Adopting Description Logics to NLP: Towards the Formal Semantics of the Natural Language

2021. 01. 14

신승우, 오주민, 노형종, 이연수

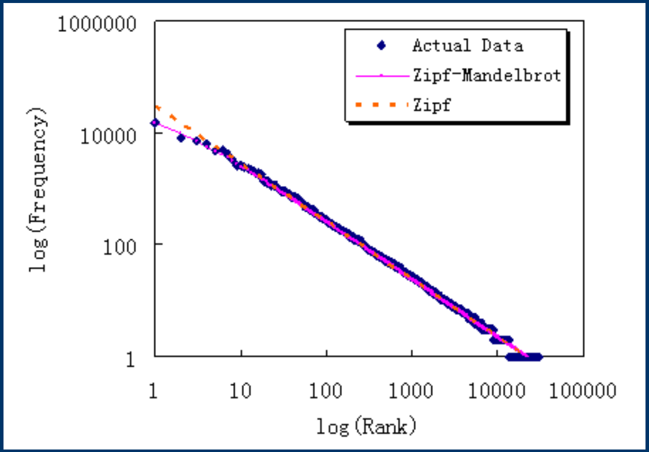
What is Natural Language Processing(NLP)?

Teaching Natural Language to Computers

- NLP : Teaching computers to understand language
    - Humans can understand natural language easily
    - However, computers are not…
- Major Difficulties in NLP
    - Vague objective : What is the *Semantic* of natural language?
    - Data sparsity : Zipf's Law

What is Natural Language Processing(NLP)?

Zipf’s Law : Sparsity of the Natural Language



The distribution of word frequencies in the novel "Ulysses”

→ Most words are very ‘Rare’!!

Meaning Representation(MR) of Natural Language

Two Major Paradigms

- Why MR?
    - A computable, standardized representation of natural language
    - "Believed" to carry semantics of natural language
- Two Major Approaches
    - Symbolic Representation Approach
    - Vector Representation Approach

MR of Natural Language: The Symbolic Approach

Symbolic Approach

- Symbolic MR
  - Adopting meta-language for Semantic Representation
  - Usually First-Order Logic, Lambda-Calculus Style

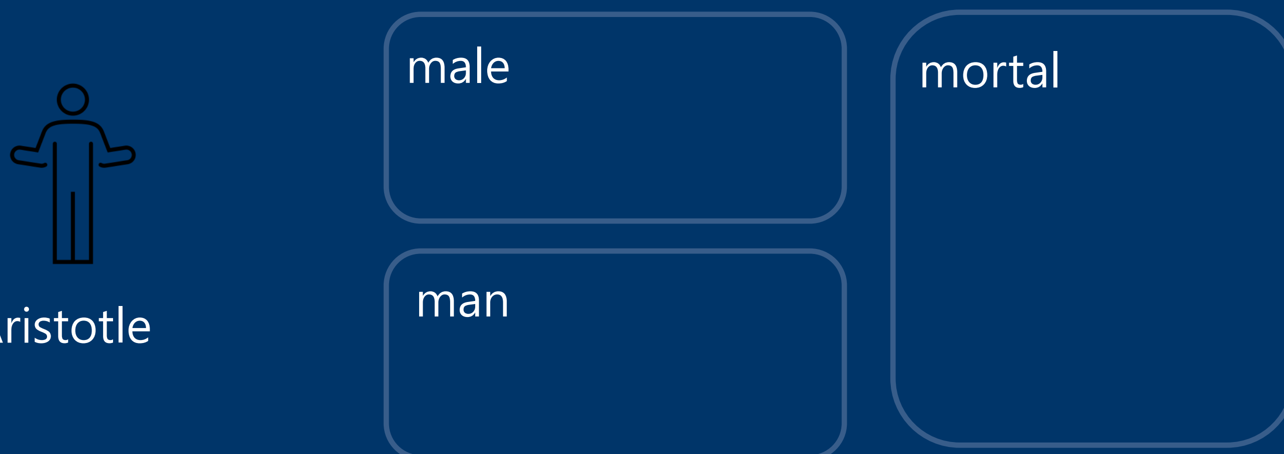
  ex) Aristotle is a human → $Aristotle \subseteq human$
- Pros & Cons
  - Pros : Human-readable
  - Cons : Difficulty in sparsity handling / verification

Cons of Symbolic MR

Why does sparsity matter?

