Master's Thesis Presentation for the 139th Dissertation Evaluation

Effective Hypernym Discovery Method based on Masked Language Modeling with Hearst Patterns

허스트 패턴을 이용한 마스크 언어 모델링 기반 효과적인 상위어 탐색 기법

Geonil Yun

rjsdlf45@gmail.com

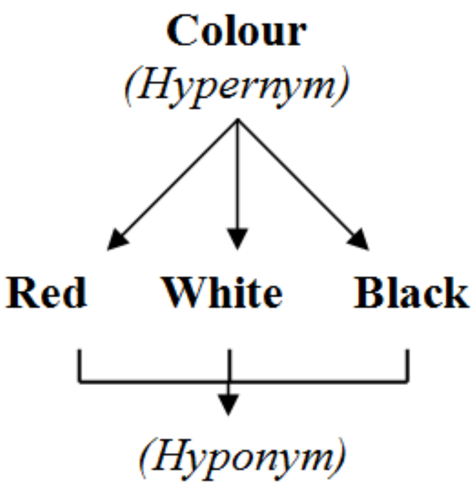
18th May, 2023

중앙대학교 CHUNG-ANG UNIVERSITY AutoML

Introduction

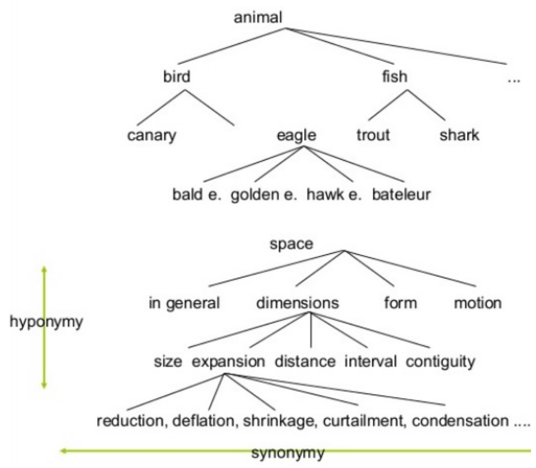
Hypernym relation

- A hypernym is a general term of the hyponym.
- The hypernym relation also called "is-a" relation.(i.e. "A car is a vehicle")
- Hypernym relation is used as important information in many NLP tasks[1].
  ex) Question-answering, ontology construction, textual entailment, and lexicon augmentation

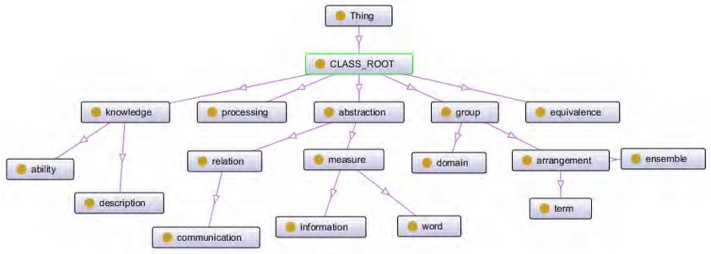


Hypernym Relation

Lexicon Augmentation



Ontology Construction



Textual Entailment

Relationship		Premise & Hypothesis
Entailment	Premise:	This church choir sings to the masses as they sing joyous songs from the book at a church.
	Hypothesis:	The church is filled with song.

NLP systems

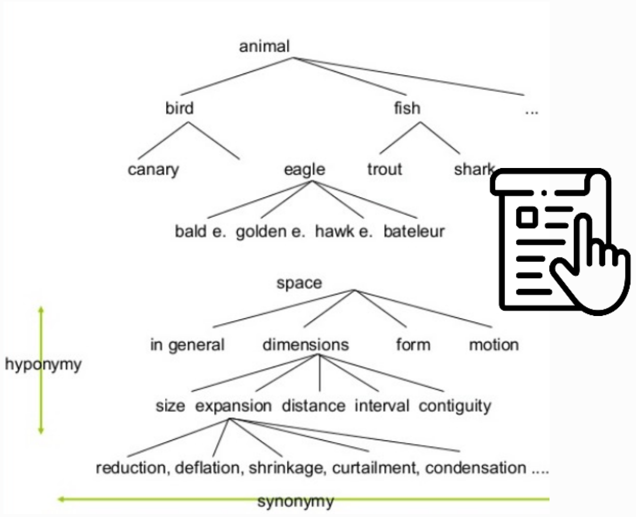
Introduction

Building hypernym relation taxonomy

- Manually building hypernym relation taxonomies is very labor-intensive and time-consuming (WordNet)[2].
- Thus, Many studies attempted to automatically extract hypernym-hyponym pair from the large corpora.

Construct manually

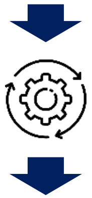
WordNet



- ✓ **Labor intensive**
- ✓ **Time consuming**
- ✓ **Diffiulty of maintenance**

Corpus

Johnson says paying bonuses to employees is not an unusual practice , especially as the state tries to compete with the private sector to keep good employees …



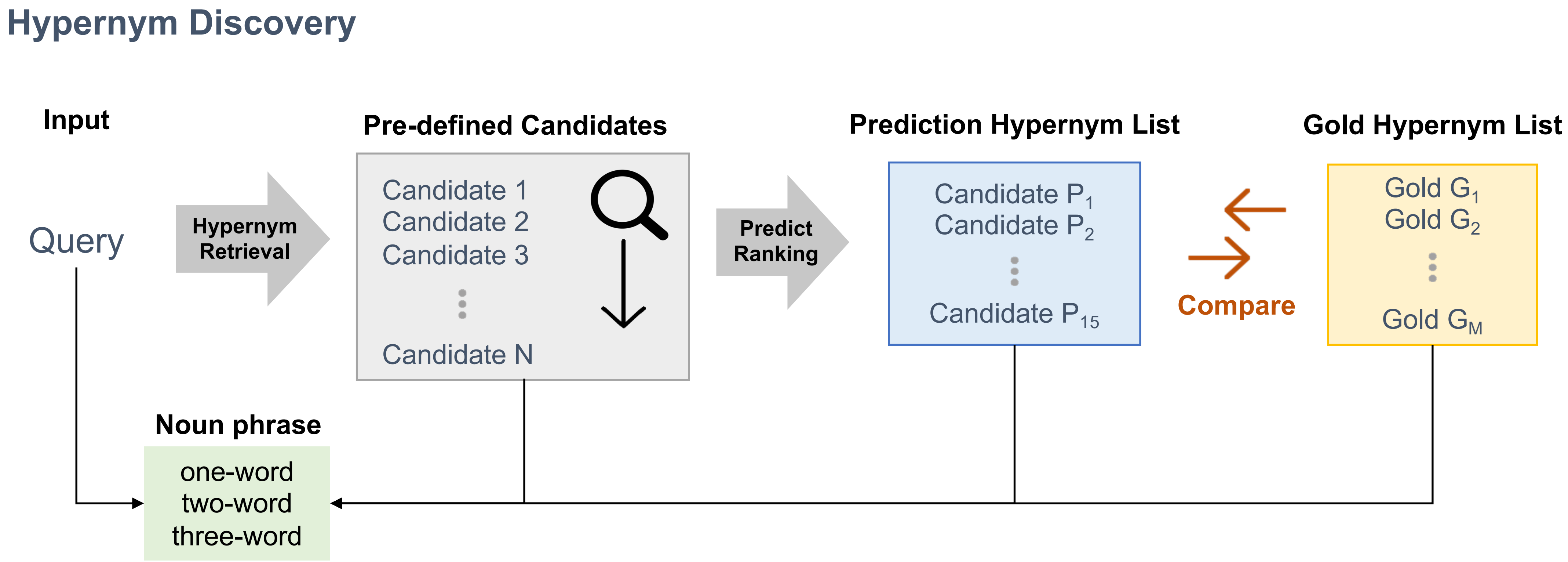
Hypernymy pair extraction

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

Introduction

Hypernym Discovery

- Corpus, training set, validation set, testing set, and candidate hypernym vocabulary are given.
- Query(hyponym) of each splits are mutual exclusive.
- It is important to capture the relation between the query and candidate hypernyms.



