Data Mining
Quality Analytics

Open DMQA Seminar

Knowledge Enhanced NLP

2020. 7. 31

Da Bin Min
Data Mining & Quality Analytics Lab.

Contents

I. Introduction

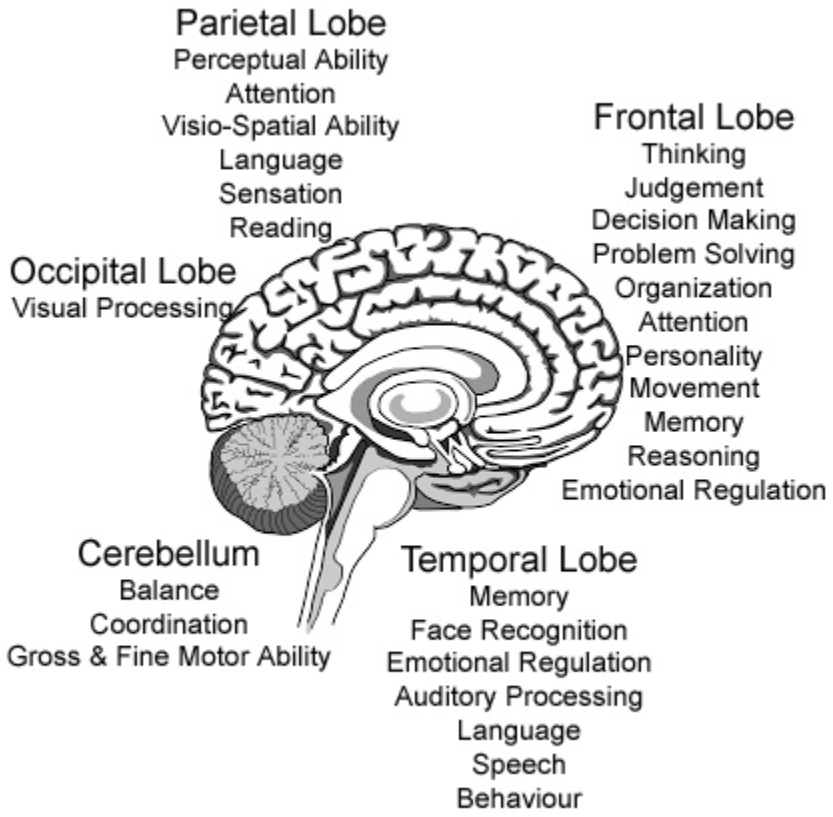
- 언어란 무엇인가?

I. Introduction

- 언어란 무엇인가?
  → 인간 사고 결과물의 표현

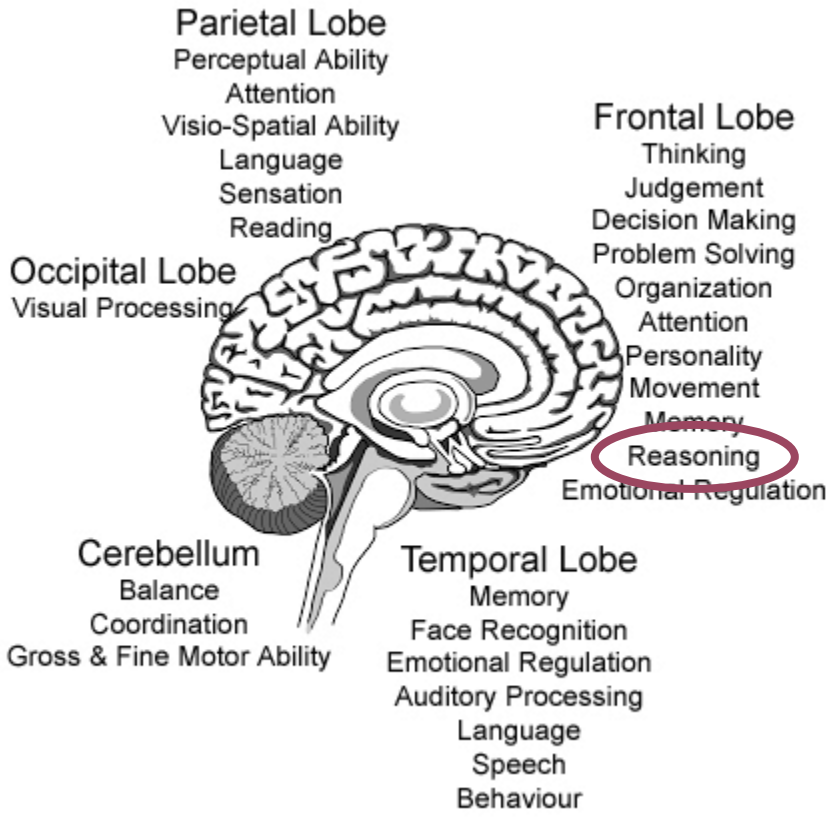
I. Introduction

- 인간의 사고는 다양한 뇌 인지기능들의 복합적인 작용을 통해 수행
  → 인간의 언어는 매우 복잡한 의미를 갖는다



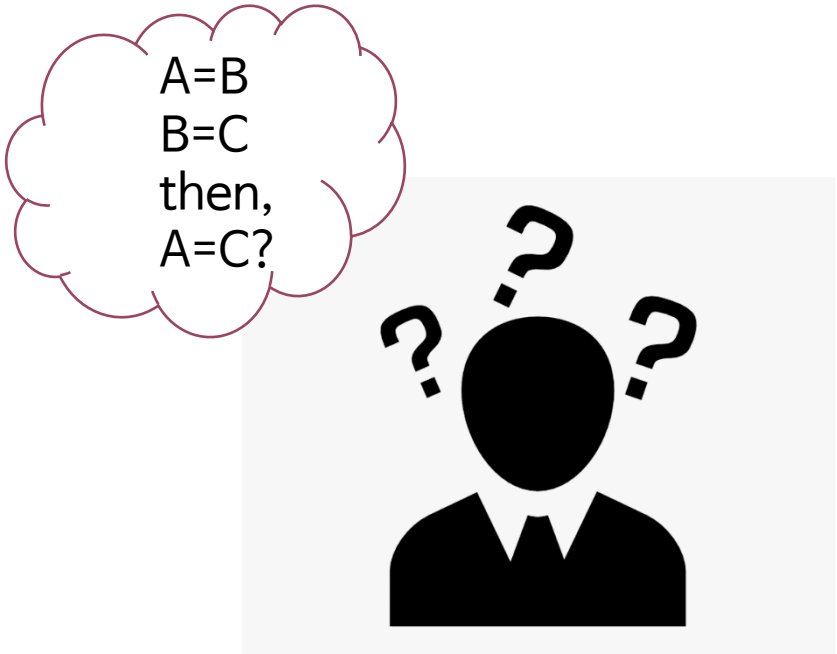
I. Introduction

- 인간의 사고는 다양한 뇌 인지기능들의 복합적인 작용을 통해 수행
  → 인간의 언어는 매우 복잡한 의미를 갖는다



I. Introduction

- **Reasoning(추론)**
  - 이미 알고있거나 확인된 정보로부터 논리적 결론을 도출하는 기능
  - 뇌의 추론기능을 담당하는 부분의 손상은 언어 및 의사소통 능력의 장애 유발(Benton E at al., 1996)



Benton E, Bryan K. Right cerebral hemisphere damage: incidence of language problems. International Journal of Rehabilitation research. Internationale Zeitschrift fur Rehabilitationsforschung. Revue Internationale de Recherches de Readaptation. 1996 Mar;19(1):47-54.