- Consider following three sentences;

1. Aristotle is a male.
2. Aristotle is a man.
3. Man is mortal.



Cons of Symbolic MR

Why does sparsity matter?

Aristotle is a male.

male

Aristotle

$$Aristotle \subseteq male$$

Cons of Symbolic MR

Why does sparsity matter?

Aristotle is a male.

male

Aristotle

$$Aristotle \subseteq male$$

Aristotle is a man.

man

Aristotle

$$Aristotle \subseteq man$$

Cons of Symbolic MR

Why does sparsity matter?

Aristotle is a **male**.

Aristotle is a **man**.

male

Aristotle

$Aristotle \subseteq \boldsymbol{male}$

man

Aristotle

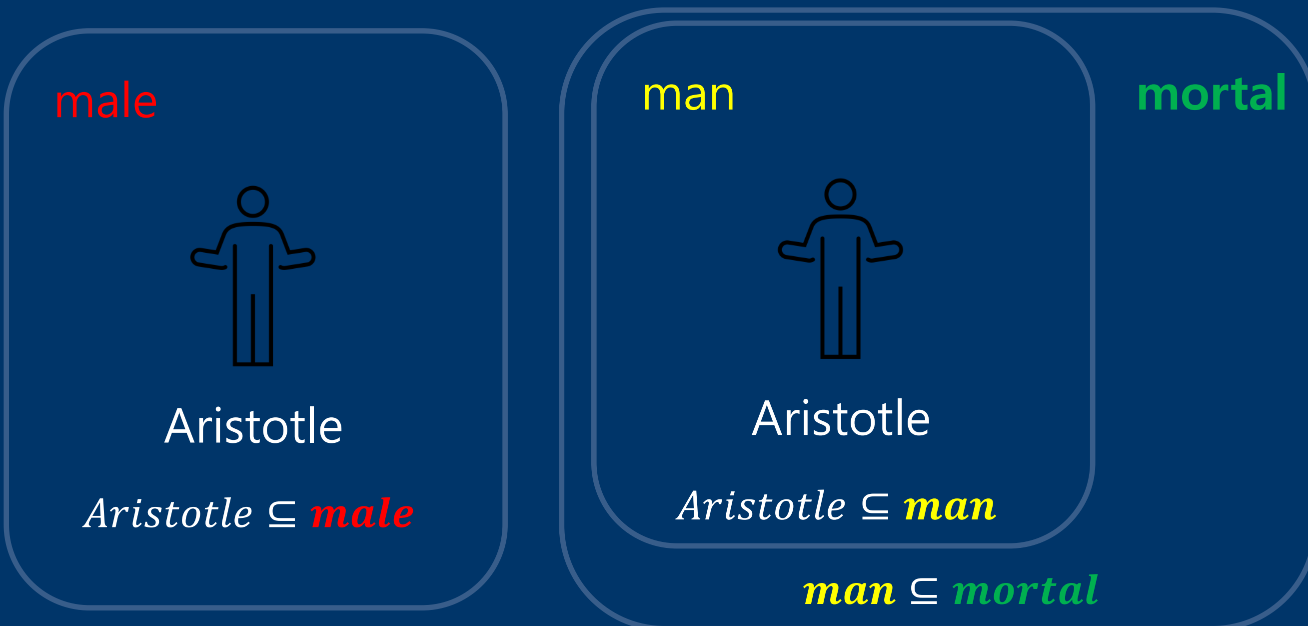
$Aristotle \subseteq \boldsymbol{man}$

Cons of Symbolic MR

Why does sparsity matter?

Aristotle is a **male**.

Aristotle is a **man**.
Man is **mortal**.



