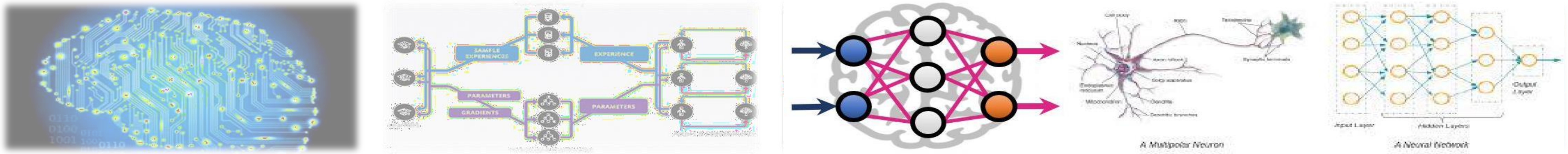


Review : Automatic Hypernym Detection from Large Text Corpora

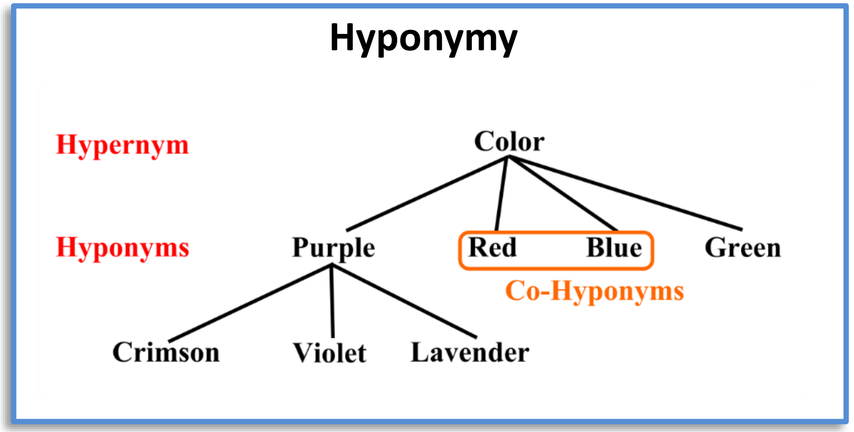
Automatic Hypernym Detection from Large Text corpora 리뷰

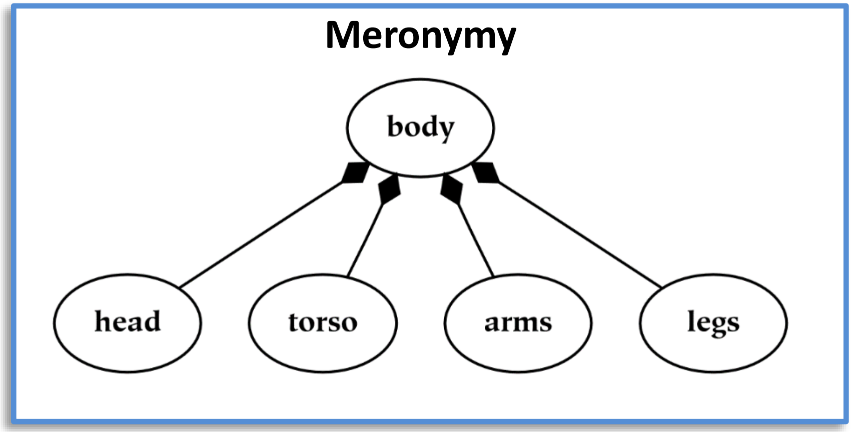
Geonil Yun,
rjsdlf45@gmail.com
2021. 09. 07.

Introduction

- **Hierarchical relationship**

  - Important task in NLP





Introduction

- **Hierarchical relationship**
  - Hypernym detection
    - Pattern-based method (Hearst)