추론 능력은

언어의 논리적 이해와 표현에

매우 중요

I. Introduction

Bob Dylan wrote Blowin' in the Wind

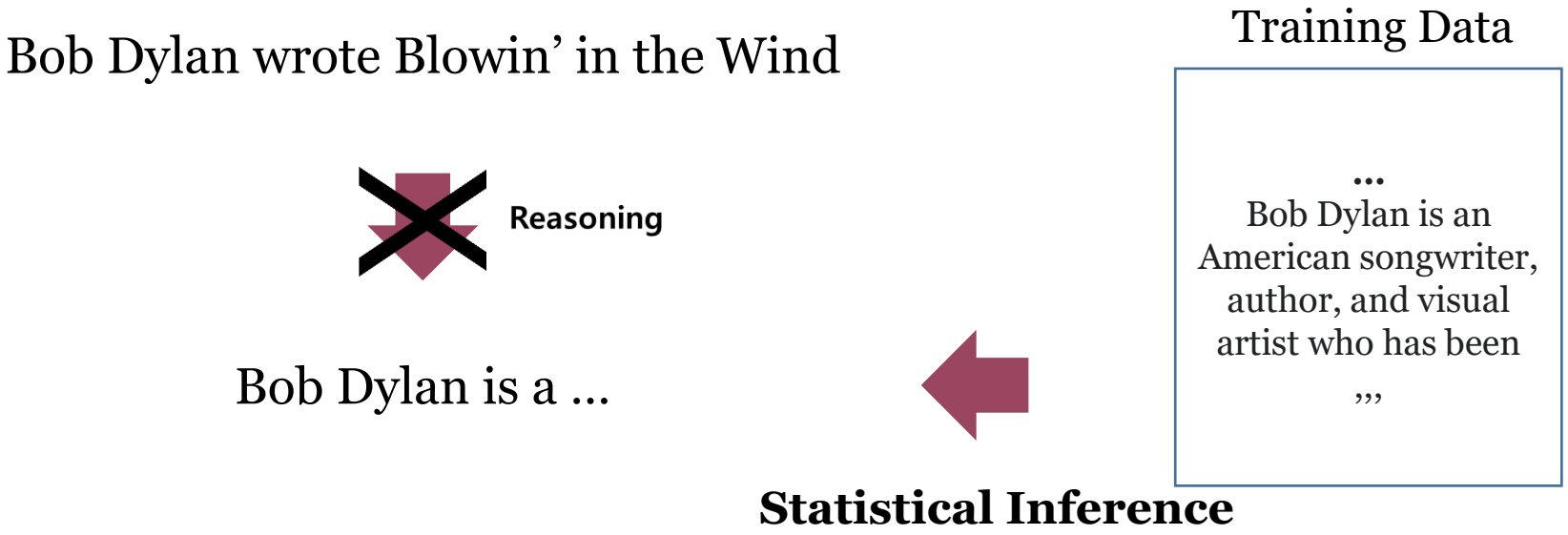
Reasoning

- 노래를 쓰는 사람은 작곡가다.
- Bob Dylan은 노래를 썼다.

Bob Dylan is a songwriter

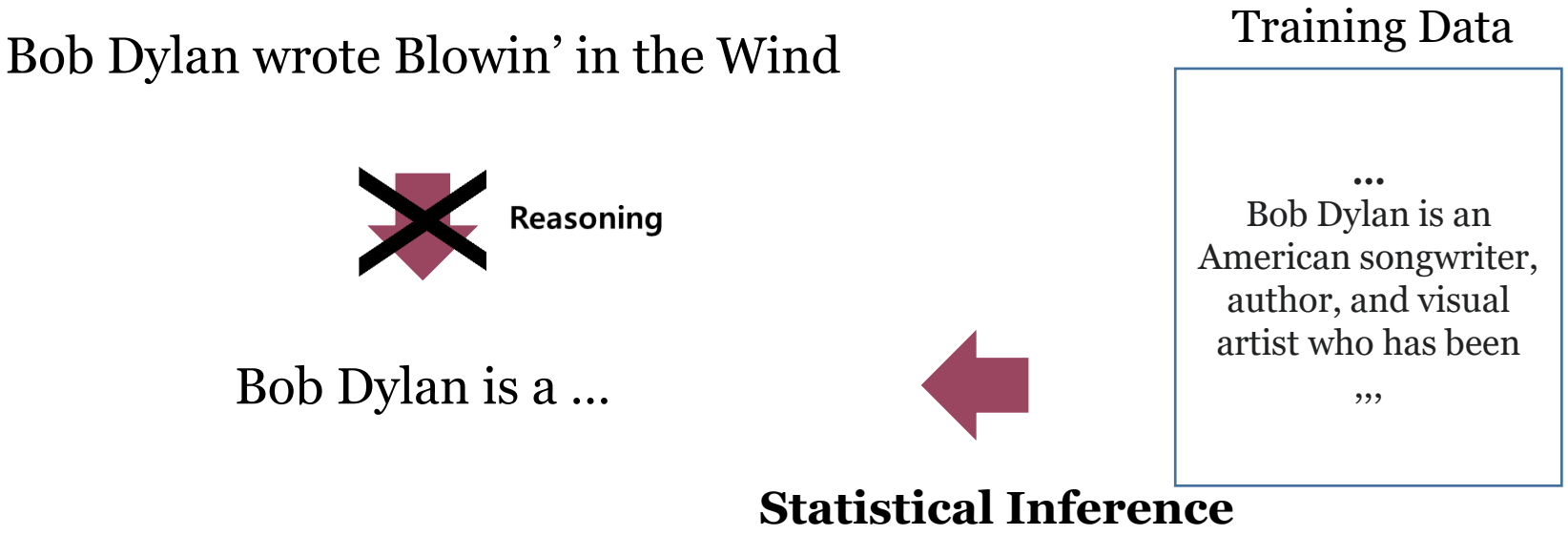
I. Introduction

- 일반적인 NLP 모델들의 한계점
  - 단어들 간의 Co-occurrence에 기반하여 단어와 문장의 의미를 이해
  - 현실 세계의 지식에 기반한 논리적 이해와 추론이 불가능
  - 학습데이터에서 관측되지 않는 수 많은 지식들을 언어 이해에 활용하지 못함