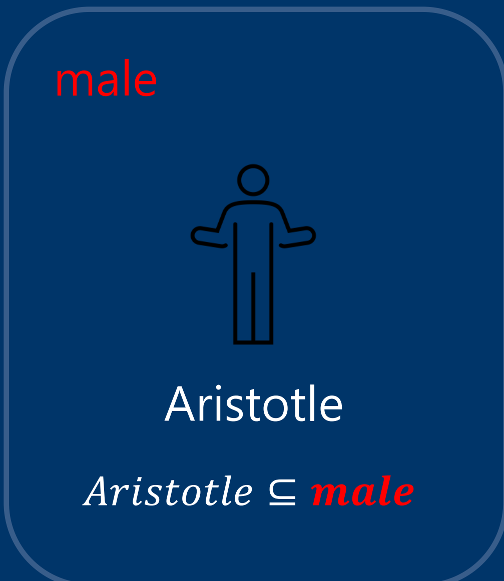
Cons of Symbolic MR

Why does sparsity matter?

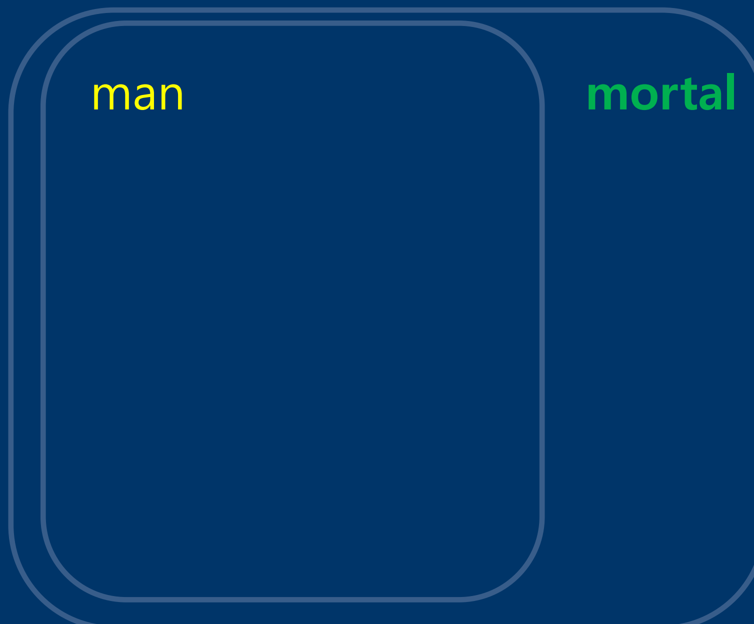
Cons of Symbolic MR

Why does sparsity matter?

Aristotle is a **male**.



Man is **mortal**.



MR of Natural Language: The Deep-Learning Approach

Deep Learning Approach

- Vector as MR
    - Extract abstract features automatically from text data
- Vector of Features
    - Example: Big Five Personality Trait Test
    - Express personality as 5-dimension vectors

Openness to experience	79 out of 100
Agreeableness	75 out of 100
Conscientiousness	42 out of 100
Negative emotionality	50 out of 100
Extraversion	58 out of 100

MR of Natural Language: The Deep-Learning Approach

Very Quick Introduction to word2vec

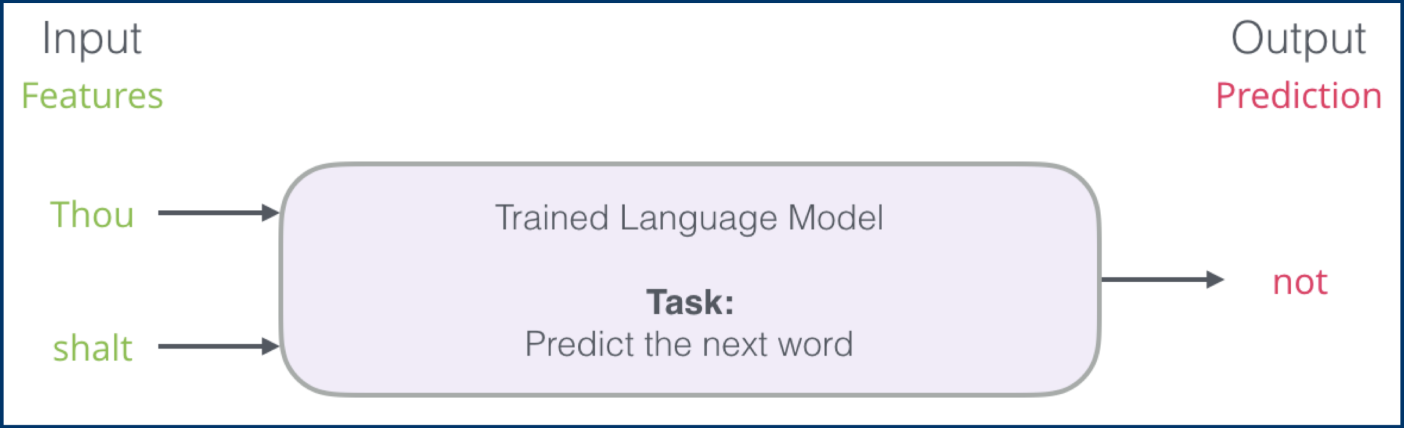
- How do we represent a word as a vector?
    - *"You shall know a word by the company it keeps"* - J.R.Firth
- Word2vec Scheme
    - Express word as a vector
    - Train the vector to contain semantics of each word

