Semantic Knowledge Graphs and Natural Language Processing

Pere-Lluís Huguet Cabot

Reading group

Knowledge Graph at Scale Innovative Training Network
22/01/2021

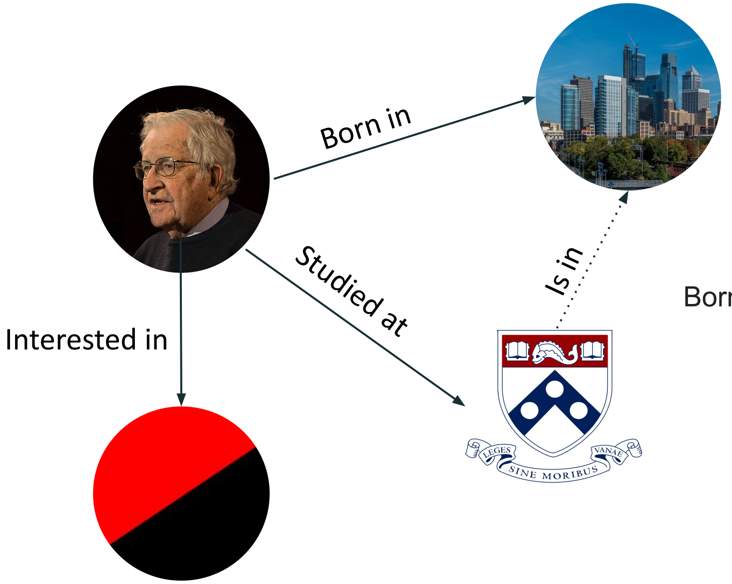
Semantic Knowledge graphs and Natural Language Processing

Knowledge Graphs include real-word information. Between the many types of relations, there can be **lexical** ones, such as those in a dictionary, or **semantical**, with relations such as **synonyms**.

I will focus on what I label as **semantic Knowledge Graphs**, although the term has been used in other contexts and there is no established definition.

A semantic relation is the relationship between **meanings**. Hence when talking about a semantic KG, I refer to a KG that entails information related to the meaning of the entities in it in the form of triplets.

Semantic relations

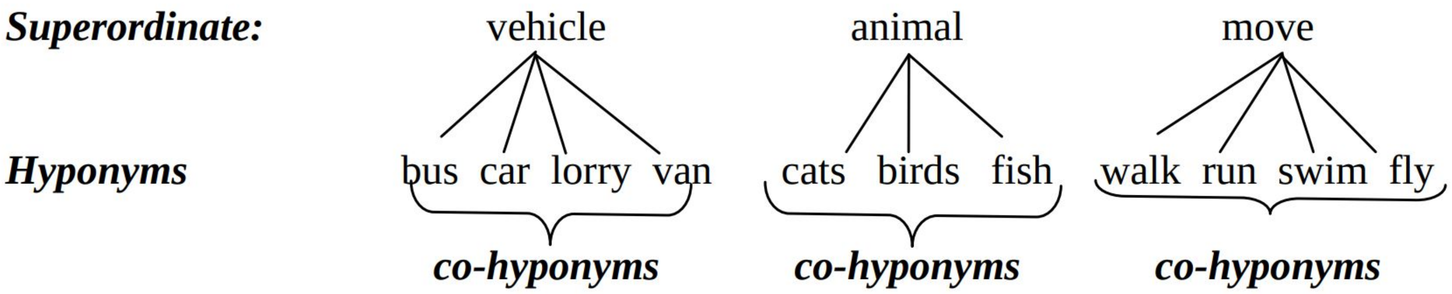


Born in   Interested in

Born to Jewish immigrants in **Philadelphia**, **Chomsky** developed an early interest in **anarchism** from alternative bookstores in New York City.