Related Work

Pattern-based method

Hearst pattern[3]

- Building lexical hierarchy(hypernym relation) automatically using **predefined syntactic patterns**.

- <u>**Problem of pattern-based methods**</u> : **extremely sparsity** – The sentence in which the word pair appears must match the pattern exactly.

Predefined Hearst patterns

NP_y such as NP_x

NP_x and other NP_y

NP_y including NP_x

NP_y especially NP_x

…

NP: Noun Phrase
x : hyponym , y : hypernym

"Injury, such as broken bone, is dangerous."

$\cancel{broken + bone}$

- ✓ how to combine ?
- ✓ senses of component words

pattern

$NP_0\ such\ as\ \{NP_1, NP_2, \dots, (and|or)\}\ NP_n$

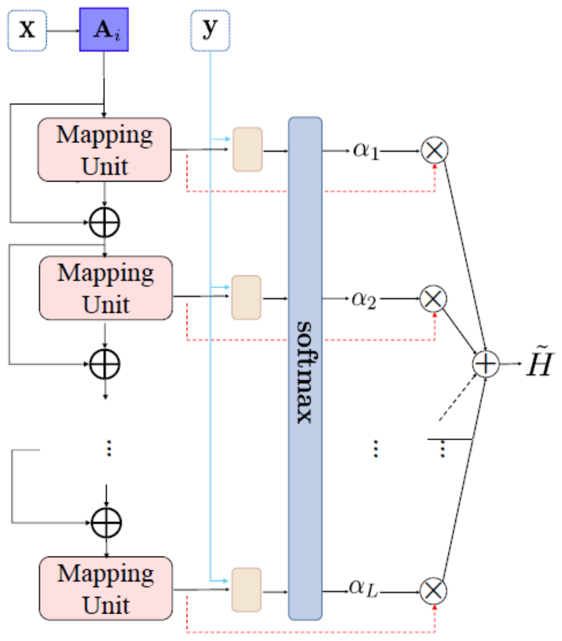
$hyponym("broken\ bone", "injury")$

Related Work

Supervised approach with word2vec embeddings

Recurrent Mapping Model (RMM)[4]

- Assume hypernym terms may come from different concept levels.
- Using recurrent **projection matrix** and **word2vec embeddings** with attention mechanism.



$$t = A_i X,\ h_0 = t$$

$$\hat{h}_{l+1} = W_\phi h_l,$$

$$h_{l+1} = \hat{h}_{l+1} + h_l$$

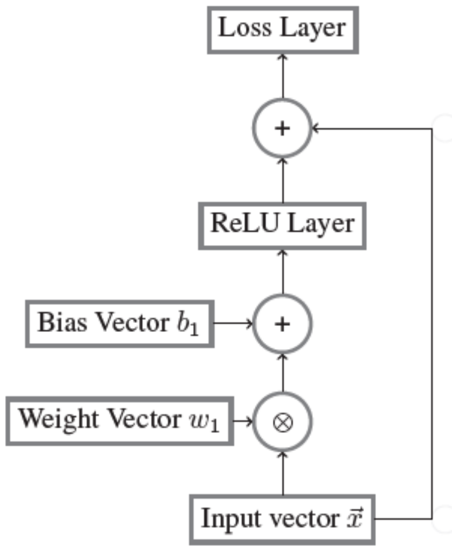
$$\alpha_l = \frac{\exp(h_l \cdot y)}{\sum_{k=1}^{L} \exp(h_k \cdot y)}$$

$$\hat{H} = \sum_{l=1}^{L} \alpha_l \cdot h_l$$

$$s_y = \hat{H} \cdot y$$

Strict Partial Order Network (SPON)[5]

- Using **word2vec embeddings** with residual connection and weight vectors.
- **Loss layer** : distance between mapped query x and candidate y



$$g(\vec{x}) = ReRU(w_1 \otimes \vec{x} + b_1)$$

$$f(\vec{x}) = \vec{x} + g(\vec{x})$$

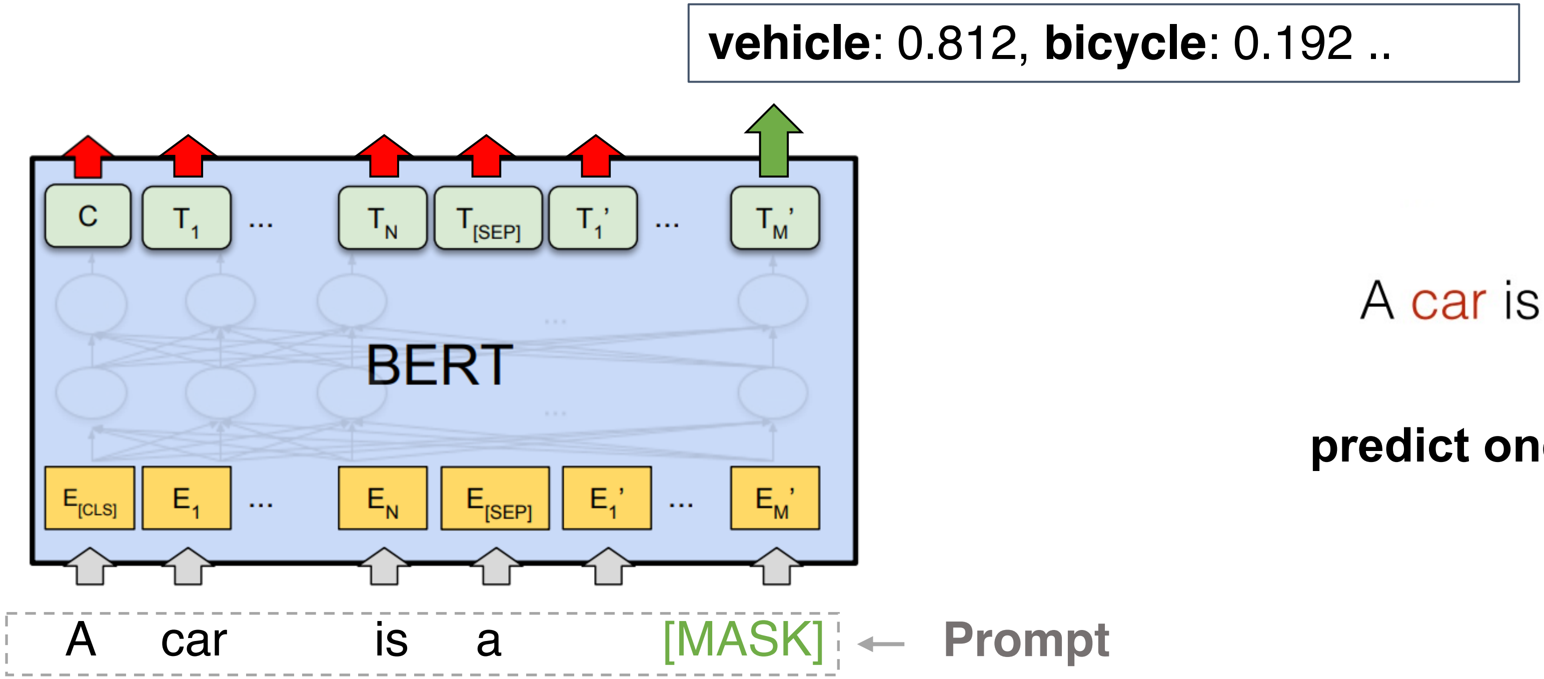
$$\psi(x, y) = \sum_{i}^{d} max(0, \epsilon + f(\vec{x})_i - \vec{y}_i)$$