MR of Natural Language: The Deep-Learning Approach

Very Quick Introduction to word2vec

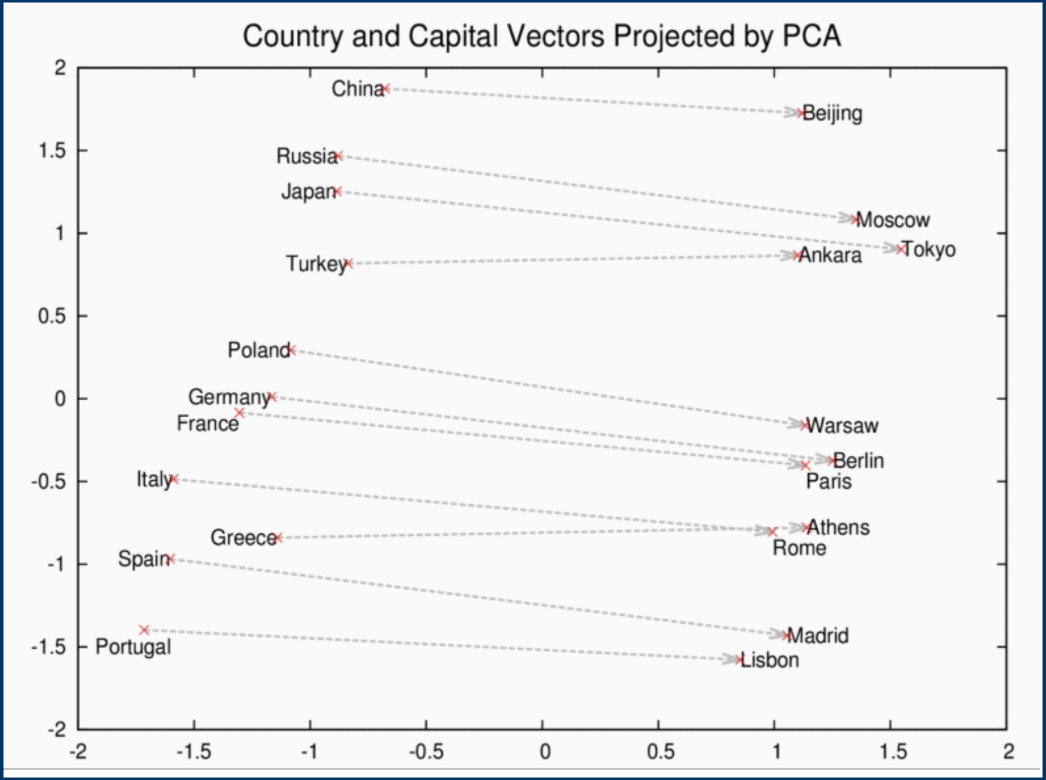
MR of Natural Language: The Deep-Learning Approach

