Knowledge-Guided NLP

Zhiyuan Liu

liuzy@tsinghua.edu.cn

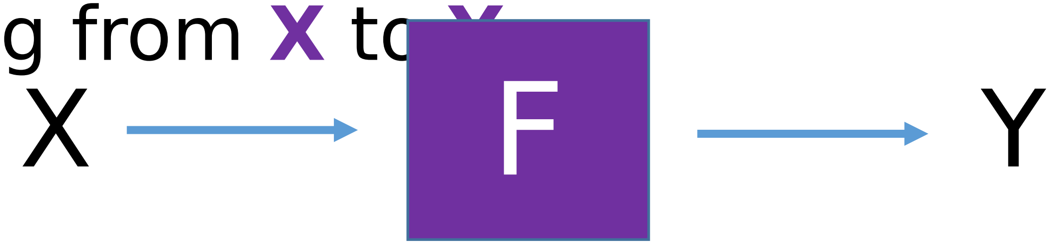
THUNLP

Language and Knowledge

THUNLP

Success of Deep Learning

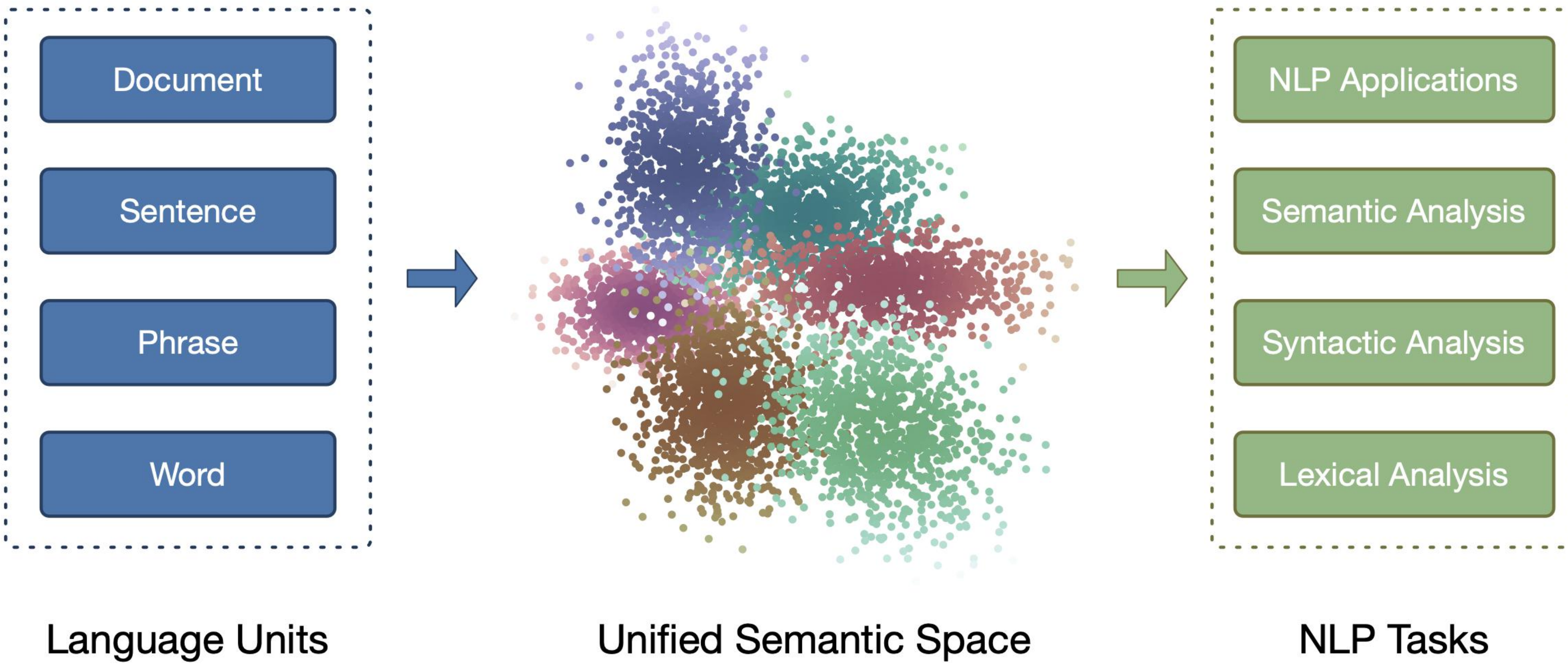
- Deep learning has achieved huge success in natural language processing by using
  - Distributed representation
  - Deep architecture
  - Large-scale data
- By deep learning, we now can learn powerful mapping from **X** to **Y**



Success of Deep Learning in NLP

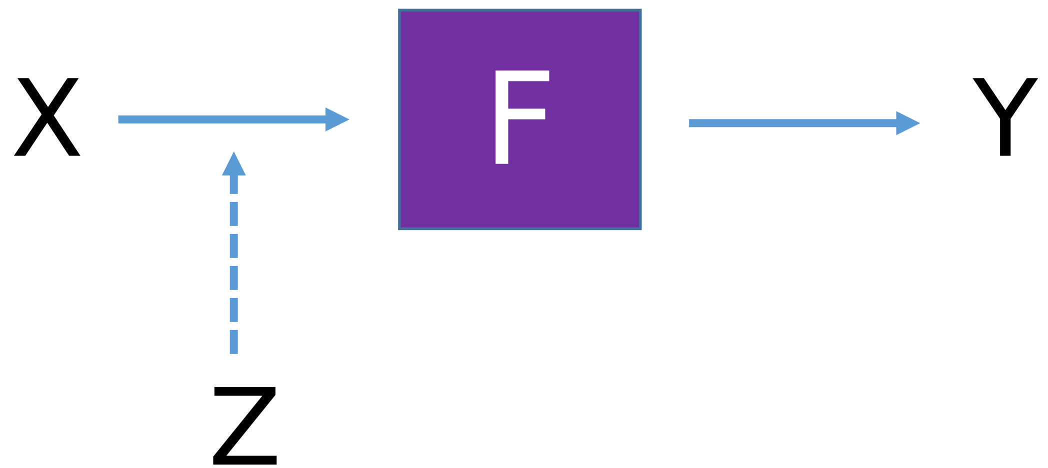
Machine Translation	**Chinese**: 我很开心	**English**: I am happy
Sentiment Classification	**Text**: I like this book	**Sentiment**: Satisfaction
Information Extraction	**Text**: Prof. Liu teaches NLP	**Relation**: Prof. Liu – NLP class
Reading Comprehension	**Text and query**: …	**Answer**: …

Success of Deep Learning in NLP



- By mapping all the language units to a **unified semantic space**, we can now learn complex interactions between these units and boost downstream NLP tasks

Challenge of Deep Learning



- However…
  - Sometimes, we need **extra knowledge** (**Z**) to understand **X** and map it to the correct **Y**.

Challenge of Deep Learning

夏天的天气像个大烤箱。

Challenge of Deep Learning

夏天的天气像个大烤箱。

Translate it to English

Challenge of Deep Learning

夏天的天气像个大烤箱。

Translate it to English

Linguistic Knowledge

The summer is like an oven.

Challenge of Deep Learning

夏天的天气像个大烤箱。

Translate it to English

Linguistic Knowledge

The summer is like an oven.

"oven" means …

Challenge of Deep Learning

夏天的天气像个大烤箱。

Translate it to English — **Linguistic Knowledge**

The summer is like an oven.

“oven” means … — **Commonsense Knowle**

The summer is hot.

Challenge of Deep Learning

夏天的天气像个大烤箱。

Translate it to English

Linguistic Knowledge

The summer is like an oven.

“oven” means …

Commonsense Knowle

The summer is hot.

So where to stay cool …

Challenge of Deep Learning

夏天的天气像个大烤箱。

Translate it to English — **Linguistic Knowledge**

The summer is like an oven.

"oven" means … — **Commonsense Knowle**

The summer is hot.

So where to stay cool — **World Knowledge**

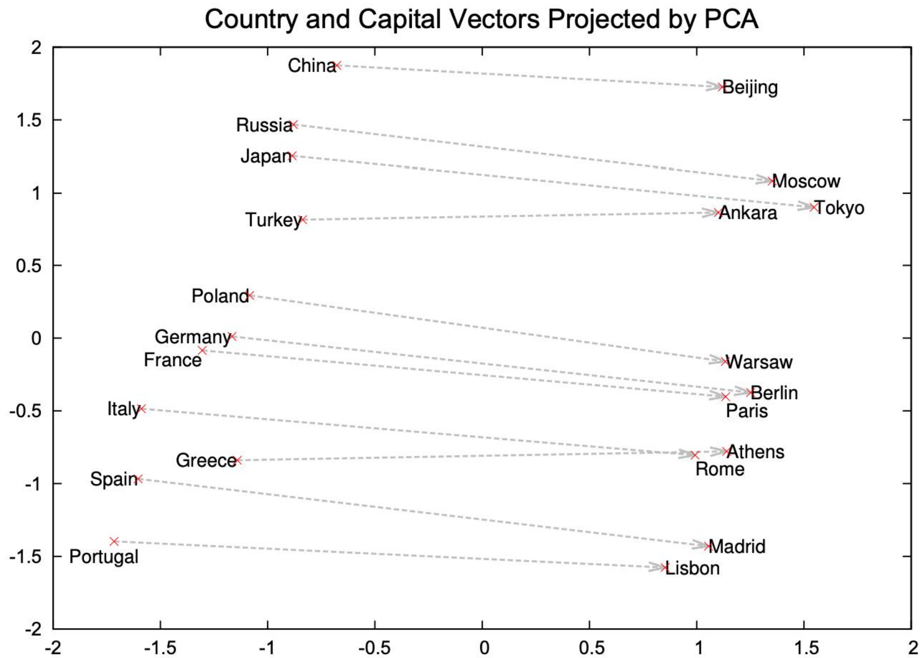
The North

Knowledge for NLP

- We need knowledge to handle certain tasks
  - **Linguistic knowledge**
    - How to use a language
    - Phonetics, morphology, syntax, semantics …
  - **Commonsense knowledge**
    - Something we do not say it out loudly, but everyone knows
    - Water is drinkable. Apple is eatable. Gun is dangerous.
  - **World knowledge**
    - Facts about the world.
    - Bill Gates founded Microsoft. NLP class starts at 19:20.