I. Introduction

- 일반적인 NLP 모델들의 한계점
  - 단어들 간의 Co-occurrence에 기반하여 단어와 문장의 의미를 이해
  - 현실 세계의 지식에 기반한 논리적 이해와 추론이 불가능
  - 학습데이터에서 관측되지 않는 수 많은 지식들을 언어 이해에 활용하지 못함

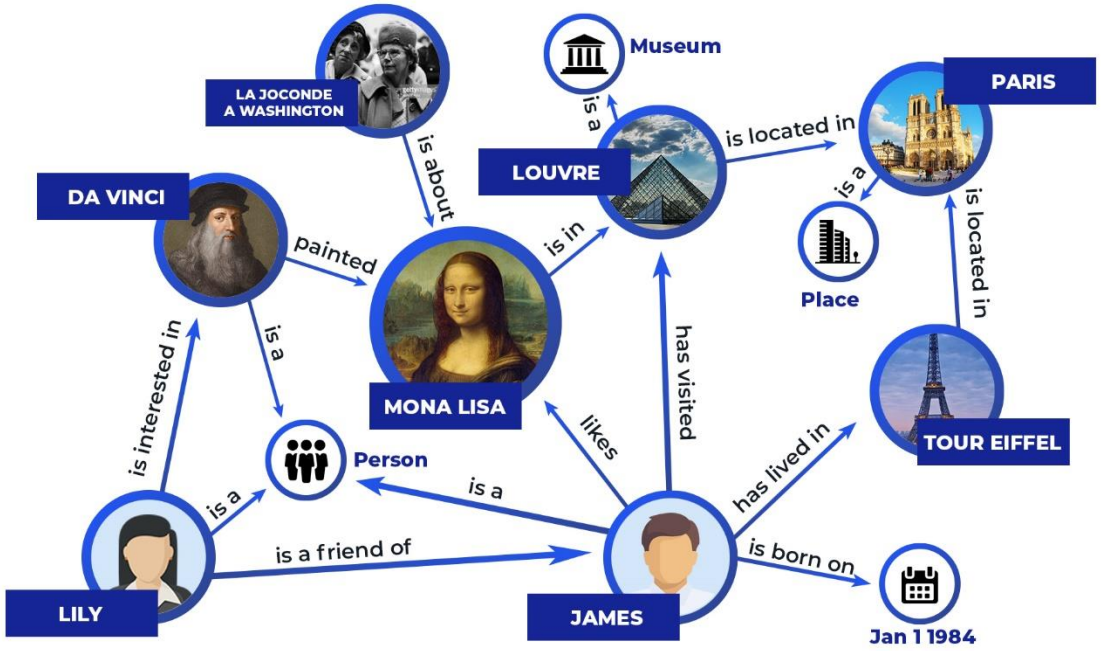


언어 이해를 위한 **인간의 추론 방식**을 모사하자

II. Knowledge Base

- Knowledge Base
  - 현실의 지식을 저장한 대규모 데이터베이스
  - Wikidata, DBPedia
- Knowledge Graph의 형식으로 지식 저장