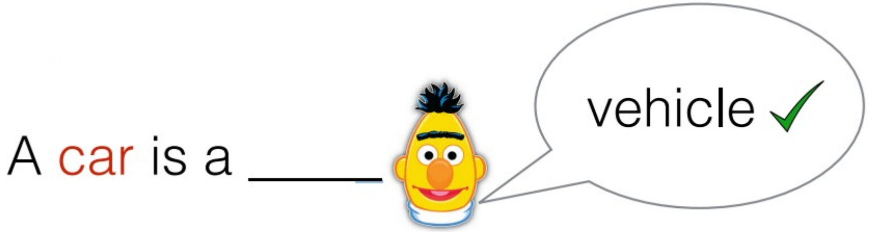
$$p(x, y) = \frac{e^{-\psi(x,y)}}{\sum_{z \in \mathcal{H}_x \cup \mathcal{H}'_x} e^{-\psi(x,z)}}$$

Related Work

Prompting Masked Language Models

- Using hypernym relation prompt to input of the language model.
- Without fine-tuning, BERT directly predicts [MASK] token of input prompt.
- Limitation of prompting masked language modeling is that the model can’t predict multiple words, but one word.



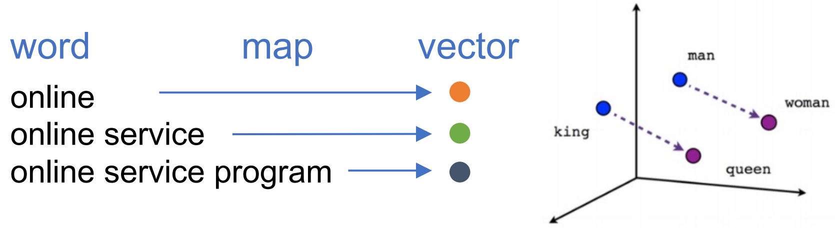


predict one word in BERT vocabulary

Common Drawback of Conventional Methods

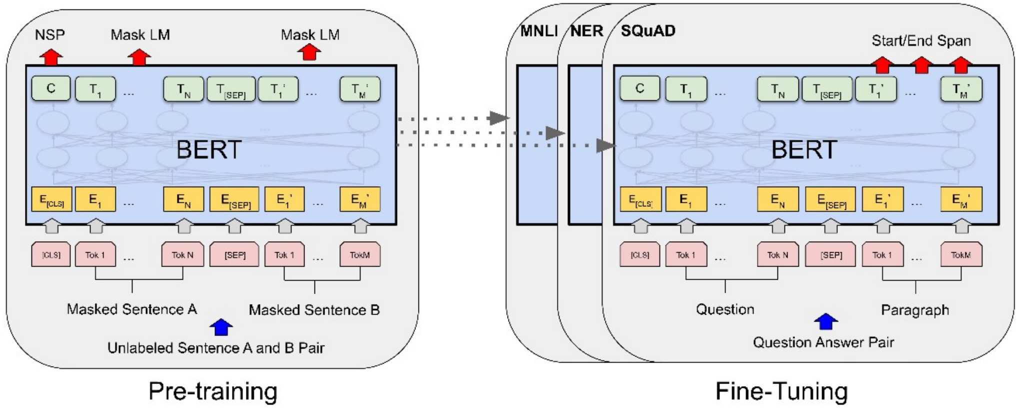
Word2vec

- Word2vec embedding maps noun phrases into a single vector.
- Thus, Word2vec embedding is vulnerable to rare or unseen word in the corpus.