Knowledge in Text

- There is **implicit knowledge** in text
  - Word embeddings reflect word analogy [1]



King – Man + Woman = Que

[1] Distributed Representations of Words and Phrases and their Compositionality. Mikolov et al. NIPS 2013

Knowledge in Text

- There is **implicit knowledge** in text
  - Pre-trained language models have good **linguistic knowledge** [1]
    - Researchers use representations from language models for various linguistic tasks and get good results.
    - POS tagging, dependency parsing, semantic role labeling…

[1] Linguistic Knowledge and Transferability of Contextual Representations. Liu et al. ACL 2019.

Knowledge in Text

- There is **implicit knowledge** in text
  - We can extract **factual knowledge** from text

William Henry Gates III (born October 28, 1955) is an American business magnate, software developer, investor, and philanthropist. He is best known as the co-founder of Microsoft Corporation.[2][3] During his career at Microsoft, Gates held the positions of chairman, chief executive officer (CEO), president and chief software architect, while also being the largest individual shareholder until May 2014. He is one of the best-known entrepreneurs and pioneers of the microcomputer revolution of the 1970s and 1980s.

- Many information could be extracted from it
  - Birth date, nationality, occupation, founder of ...

Knowledge (NOT) in Text

- Many forms of knowledge are hard to be extracted or learned from the text, or they do not exist in the text.
  - Multiple meanings for words
    - How to efficiently / accurately represent multiple meanings
  - Commonsense
    - Hardly exist in text
  - Factual knowledge
    - Still hard to extract for neural models (especially in an open scenario)
    - How to use

Knowledge in Structural Forms

- Luckily, some people have already gathered knowledge in structural forms (mainly graphs)
  - **Linguistic Knowledge**: WordNet, HowNet
  - **Commonsense Knowledge**: ConceptNet, Atomic
  - **World Knowledge**: Freebase, Wikidata
  - nowle … Financ

WORDNET

Freebase™

ATOMIC

WIKIDATA

知网

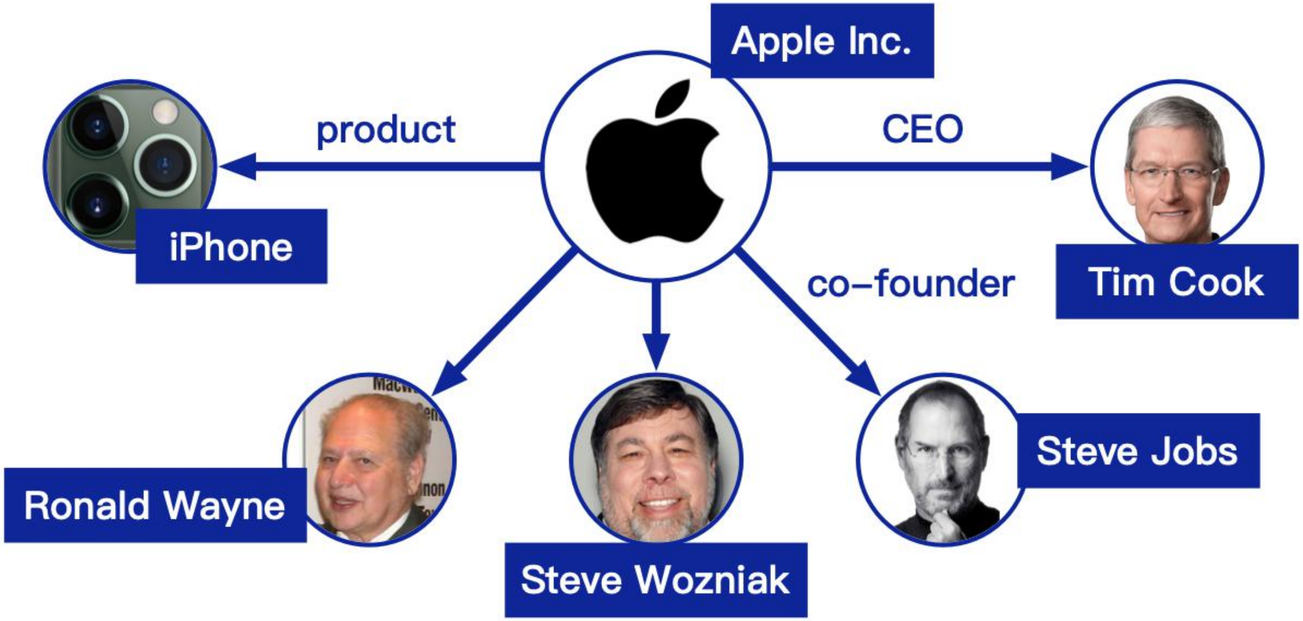
HowNet Knowledge Database

ConceptNet

An open, multilingual knowledge graph

Knowledge in Structural Forms

- World knowledge graph

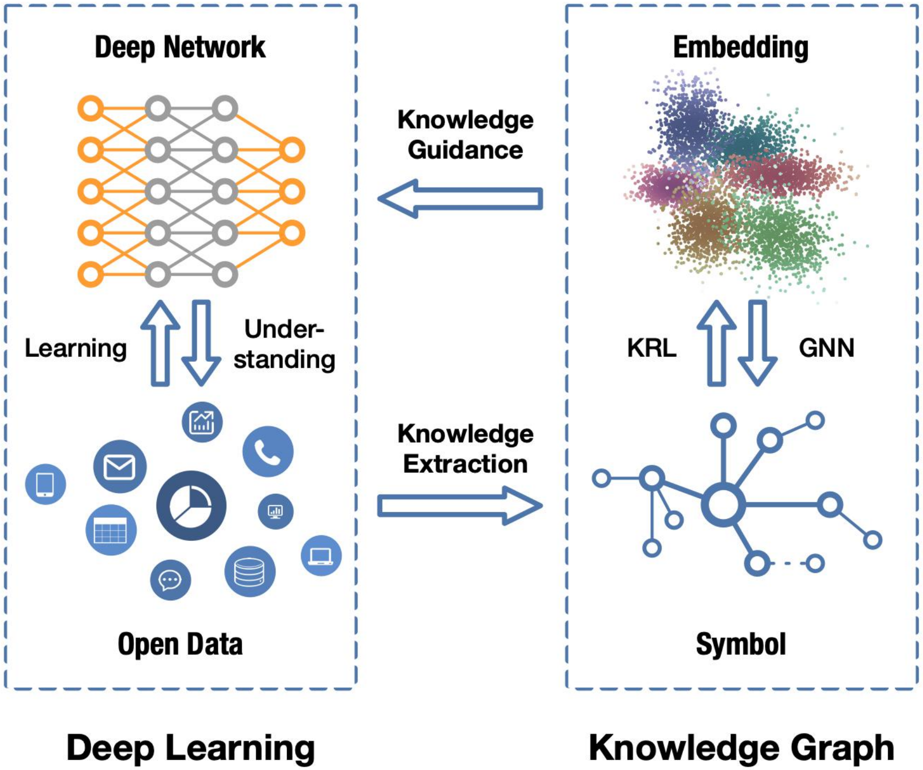


Comparison

- Knowledge from data
  - + Low cost.
  - + Good representation. Small gap to the tasks.
  - – Cannot cover all the knowledge.
  - – Not accurate. Bias.
- Knowledge from external knowledge bases
  - + Accurate. Full.
  - + Flexible.
  - – High cost to annotate.
  - – Not all the knowledge can be represented.
  - – Representation gaps to the tasks.