Very Quick Introduction to word2vec



MR of Natural Language: The Deep-Learning Approach

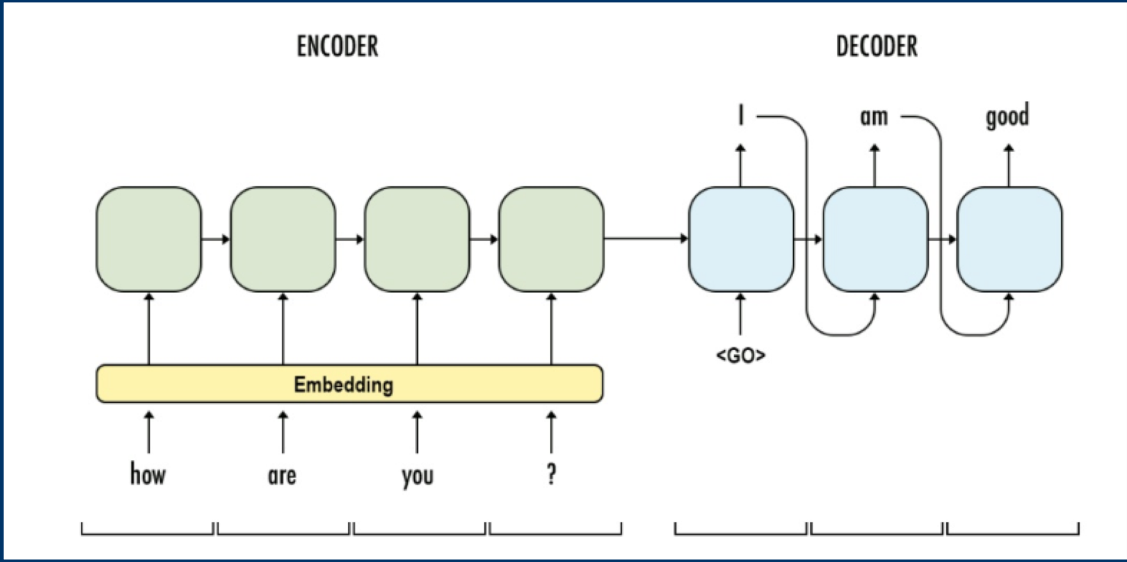
Very Quick Introduction to word2vec



MR of Natural Language: The Deep-Learning Approach

Deep Learning Approach

- Word → Vector. How about sentences?
  - Words → sequence of vectors → sentence vector!

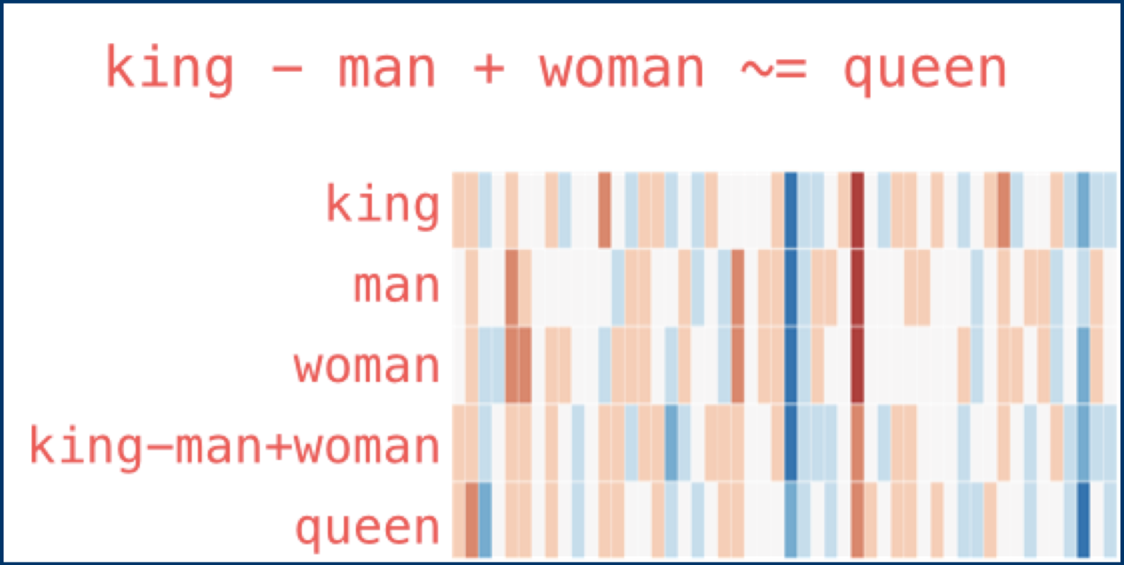


Pros of vector-based MR

The Chinese Room Nowadays….