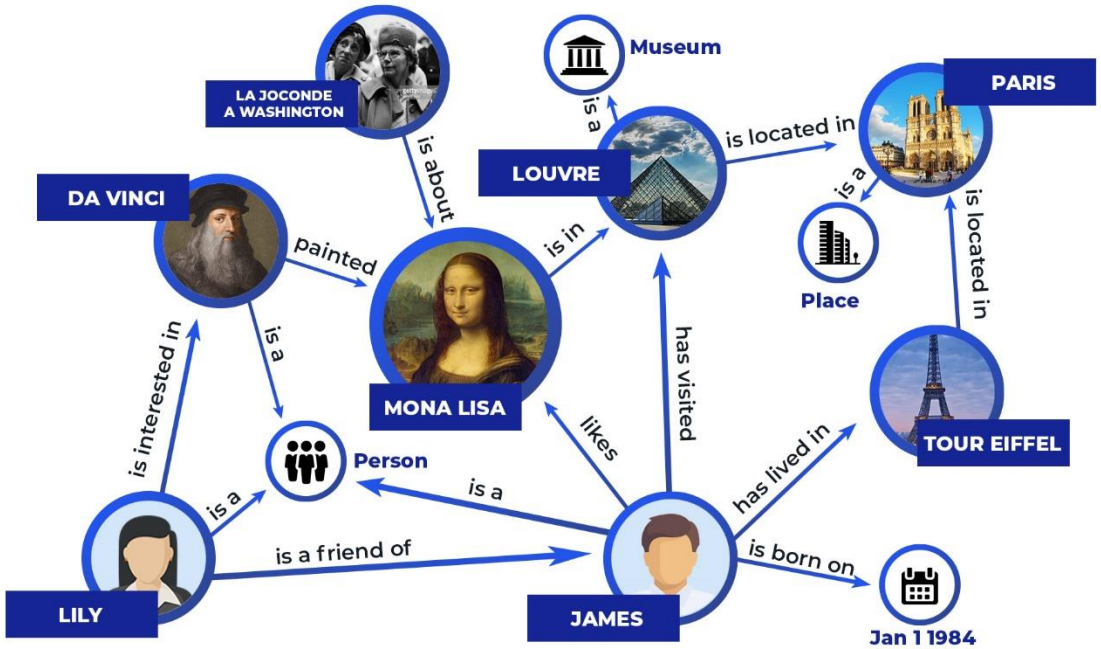
Knowledge Base



Knowledge Graph

II. Knowledge Base

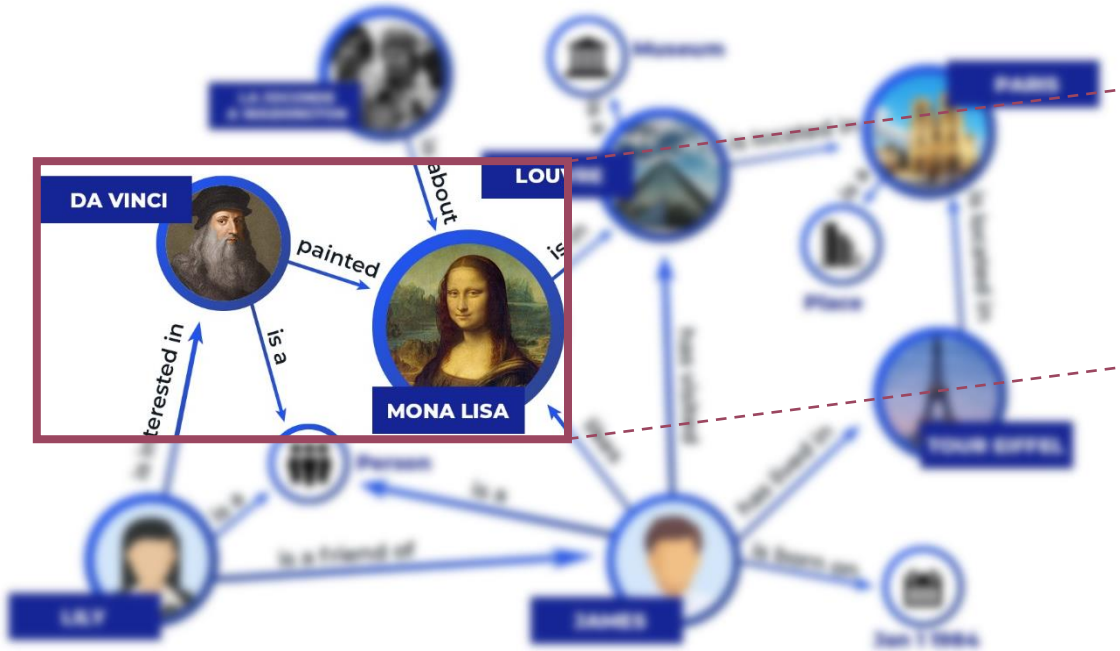
- Knowledge Graph
  - 객체(Entity)들 간의 관계(Relation)가 표현된 유향그래프
  - Node: Entity(고유명사, 연도, 대표속성 등) / Edge: Relation(관계)



Knowledge Graph

II. Knowledge Base

- Knowledge Graph
  - 객체(Entity)들 간의 관계(Relation)가 표현된 유향그래프
  - Node: Entity(고유명사, 연도, 대표속성 등) / Edge: Relation(관계)



(Da Vinci, painted, Mona Lisa)

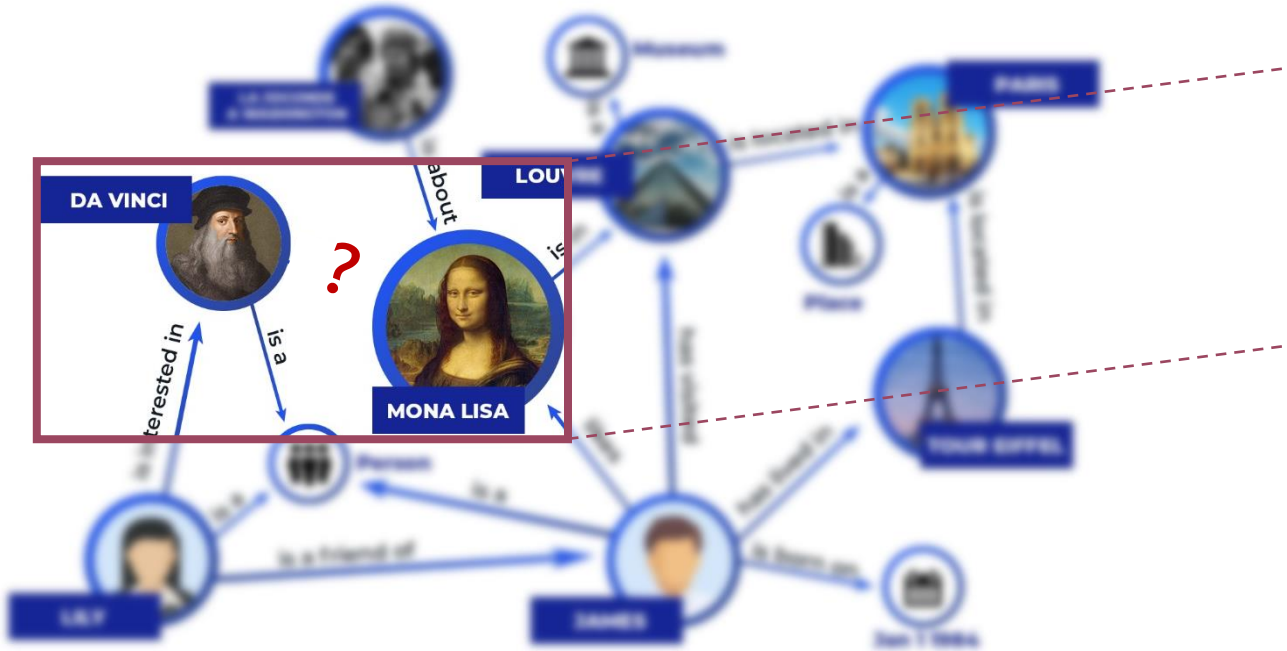
Head Entity (h) Relation (r) Tale Entity (t)

Knowledge → (h, r, t)

Knowledge Graph

II. Knowledge Base

- Knowledge Graph Embedding
  - Knowledge Graph Completion 문제 해결을 위한 방법
  - 모든 Entity와 Relation을 저차원 벡터로 표현, 벡터 연산을 통해 Graph Completion 수행
  - Knowledge Enhanced NLP 모델들에 직접적으로 사용됨



Knowledge Graph

(Da Vinci, ?, Mona Lisa)

KG Completion Problem

II. Knowledge Base

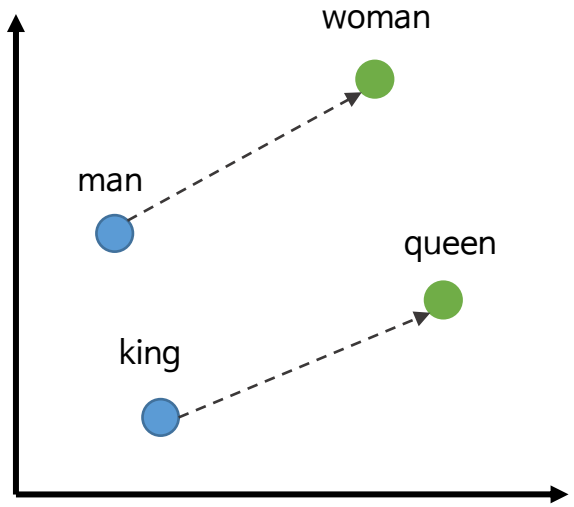
- TransE(2013)
  - 대표적인 Knowledge Graph Embedding 방법론
  - 참 (h, r, t)→ $h + r = t$ 가 되도록 유도
  - 거짓 (h, r, t) → $h + r \neq t$ 가 되도록 유도

INPUT (HEAD AND LABEL)	PREDICTED TAILS
Lil Wayne born in	**New Orleans**, Atlanta, Austin, St. Louis, Toronto, New York City, Wellington, Dallas, Puerto Rico
WALL-E has the genre	Animations, Computer Animation, *Comedy film*, *Adventure film*, *Science Fiction*, **Fantasy**, Stop motion, *Satire*, Drama

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multirelationaldata. In Proceedings of NIPS, pages 2787–2795.