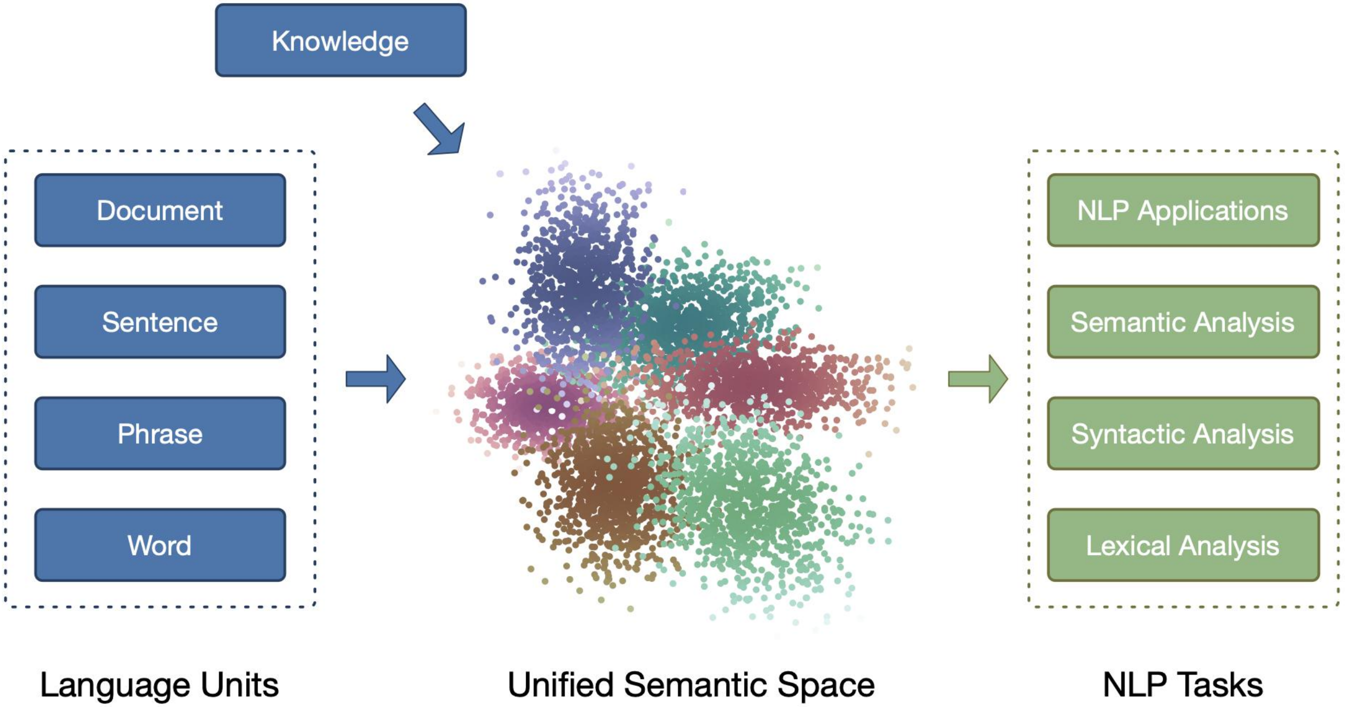
Roadmap

- A better NLP pipeline should be both
  - **Data-driven**: deep learning, large-scale data …
  - **Knowledge-guided**: knowledge graphs, symbol representation …



Roadmap

- To better combine the two
  - We should embed **knowledge** along with other language units in a unified semantic

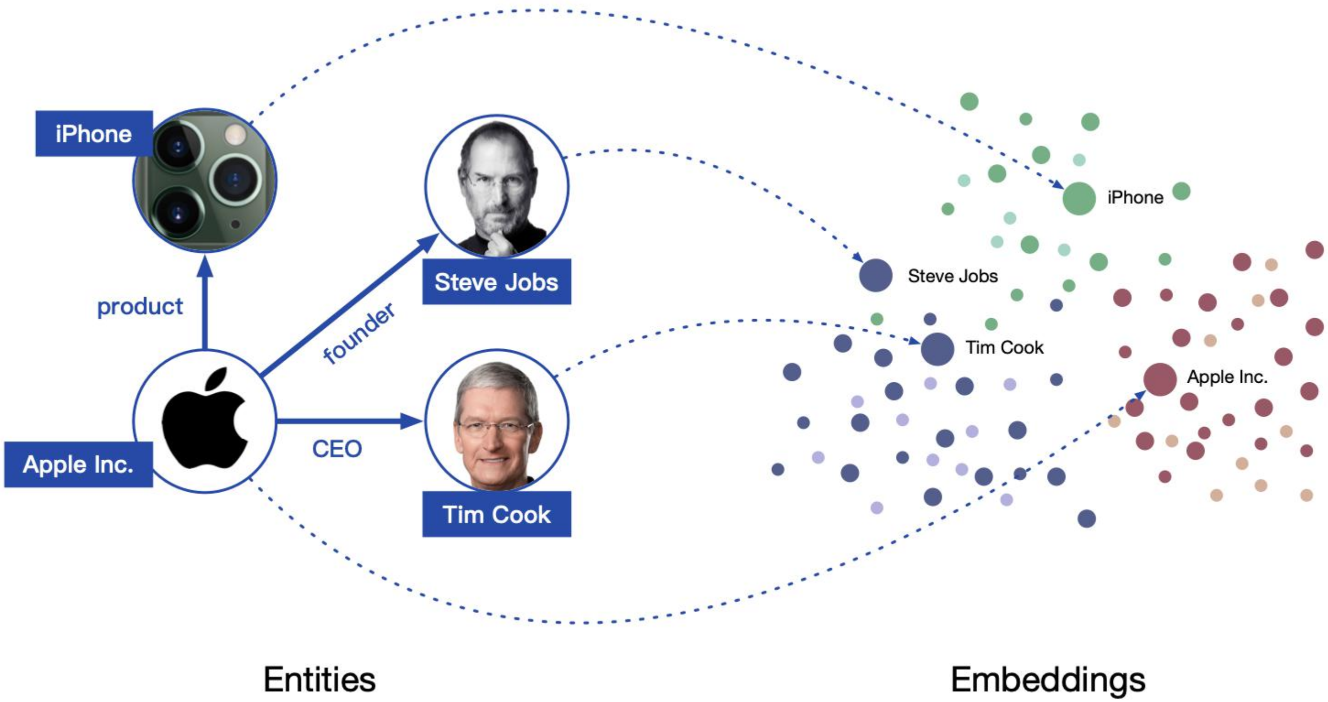


Using Knowledge Graph Embedding

THUNLP

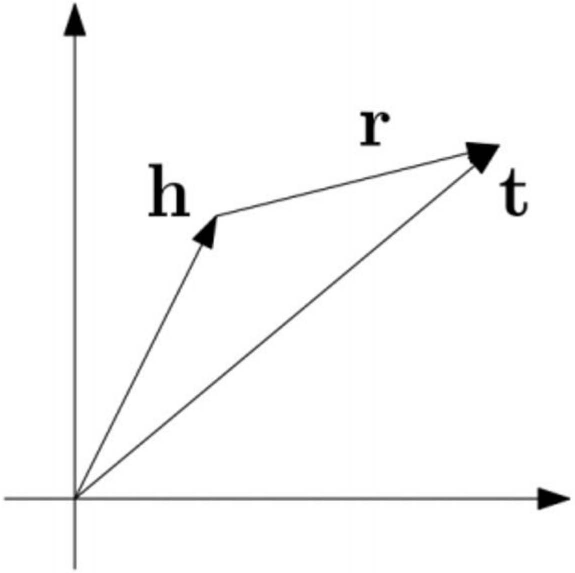
Knowledge Graph Embedding

- What is knowledge graph embedding (KGE)
  - Map entities and relations into low-dimensional embeddings



Knowledge Graph Embedding

- TransE [1]: a classical KGE method
  - For (h, r, t), see **t** as an **r** translation of **h**.

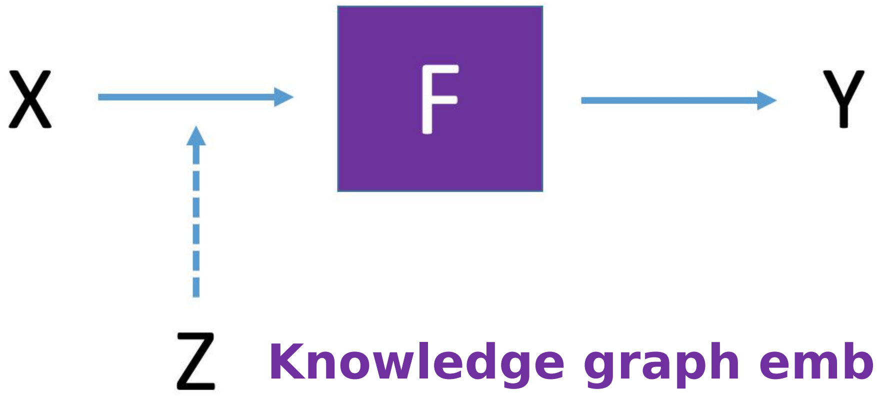


Learning objective: **h + r = t**

[1] Translating embeddings for modeling multi-relational data. Bordes, et al. NIPS 2013.