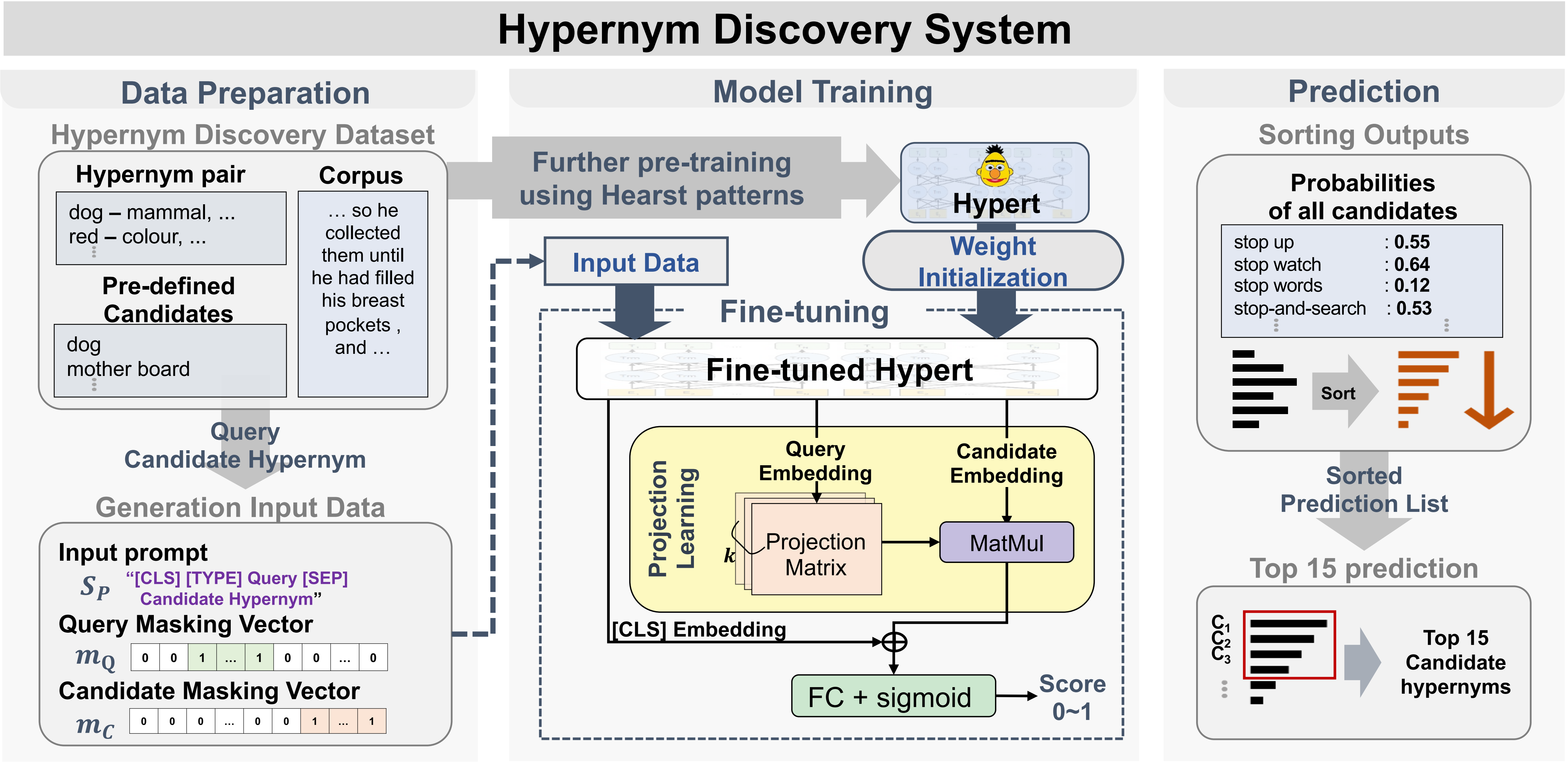
Prompting masked language models

- Language model may alleviate the problem of word2vec using subword tokenization.
- Pre-trained language model is trained on general sentences.
- General process of domain adaptation : **Pre-training** and **Fine-tuning**



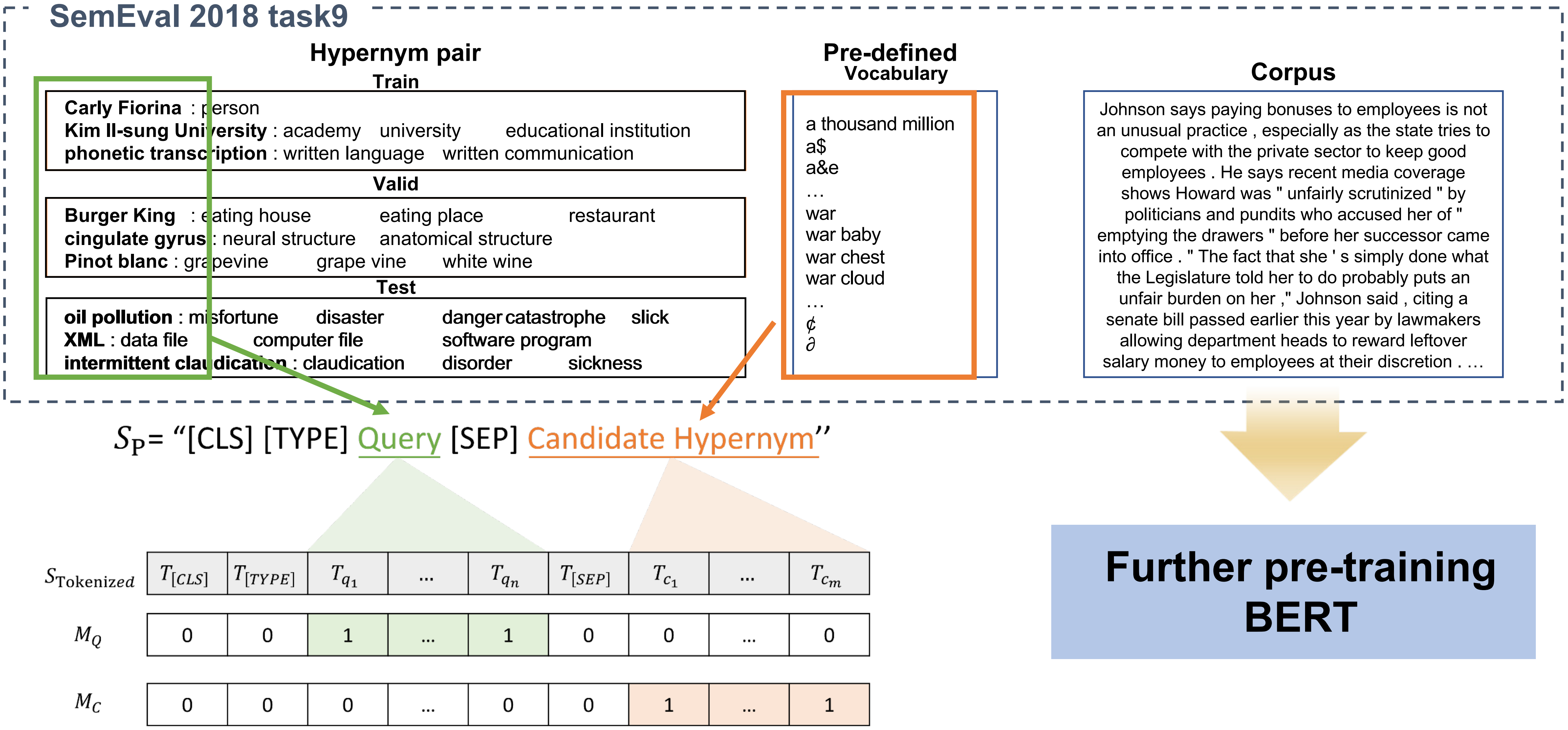
- Considering **hypernym relation as a linguistic domain**.
- **Follow the general process** of domain adaptation like BioBERT[9] and FinBERT[10].
- To construct hypernym related corpus, extended **Hearst patterns** are used for sentence extraction.

