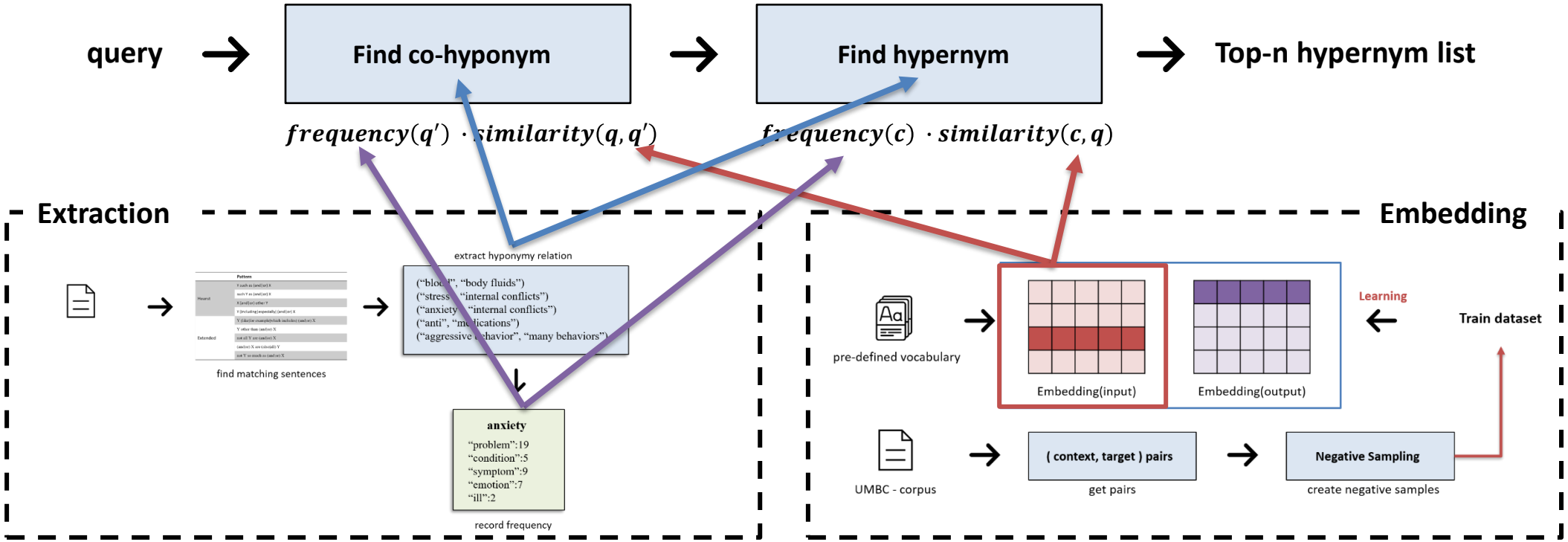
"problem":19
"condition":5
"symptom":9
"emotion":7
"ill":2

record frequency

- Use **pre-defined corpus**(UMBC 3-billion-word corpus).
- Find **(hyponym, hypernym)** pairs and record **frequency**.
- Also find **co-hyponym** pairs and record frequency.

Experiment setting

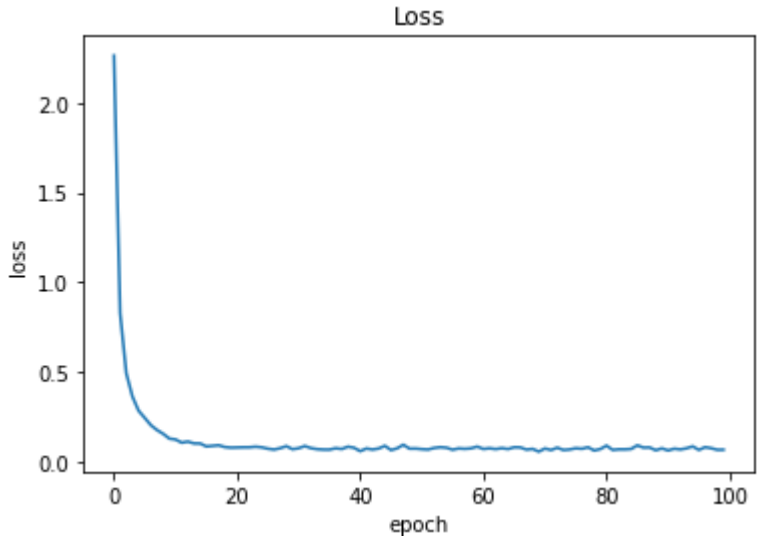
- **Pattern-based**
  - ➢ **Pipeline**



- 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 result

- **Pattern-based**
  - ➢ Applying small dataset
    - ▪ Use 1/403 corpus data to learn embeddings