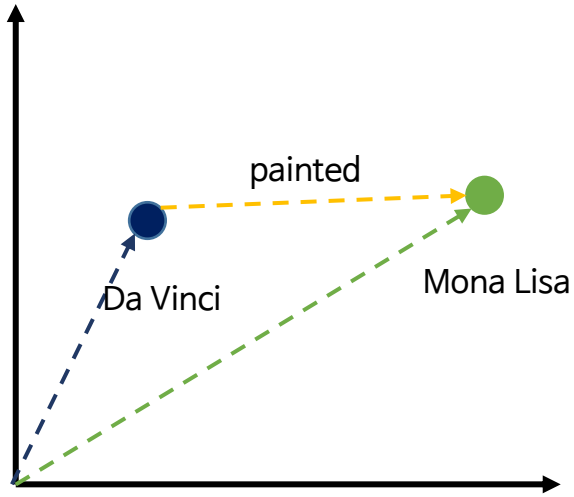
II. Knowledge Base

Word Embedding
(Word2Vec, Glove 등)



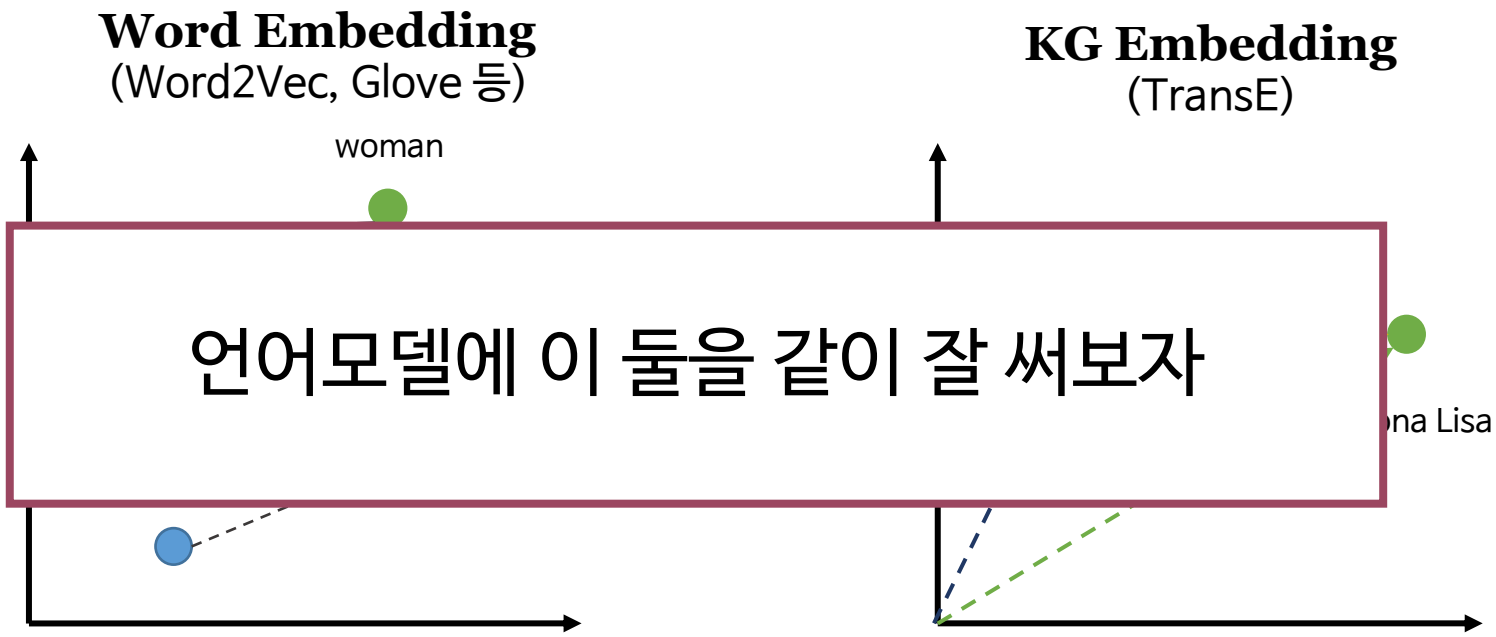
- 의미구조 유지
- Co-occurrence 기반

KG Embedding
(TransE)