Proposed Method

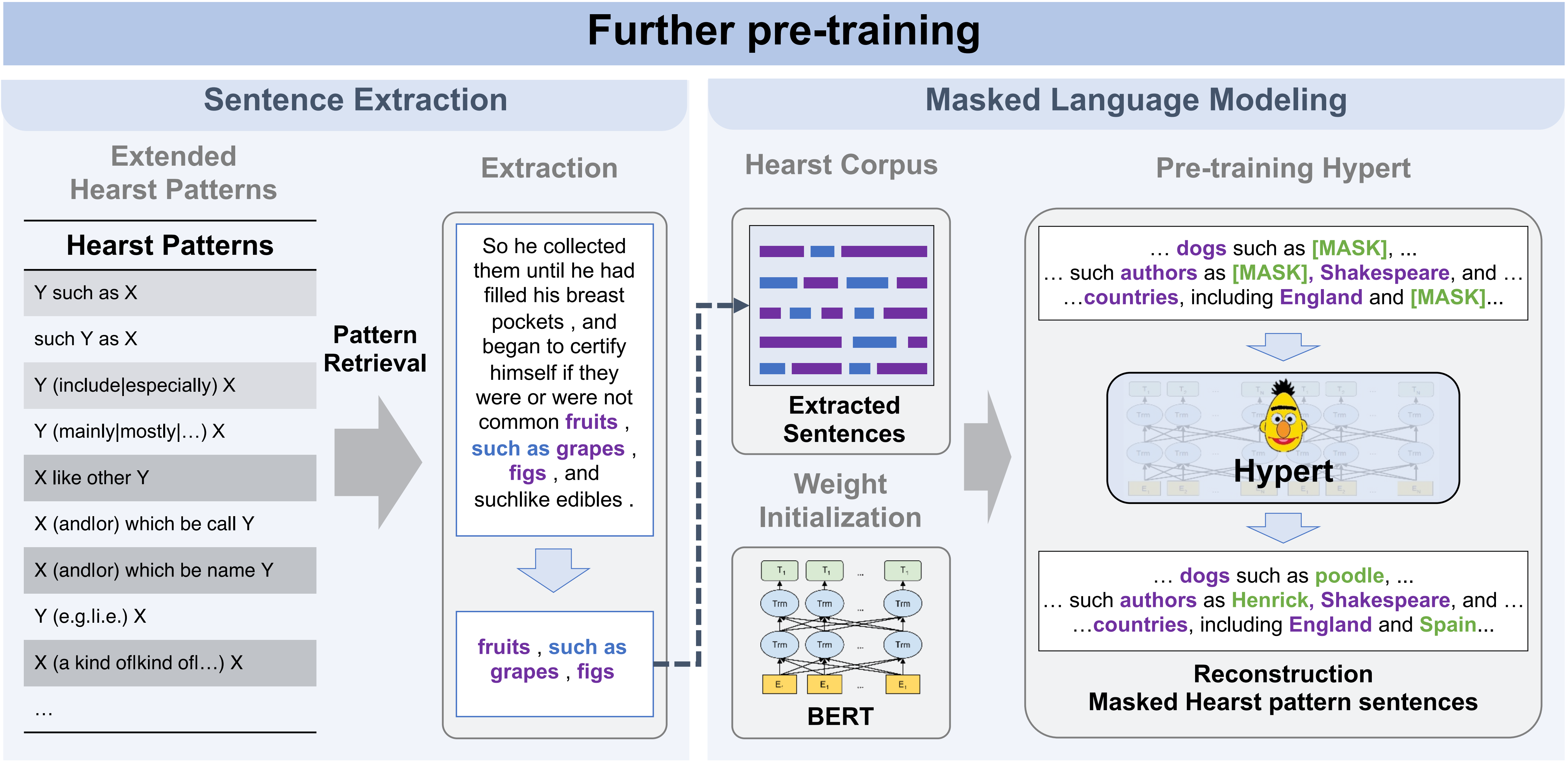


Proposed Method

Data Preparation



Proposed Method



Experimental Settings

Dataset

- SemEval-2018 task9: Hypernym Discovery[11].
- Each query has corresponding gold hypernyms up to 15.
- There are 3 sub-tasks. – 1A: General, 2A: Medical, 2B: Music

Subtask	Corpus size	# of			
		vocabulary	train	valid	test
1A (general)	16G	218,753	1,500	50	1,500
2A (Medical)	800M	93,888	500	15	500
2B (Music)	500M	69,118	500	15	500

Subtask	**Query**	**Query Type**	**Gold Hypernym(s)**	**Source**
1A General	fuse	Concept	igniter, microcomputer, electrical conductance, electronic component, party, resistor	WordNet
2A Medical	solitary pulmonary nodule	Concept	clinical finding, lung mass, single lesion, nodule	SnomedCT
2B Music	Louis Armstrong	Entity	cornetist, trumpeter, jazzman, jazz musician, musician, person	MusicBrainz

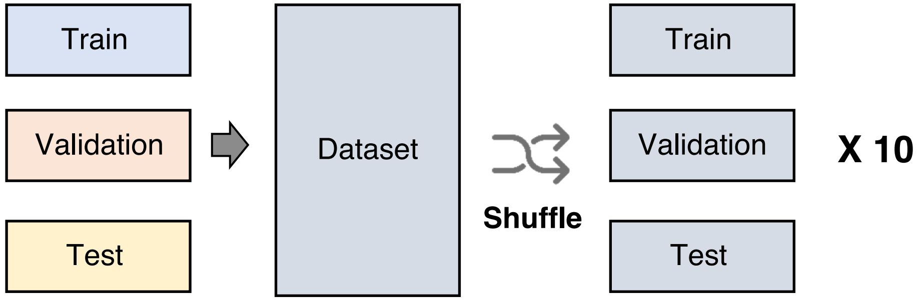
Comparison Methods

- Recurrent Mapping Model (RMM)[4]: Projection learning using word2vec embeddings.
- Strict Partial Order Network(SPON)[5]: Asymmetry and transitive properties to simple network with word2vec embeddings.
- Prompting Masked Language Model[7]: Inputting prompts with mask token into the language model.

Experimental Settings

Hyperparameter settings

- Negative sampling : 50 samples per each pair
- Maximum 1000 epochs with 200 patience for the comparison methods.
- Maximum 15 epochs for the proposed method
- 10 iterations for each dataset with hold-out cross validation.



Performance measures

- Mean Average Precision (MAP) $$\mathrm{MAP} = \frac{1}{|Q|}\sum_{q\in Q}^{|Q|}\mathrm{AP}(q)$$
- Mean Reciprocal Rank (MRR) $$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$
- Precision at k $(P@k)$ $k \in \{1, 3, 5, 15\}$ $$\mathrm{P}@k = \frac{\mathrm{TP}@k}{(\mathrm{TP}@k) + (\mathrm{FP}@k)}$$

Overall Experimental Results

Experimental Results of model performance

Subtask	Evaluation measures	Proposed method	RMM [13]	SPON [14]	Prompting BERT [16]	
					is a	such as
1A General	MRR	**38.68** ± **2.00**	27.21 ± 3.50	24.94 ± 4.10	19.74 ± 0.44	19.77 ± 0.41
	MAP	**24.17** ± **1.26**	18.25 ± 1.16	15.72 ± 1.75	11.43 ± 0.16	10.93 ± 0.19
	P@1	**29.57** ± **1.98**	18.71 ± 4.40	17.02 ± 3.95	12.10 ± 0.58	13.54 ± 0.50
	P@3	**21.56** ± **1.38**	15.57 ± 1.76	13.53 ± 2.33	10.19 ± 0.22	9.64 ± 0.25
	P@5	**21.27** ± **1.25**	16.14 ± 1.20	13.77 ± 1.87	10.20 ± 0.19	9.31 ± 0.14
	P@15	**27.52** ± **1.35**	21.47 ± 2.12	18.42 ± 1.16	13.02 ± 0.23	12.73 ± 0.22
2A Medical	MRR	**64.83** ± **3.32**	42.25 ± 2.01	46.54 ± 2.89	49.28 ± 1.60	41.40 ± 1.53
	MAP	**50.24** ± **2.29**	32.44 ± 2.72	34.19 ± 1.87	21.99 ± 0.73	18.75 ± 0.67
	P@1	**53.28** ± **3.86**	30.12 ± 2.44	35.26 ± 3.31	38.86 ± 1.65	30.62 ± 1.53
	P@3	**46.69** ± **3.52**	28.68 ± 2.82	32.63 ± 2.63	25.07 ± 1.09	19.83 ± 0.94
	P@5	**46.60** ± **2.91**	29.18 ± 2.73	32.02 ± 2.11	21.18 ± 0.92	17.30 ± 0.74
	P@15	**54.69** ± **1.74**	37.72 ± 3.40	37.35 ± 1.55	19.28 ± 0.63	17.81 ± 0.56
2B Music	MRR	**67.43** ± **2.37**	54.37 ± 3.06	60.47 ± 3.91	19.84 ± 0.98	19.54 ± 0.97
	MAP	**55.03** ± **1.98**	47.52 ± 2.75	48.38 ± 2.07	8.91 ± 0.49	8.42 ± 0.44
	P@1	**56.68** ± **2.98**	42.52 ± 3.45	48.34 ± 5.34	12.32 ± 0.65	12.42 ± 0.76
	P@3	**52.94** ± **2.05**	44.53 ± 2.51	45.29 ± 3.25	9.08 ± 0.61	8.64 ± 0.37
	P@5	**52.92** ± **2.37**	45.24 ± 2.92	46.00 ± 2.23	8.55 ± 0.47	7.72 ± 0.48
	P@15	**58.59** ± **2.05**	52.72 ± 2.90	52.81 ± 1.00	8.86 ± 0.62	8.54 ± 0.53