Knowledge Graph Embedding

- What we can do with KGE?
  - Knowledge graph applications
    - Triple classification
    - Link prediction
  - **Integrate KGE into NLP models**
    - As an external source of knowledge



Knowledge graph embedding

Knowledge Graph Embedding

- Integrate KGE into NLP models
  - In this part, we mainly focus on how to integrate **entity embeddings** (learned with some KRL methods like TransE) into NLP models.
  - Three ways to do it
    - **Feature-based methods**
    - **Attention-based methods**
    - **Knowledge-guided pre-trained models**

Feature-based Methods

- Treat KGE as a source of feature
  - What features do we already have?
    - Word embeddings
    - POS tagging *
    - Entity types *
    - [New!] **Knowledge graph embedding**
  - Entity embeddings learned by KGE methods contain semantic information about the entities
    - Type, properties, relations to others

Feature-based Methods

- Step 1: Identify entities in the text
  - Named entity recognition
  - Entity linking
- Some tasks provide annotated entities

Apple's hardware products include the iPhone smart phone, the

Apple Inc.: an American technology company

Apple: a kind of fruit

Feature-based Methods

- Step 2: Take in the entity embeddings
  - Denote the entity sequen $\{e_1, \ldots, e_m\}$
  - Formally, the output representation can be calculated by

$$\boldsymbol{O} = \mathrm{Encoder}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n; \boldsymbol{e}_1, \ldots, \boldsymbol{e}_m)$$

  - A simple way to do it is to separately calculate the output representations for the **word sequence** (text) and the **entity sequence**

Feature-based Methods

- Step 2: Take in the entity embeddings
  - Word sequence

$$\boldsymbol{O}_w = \mathrm{Encoder}_w(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$$

    - The encoder can be CNN, RNN, Transformers ...
  - Entity sequence

$$\boldsymbol{O}_e = \mathrm{Encoder}_e(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m)$$

  - Eventually, we combine the two sources by

$$\boldsymbol{O} = [\boldsymbol{O}_w, \boldsymbol{O}_e]$$

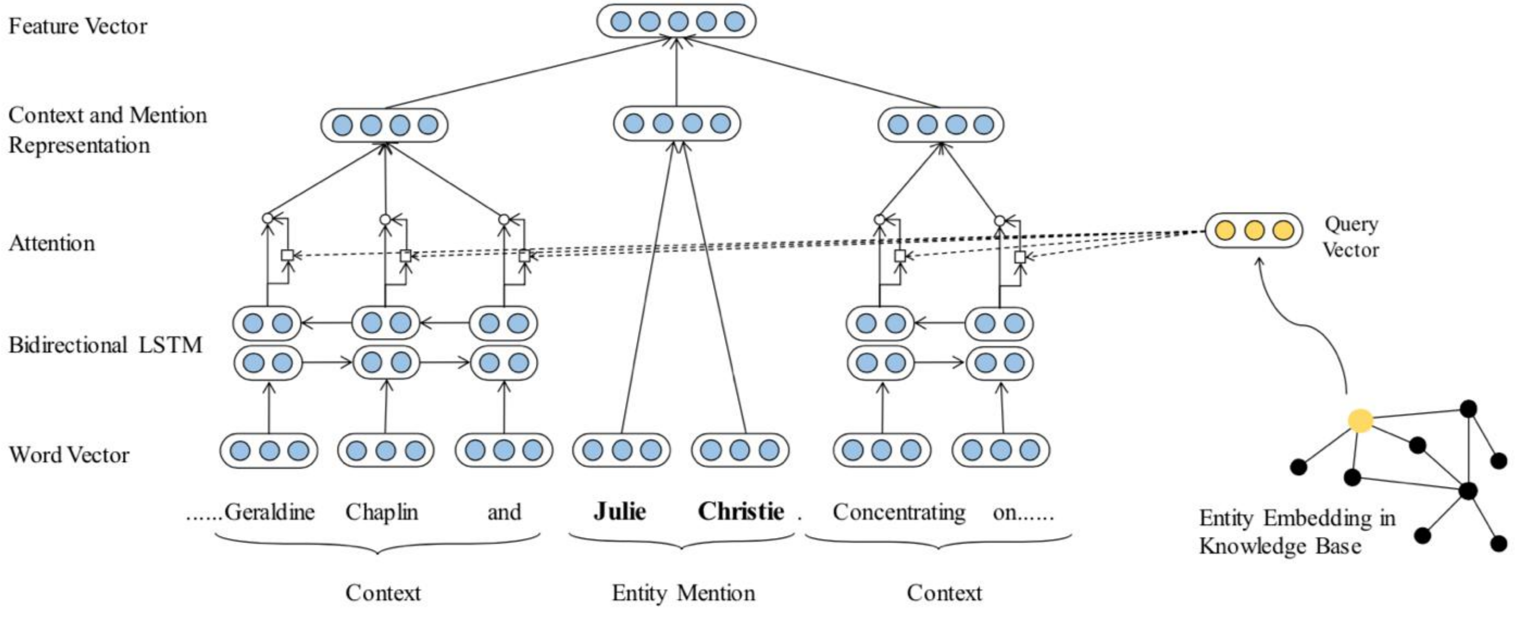
Concatenating them as one

Feature-based Methods

- Many variants for feature-based methods
  - Combine word embeddings and entity embeddings at the feature step
    - How to combine? Sum, concatenation …
  - Add projections
    - There are **gaps** between word semantic space and knowledge semantic space, since they are not trained together
    - Projections on entity embeddings, on O_e, …
  - Early fusion between words and entities
    - E.g., add entity representation to word representation

Attention-based Methods

- Attention mechanism can capture **semantic correlations** between text and entities



Improving Neural Fine-Grained Entity Typing with Knowledge Attention. Xin et al. AAAI 2018

Attention-based Methods

- This method is usually used in **entity-centered** tasks
  - Relation extraction: two entities of interest
  - Entity typing: one entity of interest
- Words that are highly informative for one task may not be relevant for other tasks
- By using attention, we can assign parts of text that have higher correlation with center entities higher weights
  - … **Bill Gates**, an American entrepreneur, …

Attention-based Methods

- Formally, for each center entity, we have

$$\boldsymbol{h}_{e_i} = \boldsymbol{W}^T \text{softmax}(\boldsymbol{e}_i \boldsymbol{A}_{\boldsymbol{k}} \boldsymbol{W})$$

  - $\boldsymbol{h}_{e_i}$: hidden representation for **entity i** after attention
  $\boldsymbol{W}$ $\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$
  - $\boldsymbol{e}_i$: word embedding sequence
  - $\boldsymbol{A}_{\boldsymbol{k}}$: entity embedding for **entity i**
  - : bi-linear matrix in attention
- If there are more than one center entities

$$\boldsymbol{O} = \text{Encoder}_e(\boldsymbol{h}_{e_1}, \ldots, \boldsymbol{h}_{e_m})$$

  - The “encoder” could be CNN, RNN, or sum / concatenation

Knowledge-guided Pretrained Models

- What if we combine both **feature-based** and **attention-based** methods?
- Take **transformer-based** models as an example
  - **Feature fusion**: aggregate both word embeddings and entity embeddings as one
  - **Self-attention**: each token attends to each other
- Simple combination may lead to overfitting
- How about integrating knowledge in **pre-trained models**?

Knowledge-guided Pretrained Models

- Formally, knowledge-guided pre-trained models can be

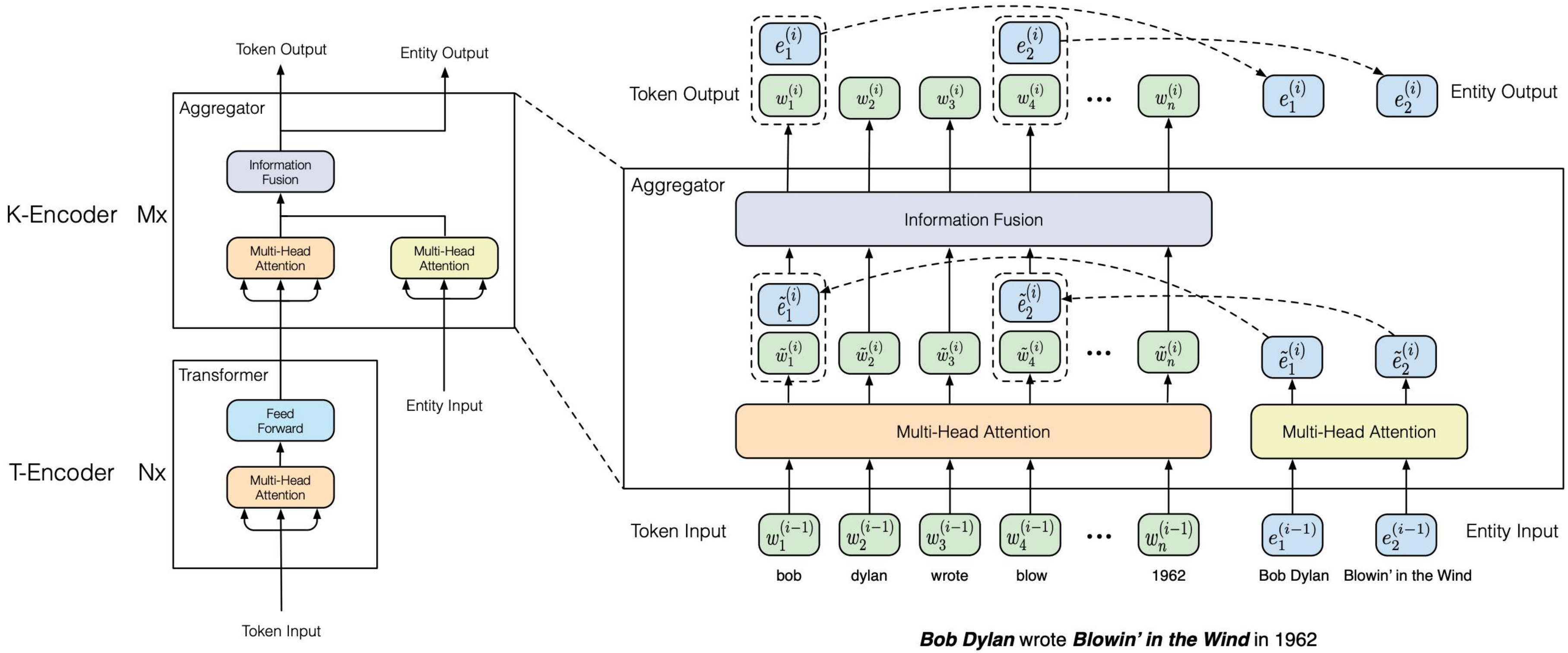
$$\boldsymbol{h}_i = f(\boldsymbol{w}_i, \boldsymbol{e}_i),$$

$$\{\boldsymbol{w}'_1, \ldots, \boldsymbol{w}'_n\} = SA(\{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n\})$$