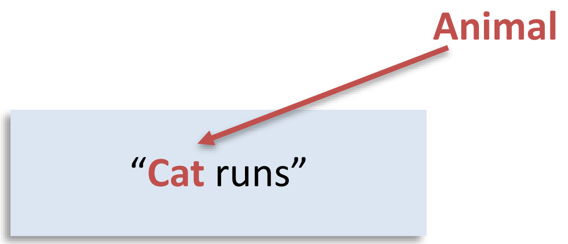
NP_y such as NP_x

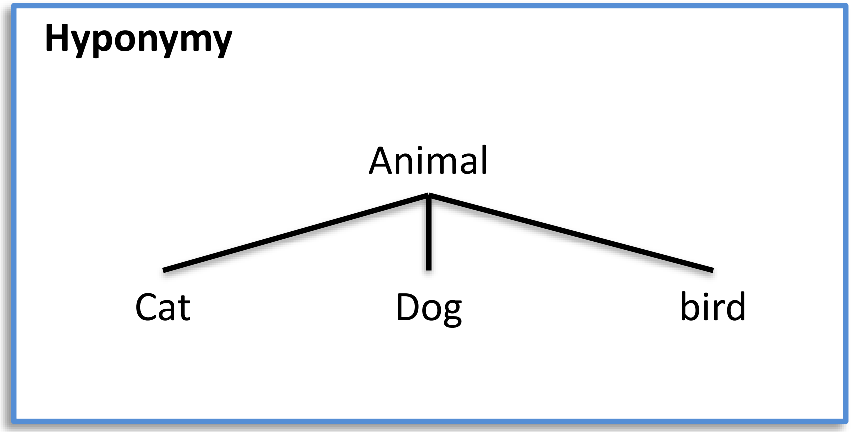
NP_x and other NP_y

- Key idea : exploit **lexico-syntactic patterns** to detect is-a relations(hyponymy relation).
- Problem of Pattern-based : **sparsity** – words must co-occur exactly the right configuration.

Introduction

- **Hierarchical relationship**
  - Hypernym detection
    - ✓ Distributional representation
      - It solves sparsity issue, but require specialized similarity measures.
      - Most successful measures are inspired by Distributional Inclusion Hypothesis(DIH)





Introduction

- **Hierarchical relationship**

  - Hypernym detection

Hypernym detection	**Pattern-based methods**	**DIH methods**
Similarity	rely on co-occurrences within certain context	
Difference	**Predefined** patterns to generate **high-precision** extractions	**Unconstrained** word co-occurrences

    - Evaluate pattern-based models and compare with DIH.
    - In detection, direction prediction, and graded entailment tasks.

  - In this paper,

    - Simple pattern-based methods consistently outperform DIH methods.
    - Low-rank embedding of pattern-based models improves sparsity problem

Models

- **Pattern-based Hypernym Detection**

  - Notation

    $P = \{(x, y)\}_{i=1}^{n}$, set of hypernymy relations. (extracted by Hearst patterns from large corpus $\mathcal{T}$)
    $w(x, y)$, count of (x, y) has been extracted.
    $W = \sum_{(x,y)\in P} w(x, y)$, total number extractions.

  - Direct application of Hearst patterns. (counts)

$$\boldsymbol{p(x, y) = \frac{w(x, y)}{W}}$$

    - **Skewed** by the **occurrence probabilities** of their constituent words.

$$hyponym(France, country) > hyponym(France, republic)$$

Models

- **Pattern-based Hypernym Detection**

  - Positive Pointwise Mutual Information (PPMI)

$$\boldsymbol{ppmi(x,y)} = \mathbf{max}\left(\mathbf{0}, \mathbf{log}\frac{\boldsymbol{p(x,y)}}{\boldsymbol{p^{-}(x)p^{+}(y)}}\right)$$

$$p^{-}(x) = \sum_{(x,y)\in P} \frac{w(x,y)}{W}$$
$$p^{+}(x) = \sum_{(y,x)\in P} \frac{w(y,x)}{W}$$

    - Cannot handle **missing data**.
    - To handle unseen pairs, **use low-rank embeddings of PPMI matrix**.

Models

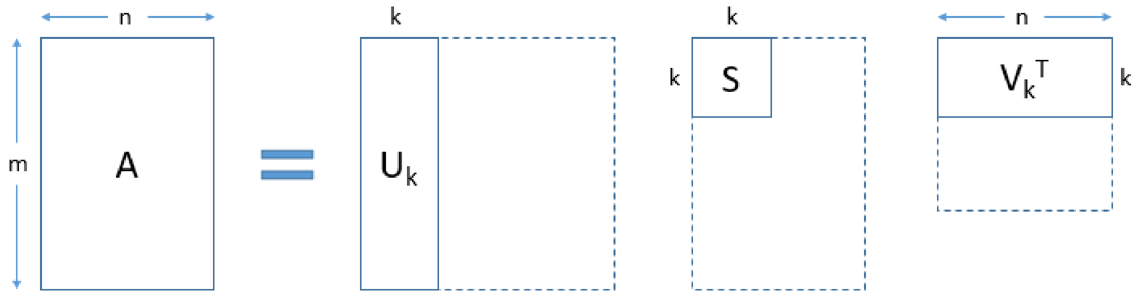
- **Pattern-based Hypernym Detection**

  - Low-rank embeddings of the PPMI matrix

    $m = |\{x : (x, y) \in P \vee (y, x) \in P\}|$, number of unique terms in P
    $X \in \mathbb{R}^{m \times m}, M_{xy} = ppmi(x, y)$

$$\boldsymbol{spmi(x, y)} = \mathbf{u}_x^{\mathbf{T}} \boldsymbol{\Sigma}_r \mathbf{v}_y$$

$$\boldsymbol{sp(x, y)} = \mathbf{u}_x^{\mathbf{T}} \boldsymbol{\Sigma}_r \mathbf{v}_y$$



▪ Using **truncated Singular Value Decomposition**
▪ Meaning x-th and y-th row of **U** and **V**