Overall Experimental Results

Experimental Results of statistical test: Wilcoxon signed-rank test

1A Dataset

Comparison methods		Evaluation measures	Proposed *versus* (*p*-value)
RMM [1]		MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
SPON [2]		MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
Prompting BERT [3]	is a	MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
	such as	MRR	**Win**(1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)

2A Dataset

Comparison methods		Evaluation measures	Proposed *versus* (*p*-value)
RMM [1]		MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
SPON [2]		MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
Prompting BERT [3]	is a	MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
	such as	MRR	**Win**(1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)

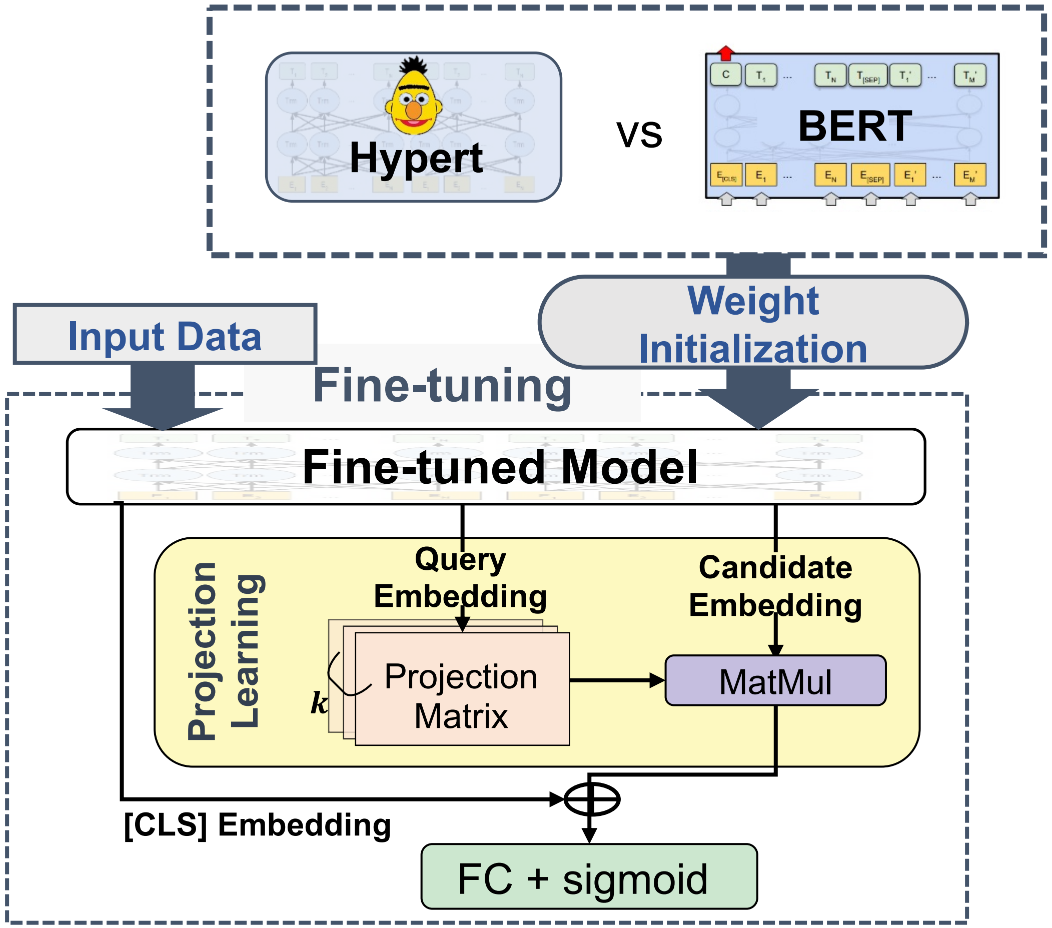
2B Dataset

Comparison methods		Evaluation measures	Proposed *versus* (*p*-value)
RMM [1]		MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
SPON [2]		MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
Prompting BERT [3]	is a	MRR	**Win** (1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)
	such as	MRR	**Win**(1.95e-3)
		MAP	**Win** (1.95e-3)
		P@1	**Win** (1.95e-3)
		P@3	**Win** (1.95e-3)
		P@5	**Win** (1.95e-3)
		P@15	**Win** (1.95e-3)

In-Depth Analysis

Effectiveness of further pre-training method

- Comparing proposed model with Hypert and BERT.



Subtask	Evaluation measures	Hypert model	BERT model
1A General	MRR	**38.68 ± 2.00**✓	36.44 ± 2.12
	MAP	**24.17 ± 1.26**✓	23.29 ± 0.78
	P@1	**29.57 ± 1.98**✓	26.38 ± 2.93
	P@3	**21.56 ± 1.38**✓	20.90 ± 1.00
	P@5	**21.27 ± 1.25**✓	20.68 ± 0.79
	P@15	**27.52 ± 1.35**✓	26.63 ± 0.75
2A Medical	MRR	**64.83 ± 3.32**✓	62.62 ± 3.20
	MAP	**50.24 ± 2.29**✓	48.85 ± 1.57
	P@1	**53.28 ± 3.86**✓	49.94 ± 4.35
	P@3	**46.69 ± 3.52**✓	45.45 ± 2.27
	P@5	**46.60 ± 2.91**✓	45.66 ± 1.74
	P@15	**54.69 ± 1.74**✓	53.36 ± 0.97
2B Music	MRR	**67.43 ± 2.37**✓	63.19 ± 5.38
	MAP	**55.03 ± 1.98**✓	49.70 ± 3.37
	P@1	**56.68 ± 2.98**✓	50.92 ± 7.31
	P@3	**52.94 ± 2.05**✓	46.88 ± 4.28
	P@5	**52.92 ± 2.37**✓	47.27 ± 3.59
	P@15	**58.59 ± 2.05**✓	53.97 ± 2.48

In-Depth Analysis

Prediction words for the rare query

- Query : "***open proxy server***" appeared 9 times in corpus.
- **bold** with ⋆ symbol for gold hypernyms, underlines for the relevant words.
- The results show that the proposed method predicts adequately for the rare word compared to others.