  - **Function f** is to aggregate information from both words and entities
  - **Function SA** refers to self-attention
- Finally, we ' $\boldsymbol{O} = \text{Encoder}_w(\boldsymbol{w}'_1, \ldots, \boldsymbol{w}'_n)$

Knowledge-guided PLM: ERNIE

- **ERNIE** is the first work about knowledge-guided PLM

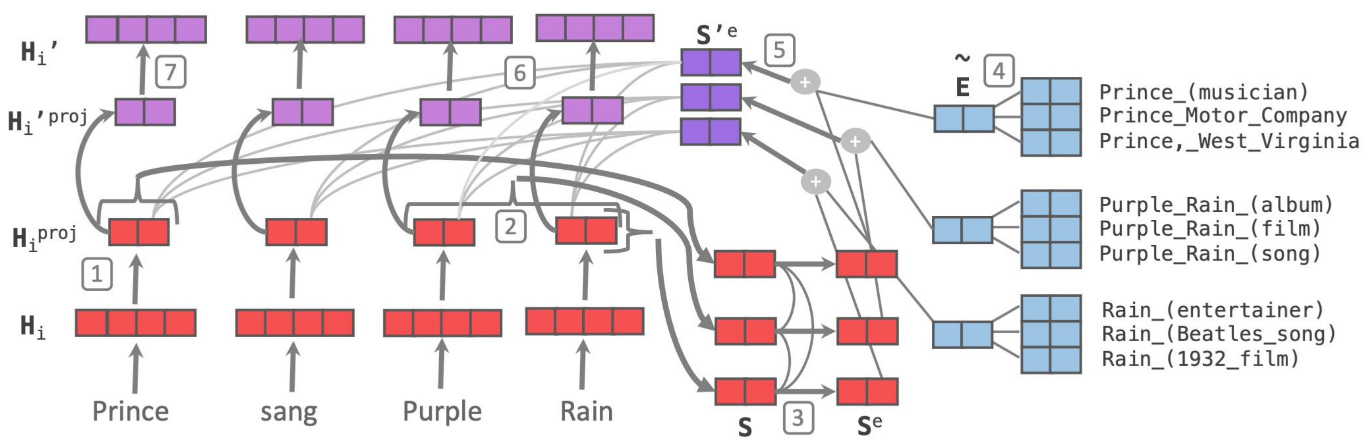


(a) Model Achitecture (b) Aggregator

ERNIE: Enhanced Language Representation with Informative Entities. Zhang et al. ACL 2019

Knowledge-guided PLM: KnowBERT

- **KnowBERT** has the similar idea as **ERNIE**
  - KnowBERT uses an integrated entity linker and adopts end-to-end training
  - Besides using world knowledge, KnowBERT also adopts linguistic knowledge by introducing **WordNet**
    - WordNet can also be organized as a knowledge



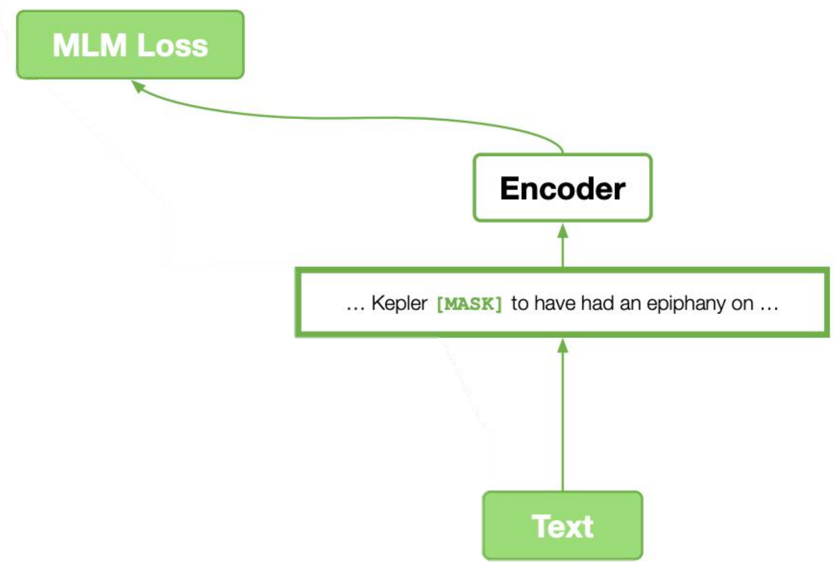
Knowledge Enhanced Contextual Word Representations. Peters et al. EMNLP 2019.

A Different Direction

- For now, all these knowledge-guided models use **pre-trained and fixed** knowledge graph embeddings.
  - External KGE **cannot be easily aligned** to language representation space
  - They require entity linkers to identify entities, which bring **error propagation** and **time overhead**
- Maybe…
  - We can design **a new training objective** to integrate world knowledge into PLMs, without using pre-trained KGE

A Different Direction: KEPL

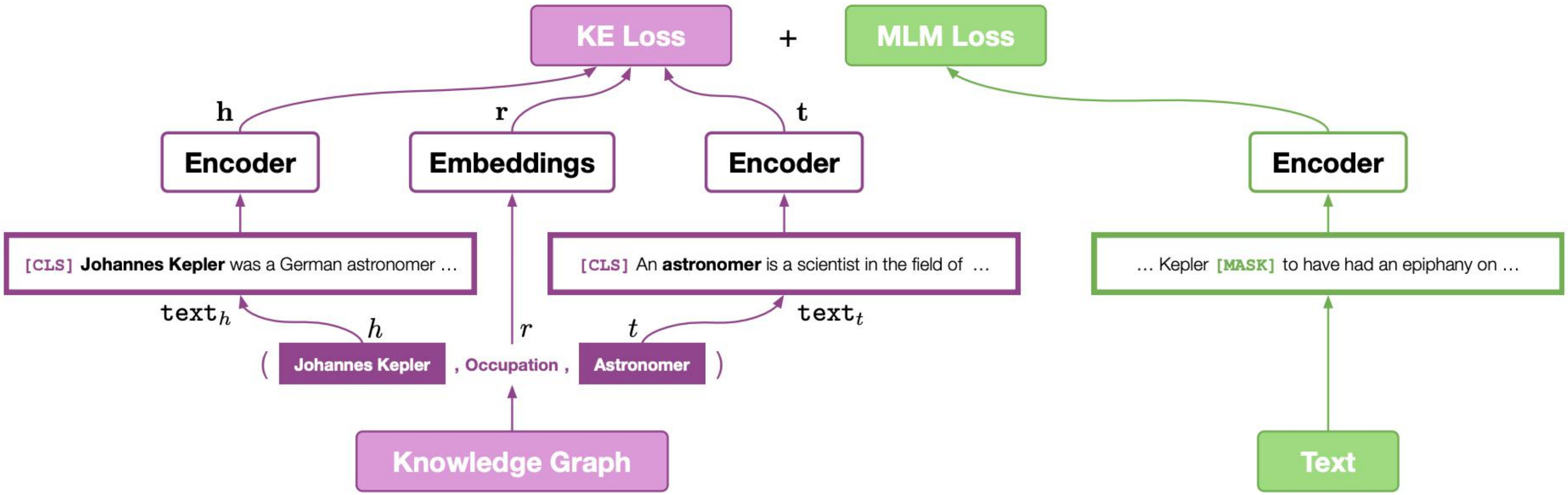
- Normal PLMs:



KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. Wang et al.

A Different Direction: KEPLER

- KEPLER:

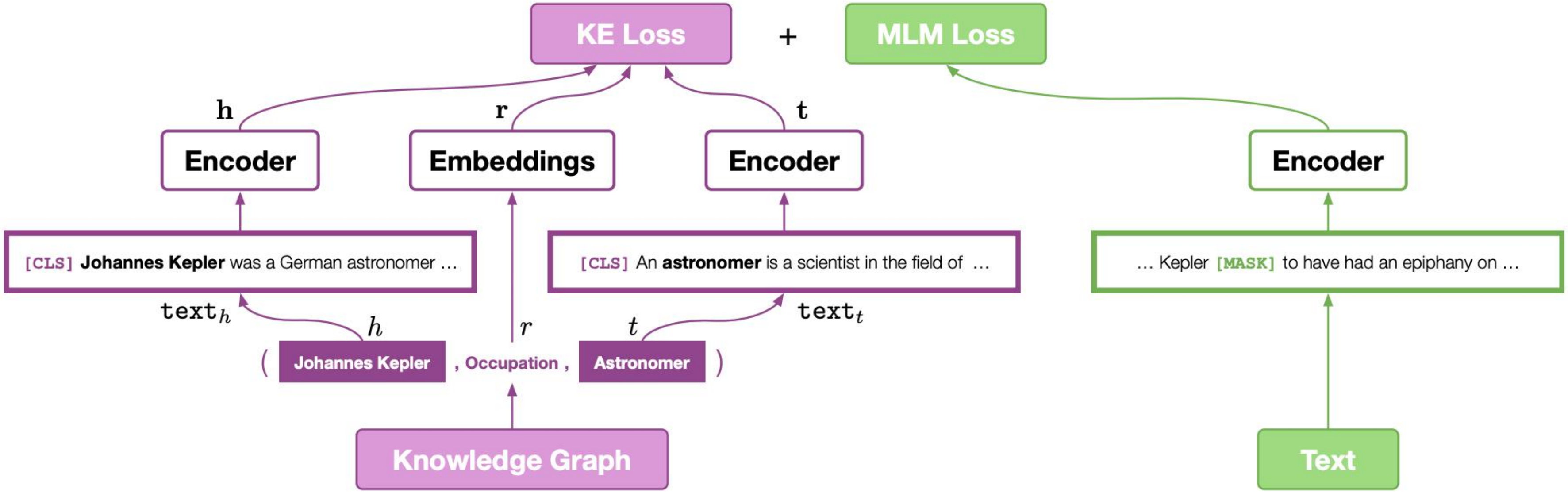


- Adding a new loss: **Knowledge embedding (KE) loss**
  - e.g., TransE loss

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. Wang et al.