- 논리구조 유지
- 지식 기반

III. Knowledge Enhanced NLP Methods

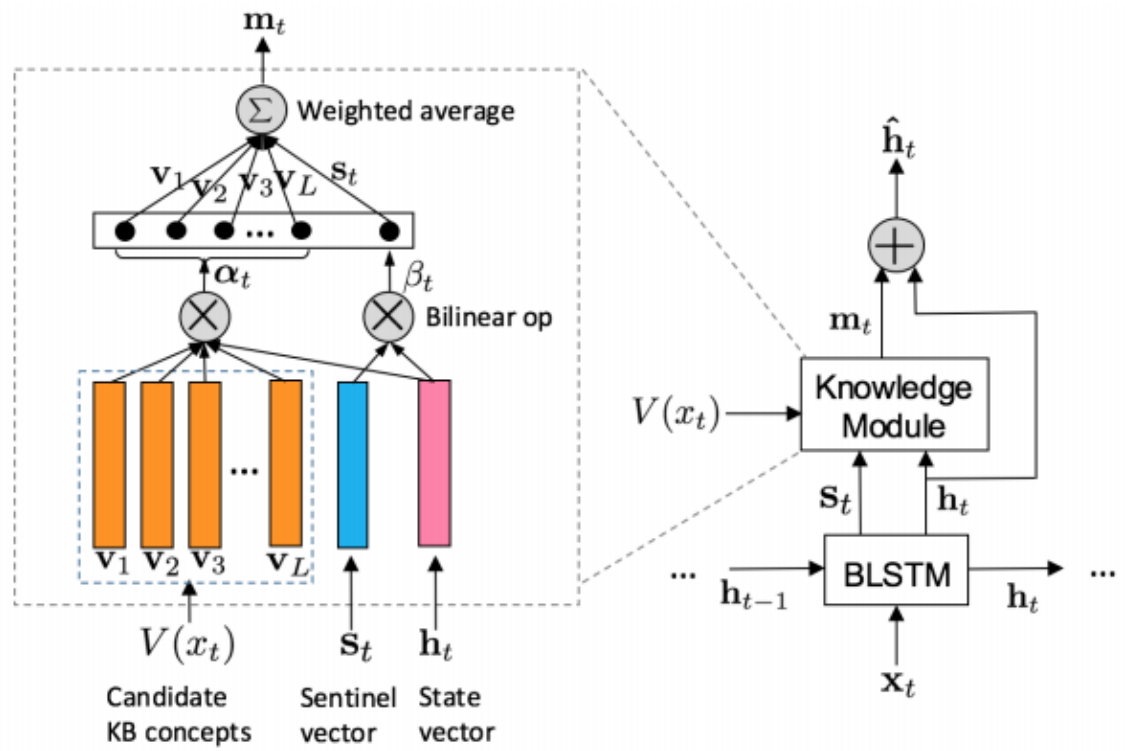


- 의미구조 유지
- Co-occurrence 기반

- 논리구조 유지
- 지식 기반

III. Knowledge Enhanced NLP Methods

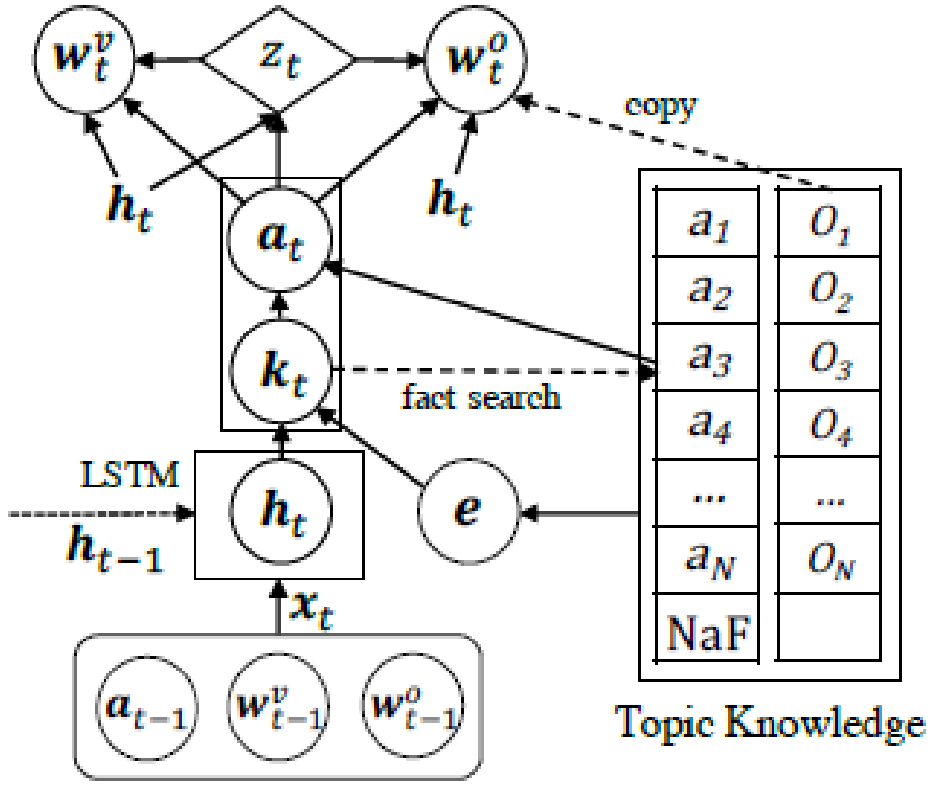
- KBLSTM(2017)
  - LSTM 기반 모델
  - 지식 (h, r, t)에 대한 임베딩 v는 h, r, t 임베딩을 이용해 생성(e.g. concat)
  - 현재 스텝의 hidden state와 관련 있는 지식을 선택하여 사용
  - 관련 있는 지식이 없을 경우 s_t가 선택됨



Bishan Yang, Tom Mitchell. 2017. Leveraging Knowledge Bases in LSTMs for Improving Machine Reading. ACL.

III. Knowledge Enhanced NLP Methods

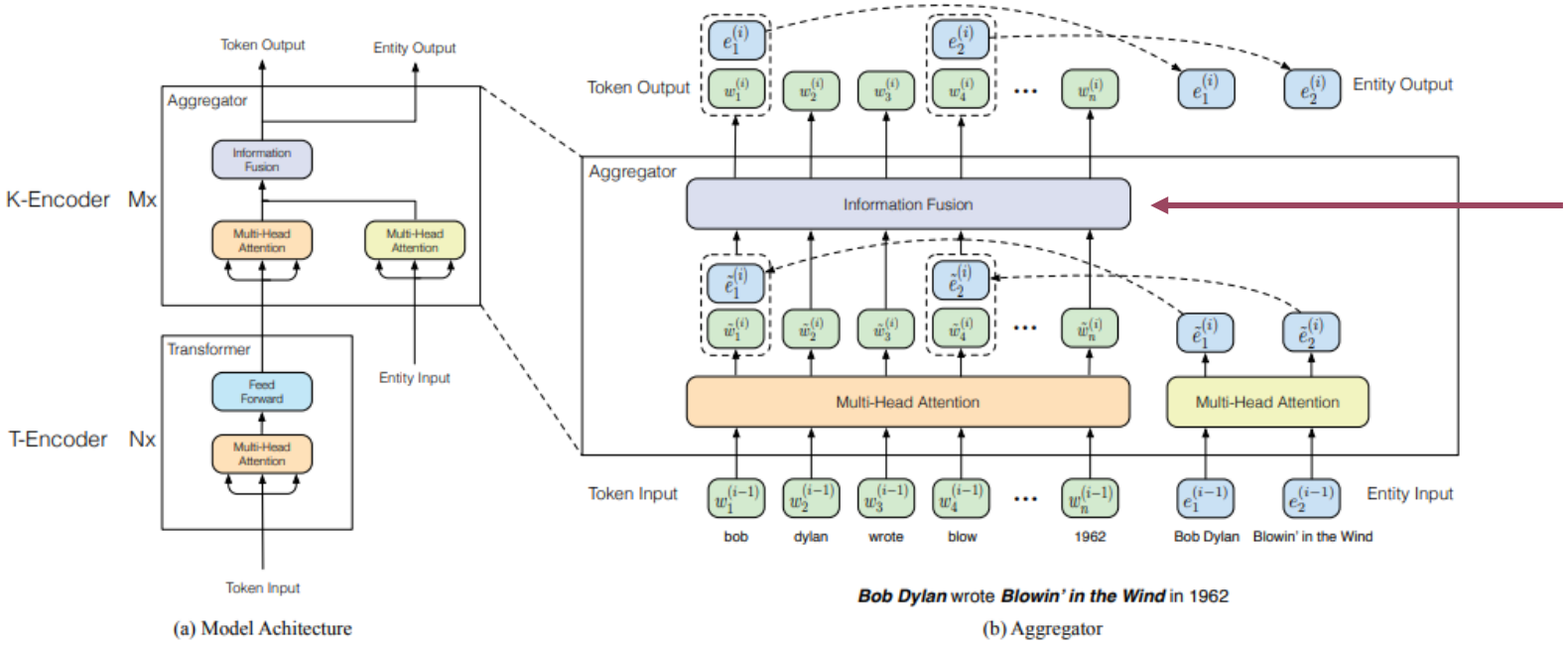
- NKLM(2017)
  - LSTM 기반 모델, KBLSTM과 유사한 구조
  - 다음 스텝의 LSTM cell에 현재 스텝 까지의 문맥적 의미와 지식을 함께 넘겨줌
  - 이전스텝 까지의 문맥적 의미와 지식을 이용해 현재 관련있는 지식을 찾아 사용



Sungjin Ahn, Heeyoul Choi, Tanel Parnamaa, and Yoshua Bengio. 2017. A neural knowledge language model. arXiv:1608.00318.