Model	Predicted
Gold	**proxy server**⋆, **service**⋆, **software program**⋆, **software package**⋆, **software**⋆, **application**⋆, **application software**⋆, **computer program**⋆, **software application**⋆
Proposed	**service**⋆, **software**⋆, server, **computer program**⋆, facility, protocol, **software program**⋆, platform, company, client, **software application**⋆, host, **application**⋆, computer system, application program
RMM	person, pseudonymized, granularly, open proxy, ad filtering, intrabank, otherwise, spoofing attack, actual, statefully, sdl plc, authenticatable, ip address spoofing, example, solely
SPON	person, constructed structure, technical specification, town, single-valued function, city, juridical person, leader, animal, state, boss, company, **computer program**⋆, measure, political leader

Conclusion and Contribution

Conclusion

- An Additional pre-training method using Hearst pattern sentence extraction is proposed to enhance hypernym awareness of the language model.
- The proposed model outperforms the existing models.
- An in-depth analysis is conducted to confirm the effectiveness of the further pre-training method compared to vanilla BERT.
- The analysis shows that the proposed model is robust for rare or unseen words.

Achievement

- In Domestic Competition
  - B-cell epitope (linear peptide) prediction competition, AI Graduate School Challenge, 2022, **2^{nd} prize (LG AI Director Award)**
- In International Conferences
  - Improved Phrase Embedding for Hypernym Relation Discovery, ***ICEIC***, 2023
- In Domestic Conferences
  - Comparison of word2vec-based methods for Hypernym Discovery, ***IEIE***, 2022
- Participation in corporate internships. – **S2W**, NLP team, 2022.Jan - 2022.Feb, cyberthreat-detection from twitter data.

Thank you

References

[1] Yamane, Josuke, et al. "Distributional hypernym generation by jointly learning clusters and projections." *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016.

[2] Oram, Peter. "WordNet: An electronic lexical database. Christiane Fellbaum (Ed.). Cambridge, MA: MIT Press, 1998. Pp. 423." *Applied Psycholinguistics* 22.1 (2001): 131-134.

[3] Hearst, Marti A. "Automatic acquisition of hyponyms from large text corpora." *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. 1992.

[4] Bai, Yuhang, et al. "Hypernym Discovery via a Recurrent Mapping Model." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021.

[5] Dash, Sarthak, et al. "Hypernym detection using strict partial order networks." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 2020.

[6] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26 (2013).

[7] Hanna, Michael, and David Mareček. "Analyzing BERT’s knowledge of hypernymy via prompting." *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. 2021.

[8] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT*. 2019.

[9] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.

[10] Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." *arXiv preprint arXiv:1908.10063* (2019).

[11] Camacho-Collados, Jose, et al. "SemEval-2018 task 9: Hypernym discovery." *Proceedings of the 12th international workshop on semantic evaluation*. 2018.

Appendix

Extended Hearst Patterns

<table>
<tr><th>No.</th><th>Pattern</th></tr>
<tr><td>1</td><td>(NP_“w+ (,)?such as (NP_“w+ ?(,)?(and |or)?)+)</td></tr>
<tr><td>2</td><td>(such NP_“w+ (,)?as (NP_“w+ ?(,)?(and |or)?)+)</td></tr>
<tr><td>3</td><td>((NP_“w+ ?(,)?)+(and |or)?other NP_“w+)</td></tr>
<tr><td>4</td><td>(NP_“w+ (,)?include (NP_“w+ ?(,)?(and |or)?)+)</td></tr>
<tr><td>5</td><td>(NP_“w+ (,)?especially (NP_“w+ ?(,)?(and |or)?)+)</td></tr>
<tr><td>6</td><td>((NP_“w+ ?(,)?)+(and |or)?any other NP_“w+)</td></tr>
<tr><td>7</td><td>((NP_“w+ ?(,)?)+(and |or)?some other NP_“w+)</td></tr>
<tr><td>8</td><td>((NP_“w+ ?(,)?)+(and |or)?be a NP_“w+)</td></tr>
<tr><td>9</td><td>(NP_“w+ (,)?like (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>10</td><td>such (NP_“w+ (,)?as (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>11</td><td>((NP_“w+ ?(,)?)+(and |or)?like other NP_“w+)</td></tr>
<tr><td>12</td><td>((NP_“w+ ?(,)?)+(and |or)?one of the NP_“w+)</td></tr>
<tr><td>13</td><td>((NP_“w+ ?(,)?)+(and |or)?one of these NP_“w+)</td></tr>
<tr><td>14</td><td>((NP_“w+ ?(,)?)+(and |or)?one of those NP_“w+)</td></tr>
<tr><td>15</td><td>example of (NP_“w+ (,)?be (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>16</td><td>((NP_“w+ ?(,)?)+(and |or)?be example of NP_“w+)</td></tr>
<tr><td>17</td><td>(NP_“w+ (,)?for example (,)?(NP_“w+ ?(,)?(and |or)?)+)</td></tr>
<tr><td>18</td><td>((NP_“w+ ?(,)?)+(and |or)?which be call NP_“w+)</td></tr>
<tr><td>19</td><td>((NP_“w+ ?(,)?)+(and |or)?which be name NP_“w+)</td></tr>
<tr><td>20</td><td>(NP_“w+ (,)?mainly (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>21</td><td>(NP_“w+ (,)?mostly (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>22</td><td>(NP_“w+ (,)?notably (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>23</td><td>(NP_“w+ (,)?particularly (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>24</td><td>(NP_“w+ (,)?principally (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>25</td><td>(NP_“w+ (,)?in particular (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>26</td><td>(NP_“w+ (,)?except (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>27</td><td>(NP_“w+ (,)?other than (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>28</td><td>(NP_“w+ (,)?e.g. (,)?(NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>29</td><td>(NP_“w+ “((e.g.—i.e.) (,)?(NP_“w+ ? (,)?(and |or)?)+(“.)?“))</td></tr>
<tr><td>30</td><td>(NP_“w+ (,)?i.e. (,)?(NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>31</td><td>((NP_“w+ ?(,)?)+(and—or)? a kind of NP_“w+)</td></tr>
<tr><td>32</td><td>((NP_“w+ ?(,)?)+(and—or)? kind of NP_“w+)</td></tr>
<tr><td>33</td><td>((NP_“w+ ?(,)?)+(and—or)? form of NP_“w+)</td></tr>
<tr><td>34</td><td>((NP_“w+ ?(,)?)+(and |or)?which look like NP_“w+)</td></tr>
<tr><td>35</td><td>((NP_“w+ ?(,)?)+(and |or)?which sound like NP_“w+)</td></tr>
<tr><td>36</td><td>(NP_“w+ (,)?which be similar to (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>37</td><td>(NP_“w+ (,)?example of this be (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>38</td><td>(NP_“w+ (,)?type (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>39</td><td>((NP_“w+ ?(,)?)+(and |or)? NP_“w+ type)</td></tr>
<tr><td>40</td><td>(NP_“w+ (,)?whether (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>41</td><td>(compare (NP_“w+ ?(,)?)+(and |or)?with NP_“w+)</td></tr>
<tr><td>42</td><td>(NP_“w+ (,)?compare to (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>43</td><td>(NP_“w+ (,)?among -PRON- (NP_“w+ ? (,)?(and |or)?)+)</td></tr>
<tr><td>44</td><td>((NP_“w+ ?(,)?)+(and |or)?as NP_“w+)</td></tr>
<tr><td>45</td><td>(NP_“w+ (,)? (NP_“w+ ? (,)?(and |or)?)+ for instance)</td></tr>
<tr><td>46</td><td>((NP_“w+ ?(,)?)+(and—or)? sort of NP_“w+)</td></tr>
<tr><td>47</td><td>(NP_“w+ (,)?which may include (NP_“w+ ?(,)?(and |or)?)+)</td></tr>
</table>