A Different Direction: KEPL

- KEPLER:



- How to get h, t, r?
  - r: like in TransE
  - h, t: encoded from **entity descriptions**

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. Wang et al.

A Different Direction: KEPI

- Benefits:
  - Bridge the gap between the language space and the knowledge space
  - Better language understanding ability towards knowledge-oriented tasks
  - NO overhead or error propagation caused by entity linkers
- There are also other models using knowledge-guided pre-training objectives instead of KGE
  - **Matching the blank** [1], **K-adapter** [2], **Pretrained Encyclopedia** [3]

[1] Matching the Blanks: Distributional Similarity for Relation Learning. Soares et al. ACL 2019.

[2] K-ADAPTER: Infusing Knowledge into Pre-Trained Models with Adapters. Wang et al.

[3] Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. Xiong et al. ICLR 2020.

Applications

- Entity-oriented application
  - Relation extraction
  - Entity typing
- General NLP application
  - Language modeling
  - Natural language understanding
  - Information retrieval
  - …

Applications

- For entity-oriented applications, knowledge-guided NLP models can bring large promotion

Model	Acc.	Macro	Micro
NFGEC (Attentive)	54.53	74.76	71.58
NFGEC (LSTM)	55.60	75.15	71.73
BERT	52.04	75.16	71.63
ERNIE	**57.19**	**76.51**	**73.39**

Results on FIGER (entity typing)

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**

Results on Open Entity (entity typing)

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**

Results on FewRel and TACRED (relation extrac

ERNIE: Enhanced Language Representation with Informative Entities. Zhang et al. ACL 2019

Applications

- However, for general NLP tasks ...

Model	MNLI-(m/mm) 392k	QQP 363k	QNLI 104k	SST-2 67k
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	-	93.5
ERNIE	84.0/83.2	71.2	91.3	93.5

Model	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k
$\text{BERT}_{\text{BASE}}$	52.1	85.8	88.9	66.4
ERNIE	52.3	83.2	88.2	68.8

Results on GLUE (language understanding)

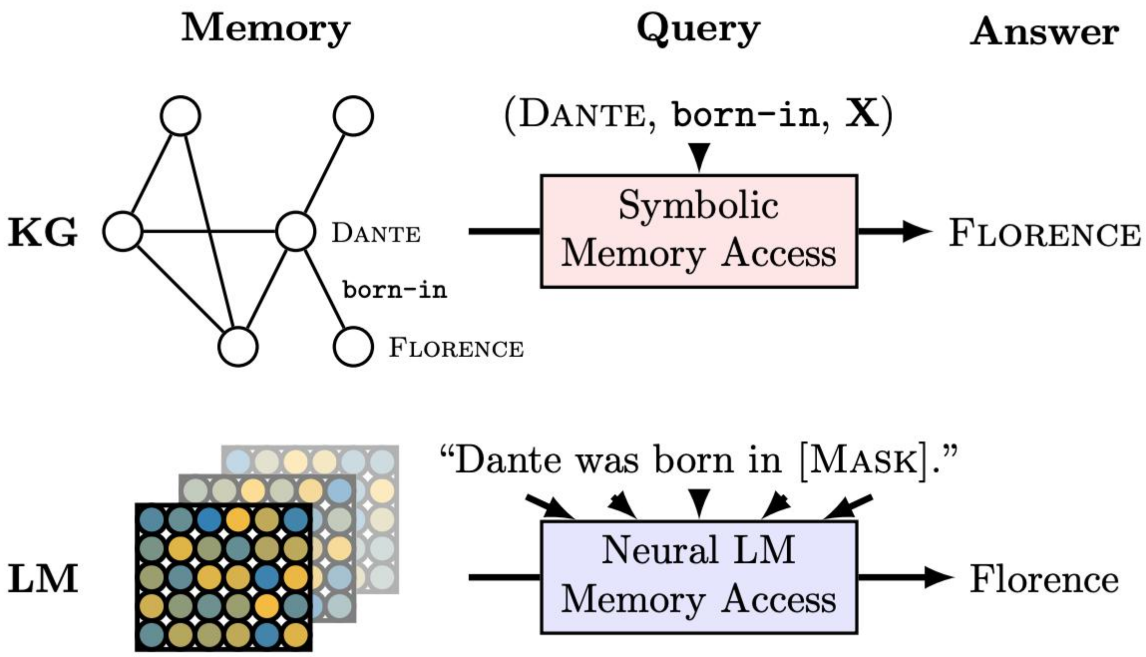
ERNIE: Enhanced Language Representation with Informative Entities. Zhang et al. ACL 2019

Applications

- For general NLP tasks, knowledge-guided models do not perform well
  - They do not need world knowledge that much
  - Few entity mentions
  - PLMs already have knowledge?

Language Models as Knowledge Bases?

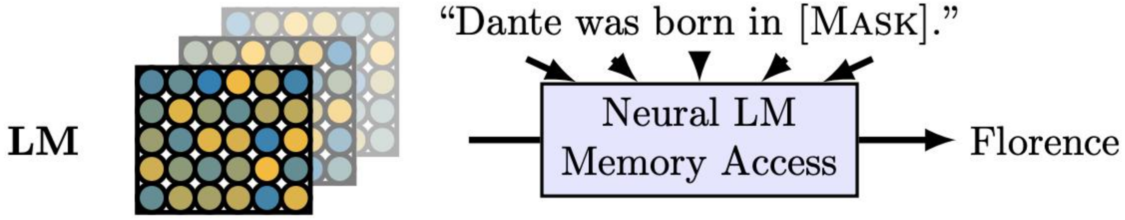
- What if pre-trained language models already have knowledge?



Language Models as Knowledge Bases? Petroni et al. EMNLP 2019.

Language Models as Knowledge Bases?

- What if pre-trained language models already have knowledge?
  - By using **masked language modeling** objective and design **templates** for retrieving knowledge, PLMs seem to contain knowledge.
  - (Bill Gates, founder, Microsoft) -> Bill Gates founded **[MASK]** **-> Microsoft**



Language Models as Knowledge Bases? Petroni et al. EMNLP 2019.

Language Models as Knowledge Bases?

- The results are … pretty positive

Corpus	Relation	Statistics #Facts	#Rel	Baselines Freq	DrQA	KB RE_n	RE_o	LM Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	`birth-place`	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	`birth-date`	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	`death-place`	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naïve entity linking (RE_n), oracle entity linking (RE_o), fairseq-fconv (Fs), Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl) across the set of evaluation corpora.

Language Models as Knowledge Bases? Petroni et al. EMNLP 2019.

Language Models as Knowledge Bases?

- The results are … pretty positive
  - But still there are a lot of problems about **LAMA**
    - The gap between the templates and training corpora
    - Cannot retrieve != do not know
  - There are some later works about probing knowledge in PLMs
    - How Can We Know What Language Models Know? Jiang et al.
    - Negated LAMA: Birds cannot fly. Kassn
    - …

Language Models as Knowledge Bases? Petroni et al. EMNLP 2019.

What's Next?

- More knowledge
  - Linguistic knowledge, commonsense knowledge, domain-specific knowledge …
- Deep fusion between features
  - Reduce the gap between language semantic space and knowledge semantic space
- Heterogeneous Modals
  - Multi-modal data
- Reasoning over knowledge
- Probing knowledge in language models

Commonsense Knowledge

THUNLP

What is Commonsense

- Some examples
  - **Water** is drinkable
  - **Apple** is eatable
  - **Bridge** is something over the river for people to travel from one side of the bank to the other
  - If you got shot by a **gun**, you would be hurt (probably you would die)

What is Commonsense

Learned from interacting with the world

- **Touching the ice, you learn that ice is cold**
- **Seeing that apple is red**

What is Commonsense

Education free

- **You do not need to go to college to have a normal conversation with other people**

What is Commonsense

Can be omitted in conversations

- **You won’t say “ice is cold” or “apple is red” in a normal conversation**

Why we need commonsens

- **It's common**
  - Mike got shot -> what will happen?
  - Eric fell into the river -> what should we do?
    - A. Eat an apple
    - B. Call the police
- **It's hard to learn**
  - Learn from interactions?
    - Cross-modal? With vision? With robots?
  - Learn from text?
    - Most commonsense can be omitted from conversations

ConceptNet

- A large commonsense knowledge graph
  - Core relations: **36**
  - Vocabulary size: **1,803,873**
- What's in the vocabulary?
  - Common objects, like apple, water, bridge, word, …
  - Different from world knowledge graph (containing entities)

ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. Speer et al. AAAI 2017

ConceptNet

Properties of apple

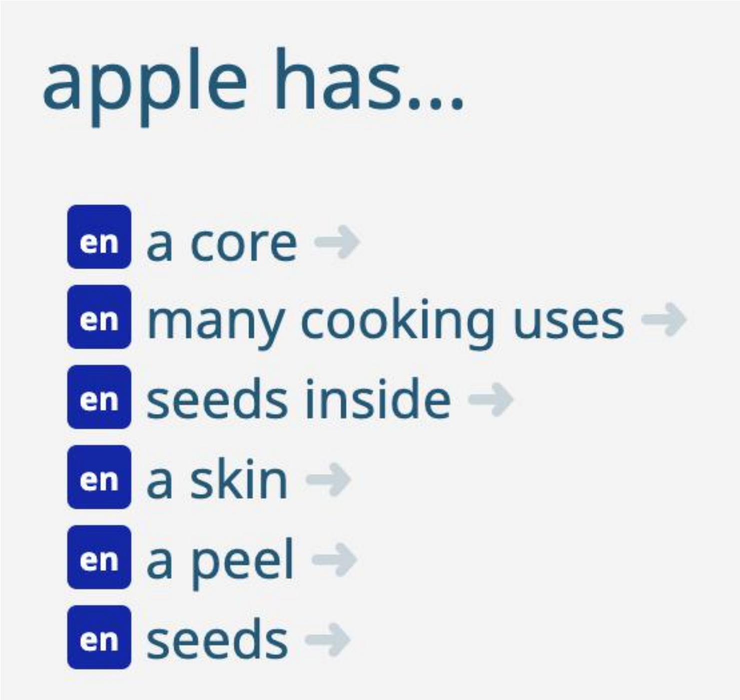
- en red →
- en green →
- en one kind of fruit →
- en opaque →
- en red or green →

Location of apple

- en apple tree →
- en a grocery store →
- en an apple orchard →
- en apple pie →
- en the fridge →
- en a greengrocer →
- en a horses mouth →
- en a house →
- en an orchard →
- en the produce section of the supermarket →
- en the refrigerator →
- en a toffee apple →

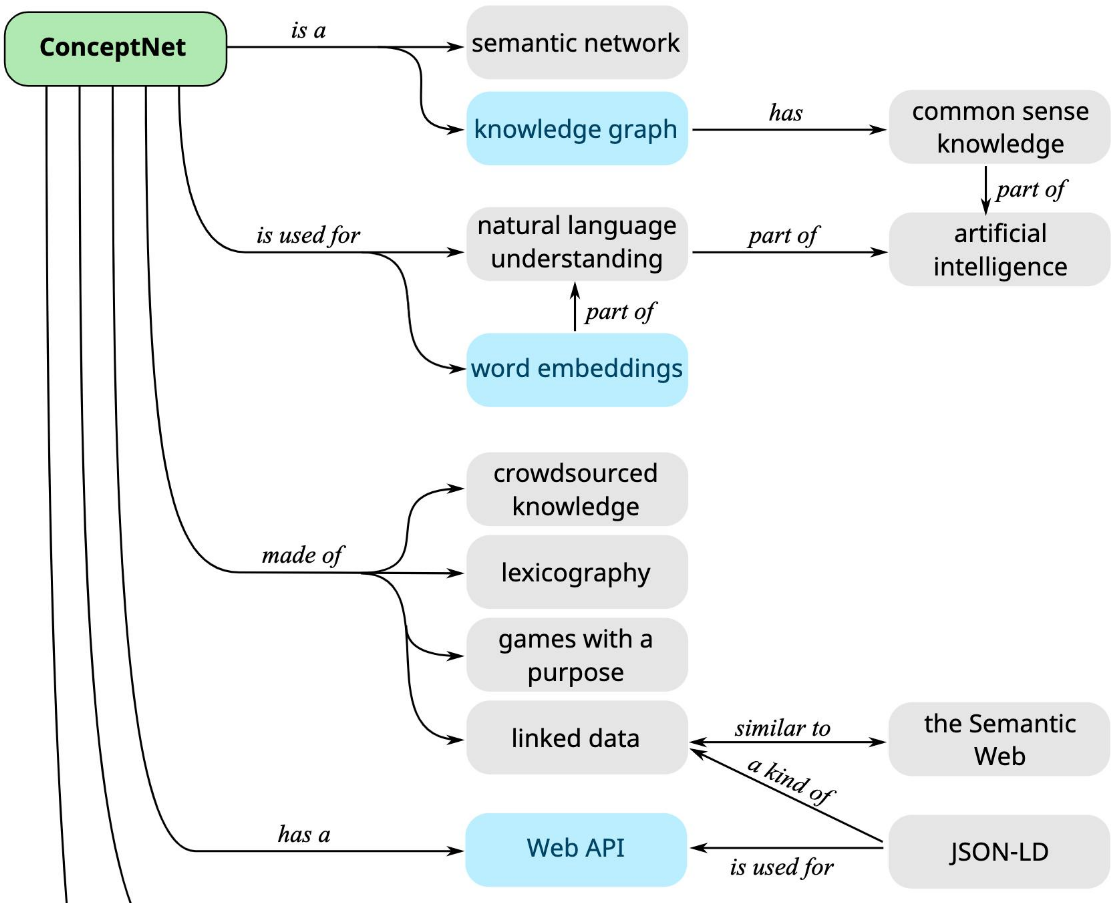
ConceptNet

- Even more interesting



ConceptNet

- ConceptNet mainly focuses on properties of objects and relations between them



Atomic

- A commonsense dataset for if-then rea

Event: `PersonX pushes PersonY down`

Annotated causes for PersonX

Because PersonX wanted (`xIntent`): to harm PersonY; to be in control; to show their power; to show their domination; to be in control; mean; hateful
Before, PersonX needed (`xNeed`): to go over to Person Y; to see Person Y; to get close to PersonY.; to put their weight into PersonY.

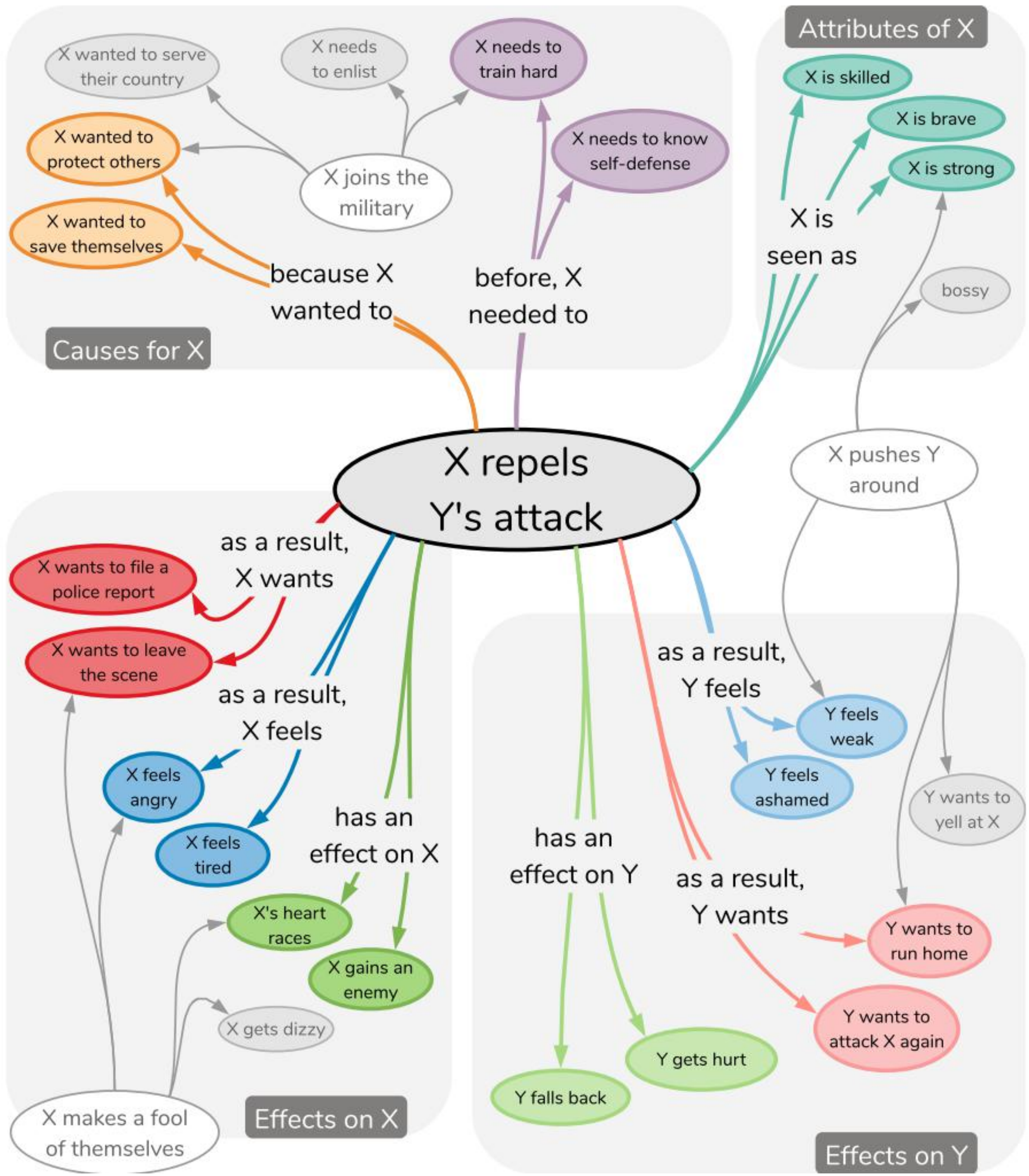
Annotated attributes of PersonX

PersonX is seen as (`xAttr`): dangerous; troublesome; NERVOUS; STUPID; Mean; Awful