III. Knowledge Enhanced NLP Methods

- EARNIE(2019)
  - BERT 기반 모델
  - 각 단어의 문맥적 의미에 KG의 지식을 주입하고자 함



$$h_i = \sigma\left(W_A \tilde{w}_i + W_B \tilde{e}_j + b\right)$$
$$w_i = \sigma(W_C h_i + b)$$
$$e_i = \sigma(W_D h_i + b)$$

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. ACL.

III. Knowledge Enhanced NLP Methods

- Advantages
  - Question Answering, Document Summarization, Entity Typing, Relation Classification
  - 고유명사가 포함된 문장의 이해와 생성 능력 탁월

Entity Typing Task

Sentence with Target Entity	Entity Types
During the Inca Empire, {the Inti Raymi} was the most important of four ceremonies celebrated in Cusco.	event, festival, **ritual, custom, ceremony, party, celebration**
{They} have been asked to appear in court to face the charge.	person, **accused, suspect, defendant**
Ban praised Rwanda's commitment to the UN and its role in {peacemaking operations}.	event, **plan, mission, action**

Relation Classification Task

Supporting Set	
(A) capital_of	(1) *London* is the capital of *the U.K.* (2) *Washington* is the capital of *the U.S.A.*
(B) member_of	(1) *Newton* served as the president of *the Royal Society*. (2) *Leibniz* was a member of *the Prussian Academy of Sciences*.
(C) birth_name	(1) *Samuel Langhorne Clemens*, better known by his pen name *Mark Twain*, was an American writer. (2) *Alexei Maximovich Peshkov*, primarily known as *Maxim Gorky*, was a Russian and Soviet writer.
Test Instance	
(A) or (B) or (C)	*Euler* was elected a foreign member of *the Royal Swedish Academy of Sciences*.

Eunsol Choi, Omer Levy, Yejin Choi, Luke Zettlemoyer . 2018. Ultra Fine Entity Typing. ACL
Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, Maosong Sun. 2018. Fewrel : A large scale supervised few shot relation classification dataset with state of the art evaluation. EMNLP

III. Knowledge Enhanced NLP Methods

- Advantages
  - Question Answering, Document Summarization, Entity Typing, Relation Classification
  - 고유명사가 포함된 문장의 이해와 생성 능력 탁월

Entity Typing Task

Model	P	R	F1
NFGEC (LSTM)	68.80	53.30	60.10
UFET	77.40	60.60	68.00
BERT	76.37	70.96	73.56
ERNIE	**78.42**	**72.90**	**75.56**

Relation Classification Task

Model	FewRel			TACRED		
	P	R	F1	P	R	F1
CNN	69.51	69.64	69.35	70.30	54.20	61.20
PA-LSTM	-	-	-	65.70	64.50	65.10
C-GCN	-	-	-	69.90	63.30	66.40
BERT	85.05	85.11	84.89	67.23	64.81	66.00
ERNIE	88.49	88.44	**88.32**	69.97	66.08	**67.97**

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. ACL.

III. Knowledge Enhanced NLP Methods

- Advantages
  - Question Answering, Document Summarization, Entity Typing, Relation Classification
  - 고유명사가 포함된 문장의 이해와 생성 능력 탁월

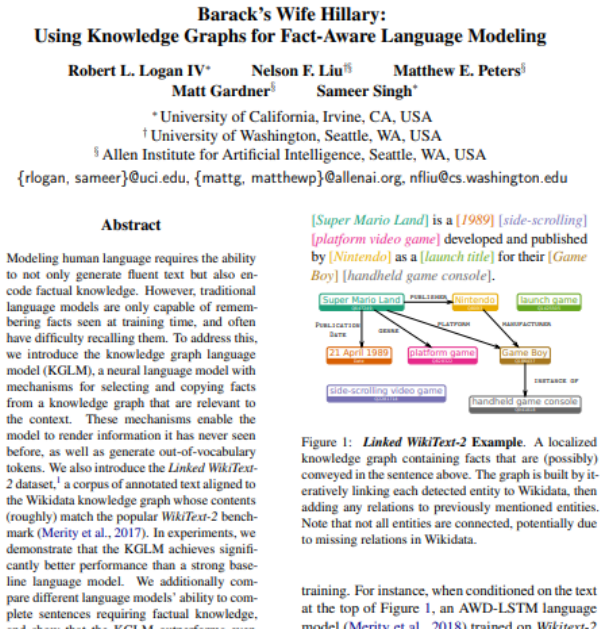
A Neural Knowledge Language Model

Warm-up	Louise Allbritton (3 july <unk>february 1979) was
RNNLM	a <unk><unk>who was born in <unk>, <unk>, <unk>, <unk>, <unk>, <unk>, <unk>
NKLM	an english [Actor]. he was born in [Oklahoma] , and died in [Oklahoma]. he was married to [Charles] [Collingwood]
Warm-up	Issa Serge Coelo (born 1967) is a <unk>
RNNLM	actor . he is best known for his role as <unk><unk>in the television series <unk>. he also
NKLM	[Film] director . he is best known for his role as the <unk><unk>in the film [Un] [taxi] [pour] [Aouzou]
Warm-up	Adam wade Gontier is a canadian Musician and Songwriter .
RNNLM	she is best known for her role as <unk><unk>on the television series <unk>. she has also appeared
NKLM	he is best known for his work with the band [Three] [Days] [Grace] . he is the founder of the
Warm-up	Rory Calhoun (august 8 , 1922 april 28
RNNLM	, 2010) was a <unk>actress . she was born in <unk>, <unk>, <unk>. she was
NKLM	, 2008) was an american [Actor] . he was born in [Los] [Angeles] california . he was born in

Sungjin Ahn, Heeyoul Choi, Tanel Parnamaa, and Yoshua Bengio. 2017. A neural knowledge language model. arXiv:1608.00318.