query Q

device
equipment
system
drug
network
material

→

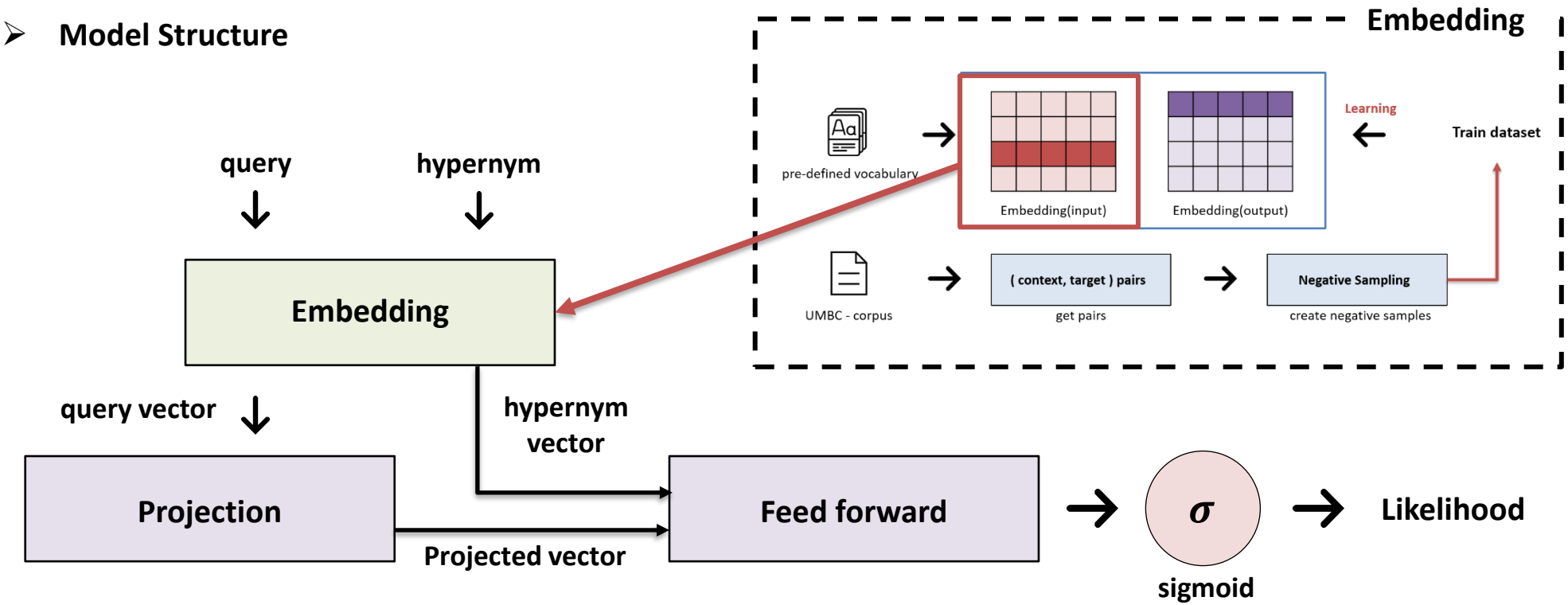
Top – 15 list

technology
application
area
infrastructure
resource
item
product
issue
field
tool
project
group
asset
facility
information

Experiment setting

- **Projection Learning**

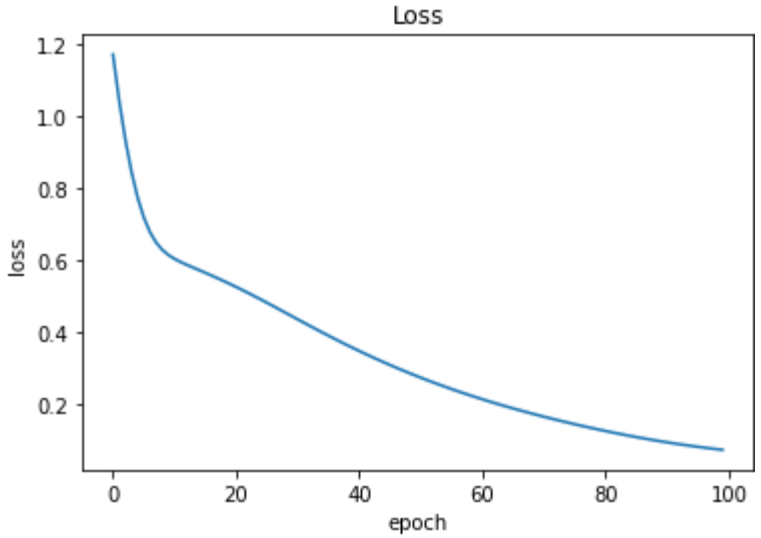
  - **Model Structure**



  - Input (query, hypernym candidate) pair
  - Projection kernel size : 24
  - Use negative sampling to train (n=5)
  - Loss function : Binary Cross Entropy
  - Adam optimizer