Cons of vector-based MR

So “what” Actually Happens in the Room?



Why Symbolic + Distributional?

Benefits of Augmenting Distributional Semantics

Distributional Semantic Augmented Formal Semantic

- **Similarity between Semantics**
- Probabilistic Deduction : Deduction to Classification

ex) $P(Aristotle \subseteq male \ \rightarrow Aristotle \subseteq man)$

$= P(mankind \subseteq human)$

- $mankind \subseteq human$: classification problem
  - Solvable by embedding
  - Interpretable reasoning

Symbolic MR of this Paper: sent2dl

Sentence → Description Logic

Symbolic MR of this Paper: sent2dl

Quick Introduction to DL

Description Logic Formula :

$$S \sqsubseteq \forall V.O \equiv S \subseteq \{x \mid \exists y, V(x,y) \land y \in O\}$$

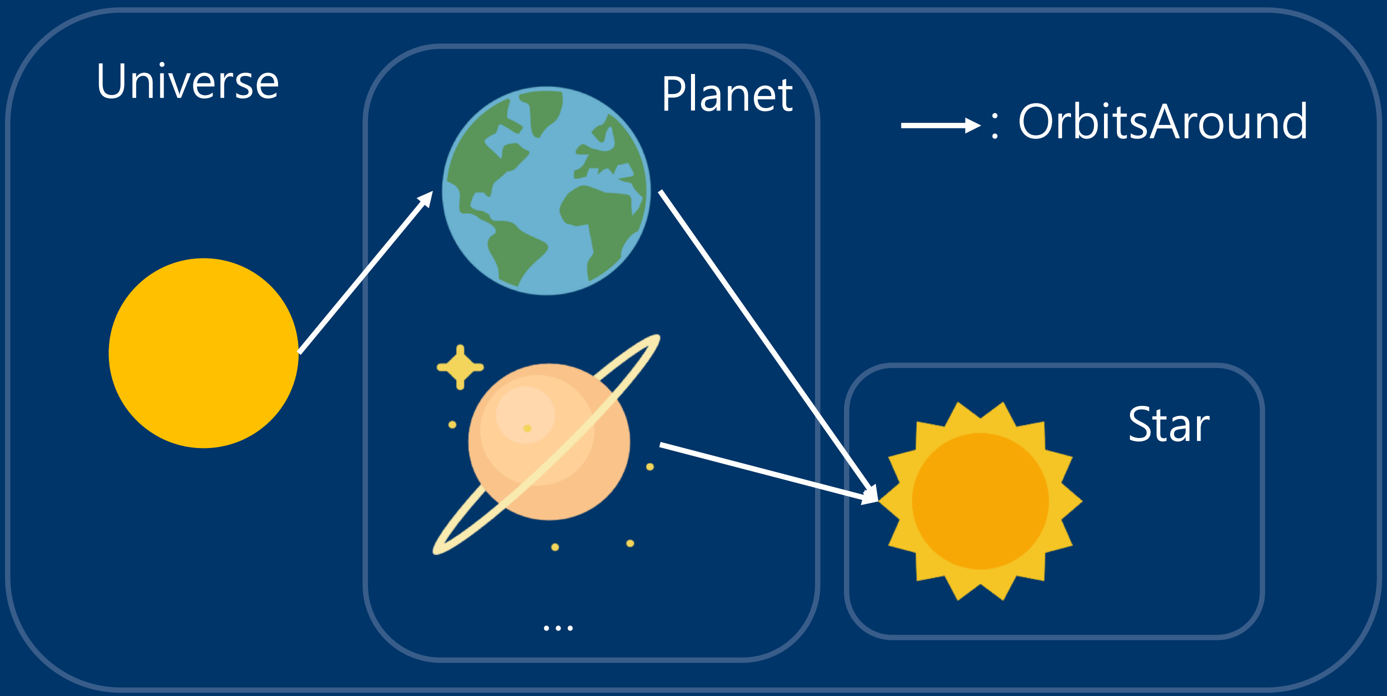
- S, O : Set
- V : Relationship (a binary predicate)