III. Knowledge Enhanced NLP Methods

- 기타 최신 연구
  - Robert Logan, et al., 2019. Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling. ACL
  - Matthew E. Peters et al., 2019. Knowledge Enhanced Contextual Word Representations. EMNLP



Barack's Wife Hillary:
Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV* **Nelson F. Liu[†§]** **Matthew E. Peters[§]**
Matt Gardner[§] **Sameer Singh***

*University of California, Irvine, CA, USA
†University of Washington, Seattle, WA, USA
§Allen Institute for Artificial Intelligence, Seattle, WA, USA
{rlogan, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nfliu@cs.washington.edu

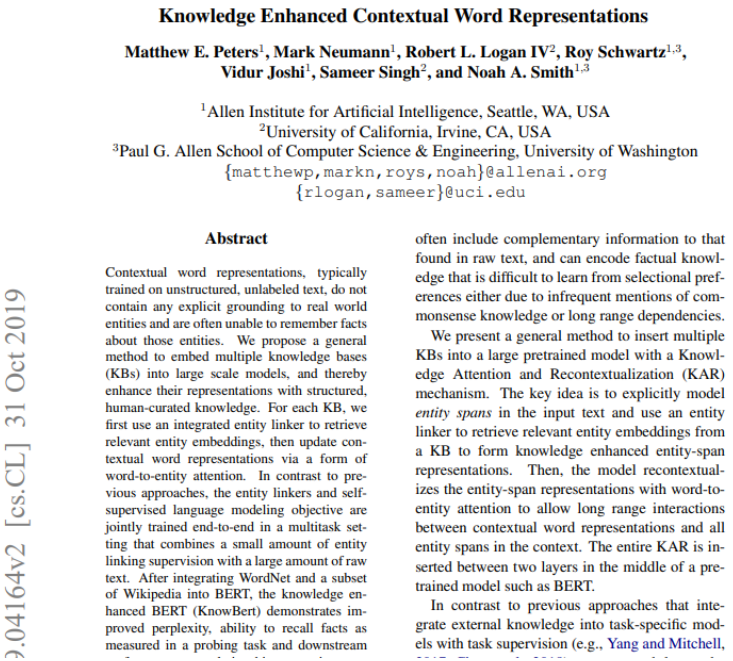
Abstract

Modeling human language requires the ability to not only generate fluent text but also encode factual knowledge. However, traditional language models are only capable of remembering facts seen at training time, and often have difficulty recalling them. To address this, we introduce the knowledge graph language model (KGLM), a neural language model with mechanisms for selecting and copying facts from a knowledge graph that are relevant to the context. These mechanisms enable the model to render information it has never seen before, as well as generate out-of-vocabulary tokens. We also introduce the *Linked WikiText-2* dataset,[1] a corpus of annotated text aligned to the Wikidata knowledge graph whose contents (roughly) match the popular *WikiText-2* benchmark (Merity et al., 2017). In experiments, we demonstrate that the KGLM achieves significantly better performance than a strong baseline language model. We additionally compare different language models' ability to complete sentences requiring factual knowledge,

[*Super Mario Land*] is a [*1989*] [*side-scrolling*] [*platform video game*] developed and published by [*Nintendo*] as a [*launch title*] for their [*Game Boy*] [*handheld game console*].

Figure 1: ***Linked WikiText-2* Example**. A localized knowledge graph containing facts that are (possibly) conveyed in the sentence above. The graph is built by iteratively linking each detected entity to Wikidata, then adding any relations to previously mentioned entities. Note that not all entities are connected, potentially due to missing relations in Wikidata.

training. For instance, when conditioned on the text at the top of Figure 1, an AWD-LSTM language



Knowledge Enhanced Contextual Word Representations

Matthew E. Peters[1], Mark Neumann[1], Robert L. Logan IV[2], Roy Schwartz[1,3], Vidur Joshi[1], Sameer Singh[2], and Noah A. Smith[1,3]

[1]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[2]University of California, Irvine, CA, USA
[3]Paul G. Allen School of Computer Science & Engineering, University of Washington
{matthewp,markn,roys,noah}@allenai.org
{rlogan,sameer}@uci.edu

Abstract

Contextual word representations, typically trained on unstructured, unlabeled text, do not contain any explicit grounding to real world entities and are often unable to remember facts about those entities. We propose a general method to embed multiple knowledge bases (KBs) into large scale models, and thereby enhance their representations with structured, human-curated knowledge. For each KB, we first use an integrated entity linker to retrieve relevant entity embeddings, then update contextual word representations via a form of word-to-entity attention. In contrast to previous approaches, the entity linkers and self-supervised language modeling objective are jointly trained end-to-end in a multitask setting that combines a small amount of entity linking supervision with a large amount of raw text. After integrating WordNet and a subset of Wikipedia into BERT, the knowledge enhanced BERT (KnowBert) demonstrates improved perplexity, ability to recall facts as measured in a probing task and downstream

often include complementary information to that found in raw text, and can encode factual knowledge that is difficult to learn from selectional preferences either due to infrequent mentions of commonsense knowledge or long range dependencies.

We present a general method to insert multiple KBs into a large pretrained model with a Knowledge Attention and Recontextualization (KAR) mechanism. The key idea is to explicitly model *entity spans* in the input text and use an entity linker to retrieve relevant entity embeddings from a KB to form knowledge enhanced entity-span representations. Then, the model recontextualizes the entity-span representations with word-to-entity attention to allow long range interactions between contextual word representations and all entity spans in the context. The entire KAR is inserted between two layers in the middle of a pretrained model such as BERT.

In contrast to previous approaches that integrate external knowledge into task-specific models with task supervision (e.g., Yang and Mitchell,

IV. Applications

- Domain Specific NLP Task
  - Biomedical 도메인 적용(Yuan Ling et al., 2017)
    → 문서 분류, 정보검색, 관계추출 테스크에 좋은 성능을 보임
  - Legal 도메인 적용 예정(Haoxi Zhong et al., 2020)
    → 판결예측, 질의응답, 유사판례검색, 판결문요약 테스크에 좋은 성능을 보일 것(KG 구축 필요)

Biomedical AI

Legal AI

Yuan Ling, Yuan An, Mengwen Liu, Sadid A. Hasan, Yetian Fan, Xiaohua Huy. 2017. Integrating extra knowledge into word embedding models for biomedical NLP tasks. IEEE.
Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. arXiv:2004.12158 (ACL 2020 accepted)

V. Conclusion

- 일반적인 NLP 방법론의 한계: 논리적인 언어 이해를 위해 필요한 **지식기반 추론** 불가
- 지식기반 추론을 위해 Knowledge Base를 이용하는 NLP 방법론 대두
- 고유명사가 포함된 문장의 이해와 생성 능력 향상
- Domain Specific NLP Task에서 특히 유망한 접근법

감사합니다