Models

- **Distributional Hypernym Detection**

  - WeedsPrec

$$\boldsymbol{WeedsPrec(x,y)} = \frac{\sum_i^n \boldsymbol{x_i} * \boldsymbol{1_{y_i>0}}}{\boldsymbol{x}}$$

  - invCL

$$\boldsymbol{invCL(x,y)} = \sqrt{\boldsymbol{CL(x,y)} * \left(\boldsymbol{1} - \boldsymbol{CL(y,x)}\right)}$$

$$CL(x,y) = \frac{\sum_i^n x_i * 1_{y_i>0}}{x}$$

  - SLQS

$$\boldsymbol{SLQS(x,y)} = \boldsymbol{1} - \frac{\boldsymbol{E_x}}{\boldsymbol{E_y}}$$

$$E_x = \text{median}_{i=1}^{N}[H(c_i)]$$

  - Use 4 methods **WeedsPec**, **invCL**, **SLQS**, **cosine**
  - SLQS model based **informativeness hypothesis**

Evaluation

- **Tasks**

  - Detection : classify whether pairs of words are in a hypernymy relation.

    - BLESS : noun-noun pairs. 14,542 total pairs with 1,337 positive examples.
    - LEDS : 2,770 nun pairs balanced between positive, negative pairs.
    - EVAL : 7,378 pairs hypernymy, synonymy, antonymy, meronymy..
    - SHWARTZ : The largest dataset(52,578 pairs), which collected from a mixture of WordNet, DBPedia and others.
    - WBLESS : 1,668 pairs subset of BLESS

  - Direction : identify which term is broader in a given pair of words.

    - BLESS : Predict all 1,337 pair positive examples.
    - WBLESS: 1,668 pairs subset of BLESS
    - BIBLESS: variant of WBLESS an

BLESS	turtle—animal	1
WBLESS	owl—creature	1
	owl—vulture	0
	animal—owl	0
BIBLESS	owl—creature	1
	owl—vulture	0
	animal—owl	-1

  - Graded Entailment : quantify the degree to which a hypernymy relation holds.

    - HYPERLEX : 2,163 noun pairs. $x\ is-a\ y$ holds on a scale of [0, 6]

Evaluation

- **Experimental setup**

  - Pattern-based models

    - Use concatenation of Gigaword and Wikipedia.
    - Remove pairs which were not extracted by at least **two distinct patterns**.
    - Remove any pair (y, x) if $\boldsymbol{p}(\boldsymbol{y}, \boldsymbol{x}) < \boldsymbol{p}(\boldsymbol{x}, \boldsymbol{y})$
    - 4.5 M matched pairs, 431K unique pairs, 243K unique terms.
    - For SVD-based models, rank $\boldsymbol{r} \in \{\mathbf{5}, \mathbf{10}, \mathbf{15}, \mathbf{20}, \mathbf{25}, \mathbf{50}, \mathbf{100}, \mathbf{150}, \mathbf{200}, \mathbf{250}, \mathbf{300}, \mathbf{500}, \mathbf{1000}\}$

Pattern
X which is a (example|class|kind|...) of Y
X (and|or) (any|some) other Y
X which is called Y
X is JJS (most)? Y
X a special case of Y
X is an Y that
X is a !(member|part|given) Y
!(features|properties) Y such as X_1, X_2, ...
(Unlike|like) (most|all|any|other) Y, X
Y including X_1, X_2, ...

  - Distributional models

    - Employ large, sparse distributional space which is computed from UkWaC and Wikipedia.
    - Targets and contexts which appear fewer than 100 times are filtered.
    - 218K words over 732K context dimensions.
    - For SLQS model, N same option SVD rank r

Results

- **Compare pattern-based models and distributional models**

	Detection (AP)					**Direction** (Acc.)			**Graded** (ρ_s)
	BLESS	EVAL	LEDS	SHWARTZ	WBLESS	BLESS	WBLESS	BIBLESS	HYPERLEX
Cosine	.12	.29	.71	.31	.53	.00	.54	.52	.14
WeedsPrec	.19	.39	.87	.43	.68	.63	.59	.45	.43
invCL	.18	.37	**.89**	.38	.66	.64	.60	.47	.43
SLQS	.15	.35	.60	.38	.69	.75	.67	.51	.16
p(x, y)	.49	.38	.71	.29	.74	.46	.69	.62	**.62**
ppmi(x, y)	.45	.36	.70	.28	.72	.46	.68	.61	.60
sp(x, y)	.66	.45	.81	.41	.91	**.96**	.84	.80	.51
spmi(x, y)	**.76**	**.48**	.84	**.44**	**.96**	**.96**	**.87**	**.85**	.53

  - Hearst patterns can provide strong performance on large corpora.

References

1) Roller, S., Kiela, D., & Nickel, M. (2018). Hearst patterns revisited: Automatic hypernym detection from large text corpora. *arXiv preprint arXiv:1806.03191*.
2) Kiela, D., Rimell, L., Vulic, I., & Clark, S. (2015). Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)* (pp. 119-124). ACL; East Stroudsburg, PA.

Thank you