Ex)

$$\{Moon\} \sqsubseteq \forall OrbitsAround.(\neg Star)$$
$$Planet \sqsubseteq \forall OrbitsAround.Star$$

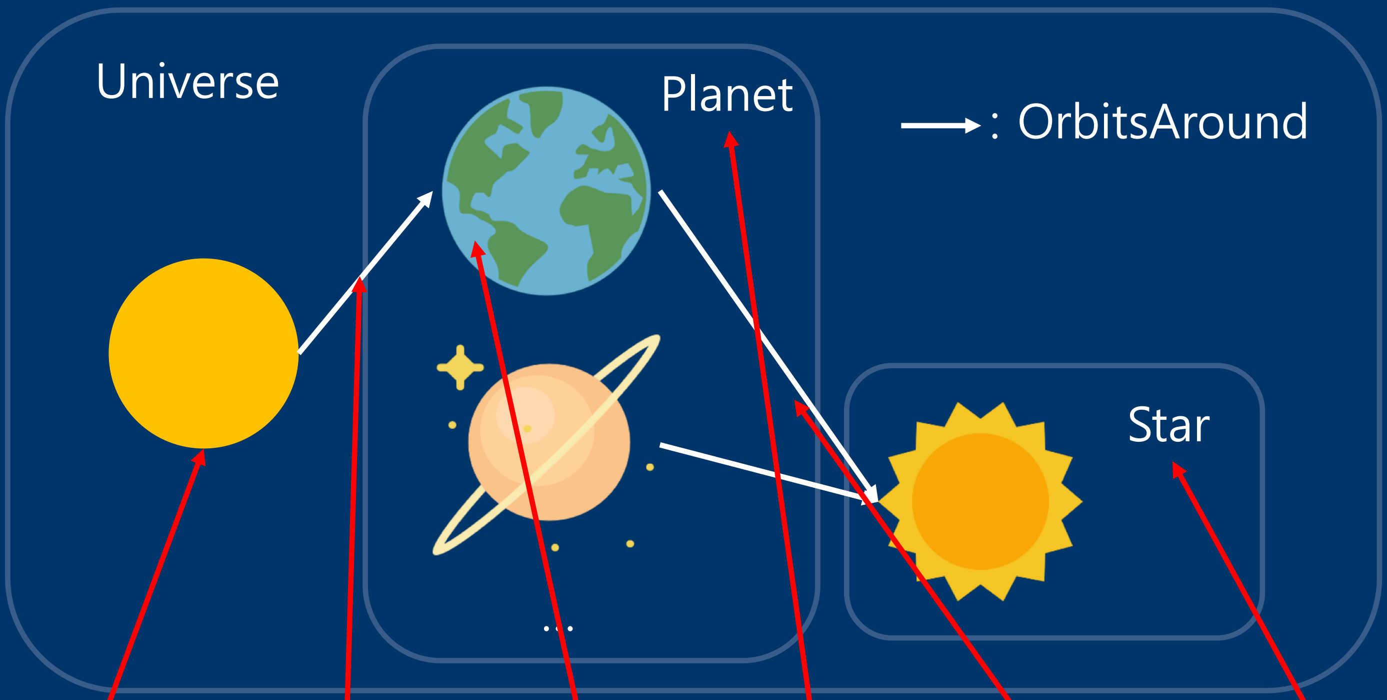
Symbolic MR of this Paper: sent2dl

Sentence → Description Logic



Symbolic MR of this Paper: sent2dl

Quick Introduction to DL : Solar System Example



$\{Moon\} \sqsubseteq \forall OrbitsAround.(\neg Star)$ $\quad Planet \sqsubseteq \forall OrbitsAround.Star$

Symbolic MR of this Paper: sent2dl

Sentence → Description Logic

Recursive / Iterative Composition

- Make parse tree of a given sentence
- Preprocessing : 안긴 문장 분할, 주어 앞 부사구 삭제 등
- Re-parse processed sentence segments
- Recursively get formula corresponding to each constituent
- Compose subtree-formulae according to the rules

Recursively transform SOV-sentence to DL Formula

$$sent2dl(SOV - sentence)$$
$$= \ sent2dl(S) \sqsubseteq \forall sent2dl(V).sent2dl(O)$$

Symbolic MR of this Paper: sent2dl

Sentence → Description Logic Example

ex) 예시 문장 변환
kt는 수원에서 열린 NC와의 시즌 14차전 경기에서 10-0으로 대승을 일궈냈다.