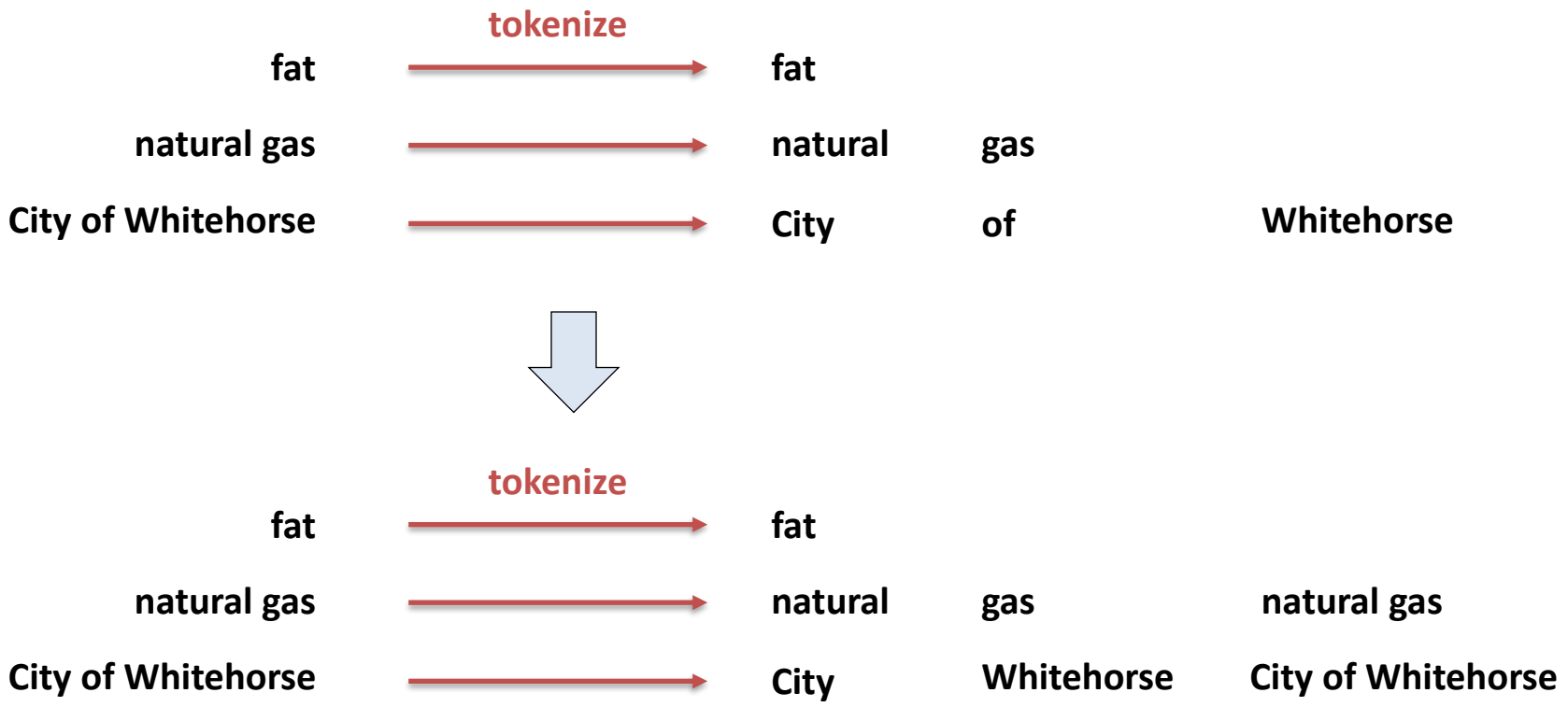
Experiment result

- **Pattern-based**
  - ➢ Applying small dataset
    - ▪ Use 1/403 corpus data to learn embeddings



Issue & Future

- **ISSUE**
  - Single word, bigram, trigram



Issue & Future

- **ISSUE**
  - Hypernym candidate
    - When predicting with projection model, use all vocabs for hypernym candidates.

Issue & Future

- **Future works**
  - 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.

Thank you