NP: Noun Phrase

PRON: Pronoun

Appendix

Model Performance: 0k to 10k pre-training step

Subtask	Pre-training steps	Evaluation measures					
		MRR	MAP	P@1	P@3	P@5	P@15
1A General	0k	36.44 ± 2.12	23.29 ± 0.78	26.38 ± 2.93	20.90 ± 1.00	20.68 ± 0.79	26.63 ± 0.75
	1k	**38.68 ± 2.00**	24.17 ± 1.26	**29.57 ± 1.98**	21.56 ± 1.38	21.27 ± 1.25	27.52 ± 1.35
	2k	37.52 ± 2.02	23.59 ± 0.98	28.37 ± 2.64	21.01 ± 1.05	20.83 ± 0.84	26.79 ± 1.16
	3k	37.71 ± 1.08	23.76 ± 0.95	28.36 ± 1.39	21.25 ± 0.99	21.04 ± 0.93	26.95 ± 1.11
	4k	37.49 ± 1.05	23.69 ± 0.86	28.03 ± 1.33	21.19 ± 0.94	20.88 ± 0.87	26.89 ± 0.95
	5k	38.06 ± 2.04	**24.51 ± 0.73**	28.42 ± 2.53	21.81 ± 1.25	21.61 ± 0.87	**27.96 ± 0.57**
	6k	38.46 ± 2.33	24.36 ± 1.18	29.13 ± 2.62	21.93 ± 1.63	21.58 ± 1.18	27.52 ± 0.96
	7k	37.75 ± 1.49	23.89 ± 1.23	28.45 ± 1.52	21.48 ± 1.16	21.12 ± 1.11	27.12 ± 1.42
	8k	37.69 ± 3.08	23.89 ± 1.63	28.46 ± 3.43	21.32 ± 2.18	21.13 ± 1.68	27.17 ± 1.10
	9k	38.67 ± 1.59	24.46 ± 0.89	29.38 ± 1.82	21.91 ± 1.08	**21.73 ± 0.80**	27.62 ± 0.91
	10k	37.85 ± 1.94	24.40 ± 0.80	27.87 ± 2.78	**21.93 ± 1.18**	21.64 ± 0.82	27.80 ± 0.70
2A Medical	0k	62.62 ± 3.20	48.85 ± 1.57	49.94 ± 4.35	45.45 ± 2.27	45.66 ± 1.74	53.36 ± 0.97
	1k	64.83 ± 3.32	50.24 ± 2.29	53.28 ± 3.86	46.69 ± 3.52	**46.60 ± 2.91**	54.69 ± 1.74
	2k	**66.52 ± 2.44**	50.48 ± 1.63	55.64 ± 3.43	46.97 ± 1.56	46.22 ± 1.86	54.80 ± 1.69
	3k	60.91 ± 4.65	47.41 ± 2.60	49.40 ± 5.70	43.50 ± 2.89	43.47 ± 2.68	52.46 ± 2.42
	4k	65.10 ± 4.13	49.60 ± 2.50	54.62 ± 4.92	45.80 ± 2.97	45.21 ± 2.48	53.98 ± 2.00
	5k	64.04 ± 3.20	49.25 ± 1.38	53.34 ± 4.37	45.32 ± 2.04	44.83 ± 1.46	53.98 ± 1.10
	6k	66.21 ± 2.66	50.15 ± 2.23	55.68 ± 2.69	**46.98 ± 2.46**	46.05 ± 2.35	54.34 ± 2.12
	7k	64.74 ± 3.23	49.62 ± 2.04	54.28 ± 3.84	45.94 ± 3.08	45.36 ± 2.57	54.42 ± 1.27
	8k	66.34 ± 3.81	**50.52 ± 2.15**	**56.16 ± 4.95**	46.98 ± 2.65	46.28 ± 2.26	54.82 ± 1.83
	9k	64.42 ± 2.52	50.15 ± 1.52	53.18 ± 4.00	46.15 ± 1.78	45.85 ± 2.06	**55.01 ± 1.06**
	10k	63.44 ± 2.91	48.56 ± 1.34	52.62 ± 3.52	44.90 ± 2.34	43.92 ± 1.72	53.40 ± 1.46
2B Music	0k	63.19 ± 5.38	49.70 ± 3.37	50.92 ± 7.31	46.88 ± 4.28	47.27 ± 3.59	53.97 ± 2.48
	1k	**67.43 ± 2.37**	**55.03 ± 1.98**	**56.68 ± 2.98**	**52.94 ± 2.05**	**52.92 ± 2.37**	**58.59 ± 2.05**
	2k	65.71 ± 3.04	53.97 ± 2.24	53.84 ± 4.03	51.13 ± 2.99	51.91 ± 2.39	58.37 ± 1.79
	3k	62.93 ± 4.22	50.53 ± 3.75	50.42 ± 5.50	47.29 ± 4.56	47.95 ± 4.17	55.36 ± 2.78
	4k	65.63 ± 4.65	53.55 ± 3.27	53.86 ± 5.92	50.98 ± 3.89	51.67 ± 3.55	57.64 ± 2.48
	5k	66.89 ± 3.31	53.96 ± 2.33	55.62 ± 4.71	51.45 ± 2.99	51.69 ± 2.47	58.25 ± 1.51
	6k	67.40 ± 3.41	54.53 ± 3.23	56.42 ± 4.46	52.24 ± 4.52	52.25 ± 3.94	58.49 ± 2.41
	7k	66.09 ± 4.51	53.43 ± 3.02	54.84 ± 5.85	51.28 ± 4.33	51.27 ± 3.25	57.46 ± 1.96
	8k	67.21 ± 3.47	54.28 ± 1.93	56.18 ± 4.74	51.76 ± 3.20	52.17 ± 2.01	58.20 ± 1.29
	9k	65.25 ± 6.19	52.37 ± 4.57	53.48 ± 8.48	49.46 ± 6.50	49.79 ± 5.01	56.87 ± 3.42
	10k	64.54 ± 5.24	51.69 ± 3.77	52.92 ± 6.93	48.86 ± 5.33	49.52 ± 4.61	56.04 ± 2.52