- 문장 Preprocessing
    - kt는 경기에서 10-0으로 대승을 일궈냈다.
    - NC와의 시즌 14차전 경기가 수원에서 열렸다.

Symbolic MR of this Paper: sent2dl

Sentence → Description Logic Example

ex) 예시 문장 변환
kt는 수원에서 열린 NC와의 시즌 14차전 경기에서 10-0으로 대승을 일궈냈다.

- 문장 Re-parse
    - (VP

        (NP-SBJ kt는)

        (NP-AJT 경기에서)

        (NP-AJT 10-0으로)

        (NP-OBJ 대승을)

        (VP 일궈냈다.))

Symbolic MR of this Paper: sent2dl

Sentence → Description Logic Example

ex) 예시 문장 변환

kt는 수원에서 열린 NC와의 시즌 14차전 경기에서 10-0으로 대승을 일궈냈다.

- 문장 형태에 맞는 rule 적용

    ex) (VP (NP-SBJ) (NP-AJT*) (NP-OBJ) (VP)) 의 경우,

    $f(NP-SBJ) \sqsubseteq \forall f(NP-AJT) \circ f(VP).f(NP-OBJ)$

    의 규칙 적용.

- 재귀적으로 depth가 2인 트리가 될 때까지 적용, 그 후에는 base-rule 적용

    ex) (NP (kt NNP) (는 JX)) → $\{kt\}$

- 그 후 병합 (Principle of Compositionality)

Augmenting Distributional Semantics to Logical Formula

Giving Similarity Between Logic Formulae

Given two DL Formulae

$$S_1 \sqsubseteq \forall V_1 . O_1$$
$$S_2 \sqsubseteq \forall V_2 . O_2$$

Distance between formulae is defined as

$$sim(S_1, S_2) + sim(V_1, V_2) + sim(O_1, O_2)$$

where

$$sim(a, b) = w2v(a, b), \qquad a, b \in Vocab$$
$$= \min_{\{\pi(a), \pi(b)\}} \Big[\sum_{i \in \pi(a), j \in \pi(b)} w2v(i, j) \Big], \qquad a, b \text{ is } Set$$

Experimental Verification

On unsupervised sentence clustering

Dataset : Baseball match recap data

Benchmark

- Benchmark 1 : Multitask-Learning
    - Sentence Representation Learning Task via Quick Thought
    - Seq2Seq based Neural Machine Translation Task
- Benchmark 2: Sentence-Bert
    - Sentence-transformer의 multilingual 옵션 적용

Metric : AMI , NMI

Experimental Verification

	Multitask Learning	Sentence-Bert	Ours
AMI	0.12 / 0.47	0.12	0.65
NMI	0.11 / 0.48	0.10	0.65

Experimental Verification

Example Results

Benchmark(Multitask-Learning)	ours(sent2dl)
- 두산은 4월 15일 고척 스카이돔에서 열린 키움과의 원정 경기에서 3-2로 승리했다. - LG는 19일 고척에서 열린 키움과의 원정 경기에서 8-3으로 이겼다. - NC는 시즌 순위 10위를 유지했다. - 한화는 5월 30일 대전 한화생명 이글스 파크에서 열린 NC와 7차전 홈 경기에서 10-4로 이겼다,	- 키움 임병욱이 1안타 1타점으로 팀 승리에 기여했다. - 7번 지명타자로 출전한 김동엽이 2타수 1득점의 활약했다. - 데뷔전을 치룬 이도윤은 1타수 무안타에 그쳤다. - 데뷔전을 치룬 이도윤은 1타수 무안타로 물러났다. - 박기혁은 4타수 1안타 1타점으로 팀 승리에 기여했다.

Conclusion

- Things done
  - Tried to Merge vector-based MR and symbolic MR
    - Made Description-Logic based MR
    - Only SYNTAX, no interpretation/formal semantic
  - Indirectly verified via unsupervised clustering task
- Further TODOs
  - Assign formal semantics to proposed language

END OF DOCUMENT