Annotated effects on PersonX

As a result, PersonX feels (`xReact`): empowered; powerful; dominant; powerful
As a result, PersonX wants (`xWant`): to apologize to Person Y; to help Person Y up; to punch personY.; to help personY up.; to continue the fight.; to go get some ice.
Has an effect on PersonX (`xEffect`): has an adrenaline rush; laughs; gets kicked by PersonY; get yelled at by PersonY; person X gets in a fight; person X inflicts pain

ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. Sap et al. AAAI 2019


X repels Y's attack
Causes for X
because X wanted to
X wanted to serve their country
X needs to enlist
X needs to train hard
X needs to know self-defense
X wanted to protect others
X joins the military
X wanted to save themselves
before, X needed to
Attributes of X
X is seen as
X is skilled
X is brave
X is strong
bossy
X pushes Y around
as a result, X wants
X wants to file a police report
X wants to leave the scene
as a result, X feels
X feels angry
X feels tired
has an effect on X
X's heart races
X gains an enemy
X gets dizzy
X makes a fool of themselves
Effects on X
has an effect on Y
as a result, Y feels
Y feels weak
Y feels ashamed
Y wants to yell at X
as a result, Y wants
Y wants to run home
Y wants to attack X again
Y gets hurt
Y falls back
Effects on Y

Tasks for commonsense

- Commonsense is important, but
  - In many cases, commonsense is not (heavily) needed for solving the problem
  - Many tasks are not designed to specifically evaluate commonsense abilities
- Commonsense datasets are needed!
  - Swag & HellaSwag
  - Winograd Schema Challenge
  - Commonsense QA
  - ...

Swag & HellaSwag

On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
d) nervously sets her fingers on the keys.

A girl is going across a set of monkey bars. She
a) jumps up across the monkey bars.
b) struggles onto the monkey bars to grab her head.
c) gets to the end and stands on a wooden plank.
d) jumps up and does a back flip.

- Natural language inference + commonsense reasoning

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. Zellers et al. EMNLP 2018.

Winograd Schema Challeng

- An alternative to the Turing test
- A **Winograd schema** is a pair of sentences that differ only in one or two words and that contain a **referential ambiguity** that is resolved **in opposite directions** in the two sentences
- **Winogrand**: an upgraded version at both scale and hardness

The Winograd Schema Challenge. Levesque et al. AAAI 2012.

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. Sakaguchi et al. AAAI 2020

Winograd Schema Challeng

1a. Pete envies Martin *because* **he** is successful.
1b. Pete envies Martin *although* **he** is successful.

Question: Is **he** Pete or Martin?
Answers: 1a – Martin, 1b - Pete

The Winograd Schema Challenge. Levesque et al. AAAI 2012.

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. Sakaguchi et al. AAAI 2020

Winograd Schema Challeng

1a. The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed.*

1b. The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired.*

Question: Is it house or tree?

Answers: 1a – Tree, 1b – house

The Winograd Schema Challenge. Levesque et al. AAAI 2012.

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. Sakaguchi et al. AAAI 2020

CommonsenseQA

*Where on a **river** can you hold a cup upright to catch water on a sunny day?*

✓ **waterfall**, ✗ bridge, ✗ valley, ✗ pebble, ✗ mountain

*Where can I stand on a **river** to see water falling without getting wet?*

✗ waterfall, ✓ **bridge**, ✗ valley, ✗ stream, ✗ bottom

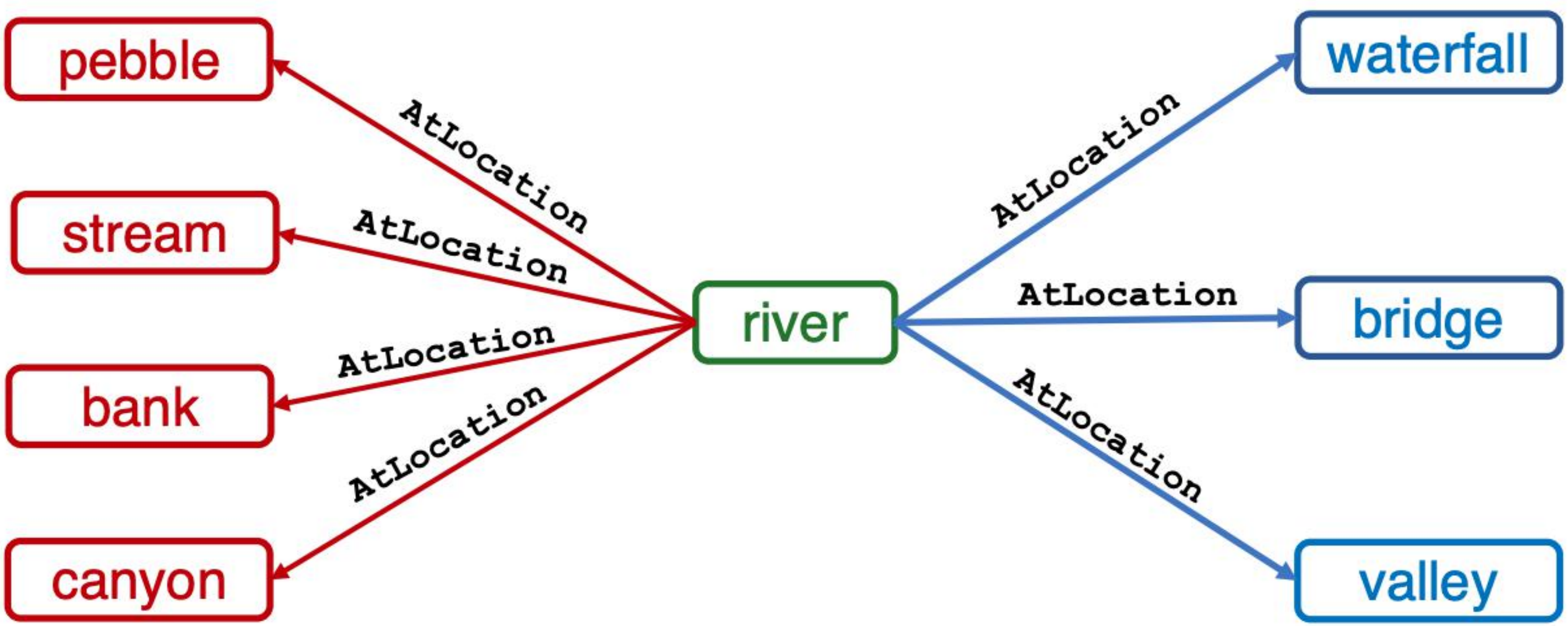
*I'm crossing the **river**, my feet are wet but my body is dry, where am I?*

✗ waterfall, ✗ bridge, ✓ **valley**, ✗ bank, ✗ island

CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge.
Talmor et al. NAACL 2019

CommonsenseQA

- How CommonsenseQA is build?
  - From ConceptNet + Human annotator



CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. Talmor et al. NAACL 2019

How to Use Commonsense

- What have been done with commonsense + NLP
  - Commonsense knowledge base completion
    - COMET : Commonsense Transformers for Automatic Knowledge Graph Construction. Bosselut et al.
    - Commonsense Knowledge Mining from Pretrained Models. Feldman et al.
  - Commonsense reasoning
    - Knowledge-Aware Graph Networks for Commonsense Reasoning. Lin et al.
    - Explain Yourself! Leveraging Language Models for Commonsense Reasoning. Rajani et al.
- Still, a long way to go …

Summary

- Knowledge-guided NLP
  - What is "knowledge"
  - Why we need knowledge
- Leverage knowledge graph embeddings
  - Feature-based, attention-based, pre-trained models
  - An implicit direction (knowledge as objective)
- Commonsense knowledge
  - What is commonsense
  - Commonsense graphs and datasets
  - Latest progress

Summary

- Knowledge is essential for many NLP tasks
- However, there are still many challenges
  - Forms of knowledge representation
  - Knowledge sources
  - Gaps between language space and knowledge space
  - Commonsense
  - ...
- More efforts are needed!

Reading Material

- Must-read papers

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

Neural natural language inference models enhanced with external knowledge. Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. ACL 2018 [link]

Neural knowledge acquisition via mutual attention between knowledge graph and text. Xu Han, Zhiyuan Liu, and Maosong Sun. AAAI 2018 [link]

For reading material recommendation of this course, please refer to our github.

Reading Material

- Further reading

Language Models as Knowledge Bases? [link]

Knowledge enhanced contextual word representations. Matthew EPeters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah ASmith. EMNLP 2019 [link]

Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. Robert Logan, Nelson FLiu, Matthew EPeters, Matt Gardner, and Sameer Singh. ACL 2019 [link]

Knowledgeable Reader: Enhancing Cloze-style Reading Comprehension with External Commonsense Knowledge. Todor Mihaylov and Anette Frank. ACL 2018 [link]

Improving question answering by commonsense-based pre-training. Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2018 [link]

Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. ACL 2018 [link]

For reading material recommendation of this course, please refer to our github.

Q&A

THUNLP

Thanks

Zhiyuan Liu
liuzy@tsinghua.edu.cn
THUNLP