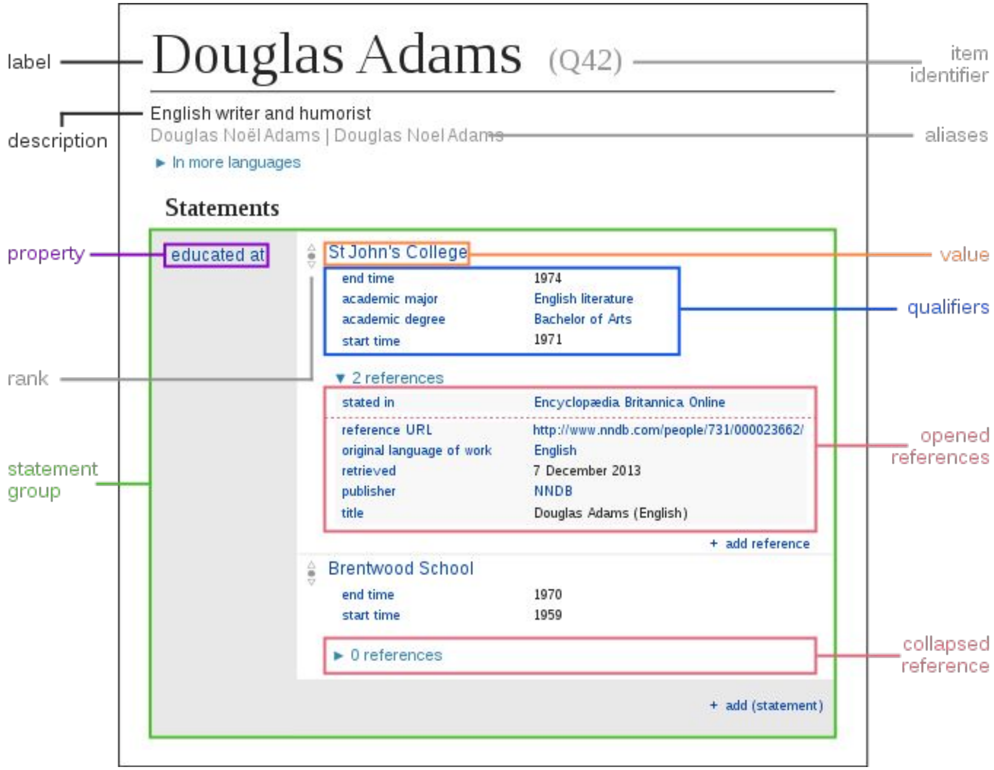
He studied at the **University of Pennsylvania**.

Studied at



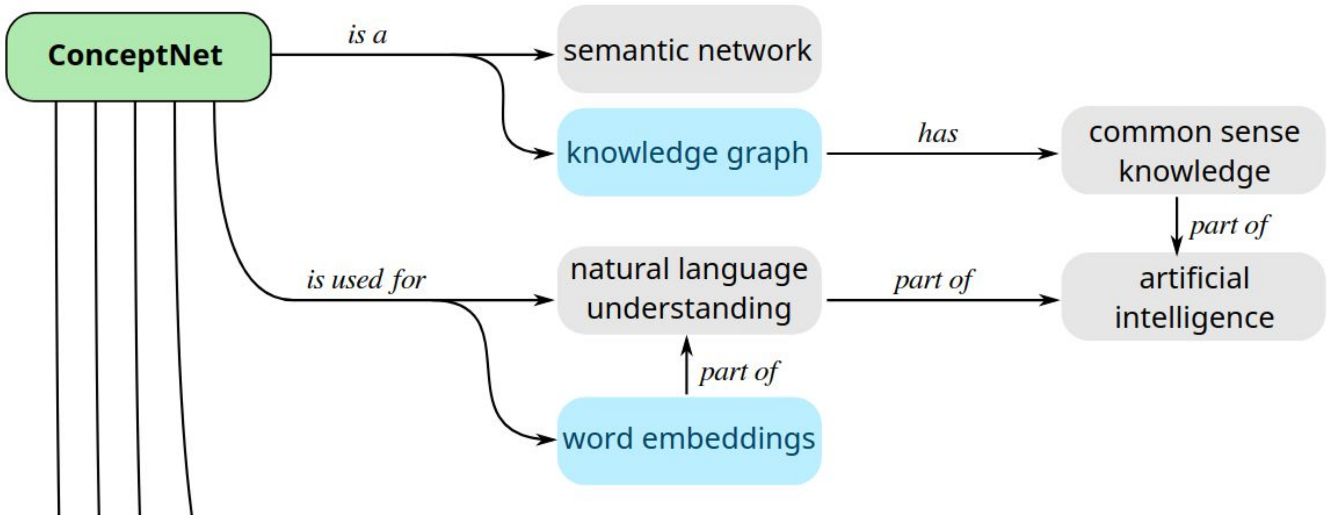
Wikidata

- http://wikidata.org
- Almost every language (by community)
- General knowledge
- JSON, XML, SQL, and RDF
- Continuously updated
- Can be freely downloaded
- Or online access through Wikibase-API, SPARQL (third party)



ConceptNet

- https://www.conceptnet.io/
- 10 core languages
- Common-sense
- TSV
- Last update Jul 3, 2019, version 5.7
- Can be freely downloaded
- Or online access through ConceptNet API



Not found
'covid' is not a node in ConceptNet.

DBpedia

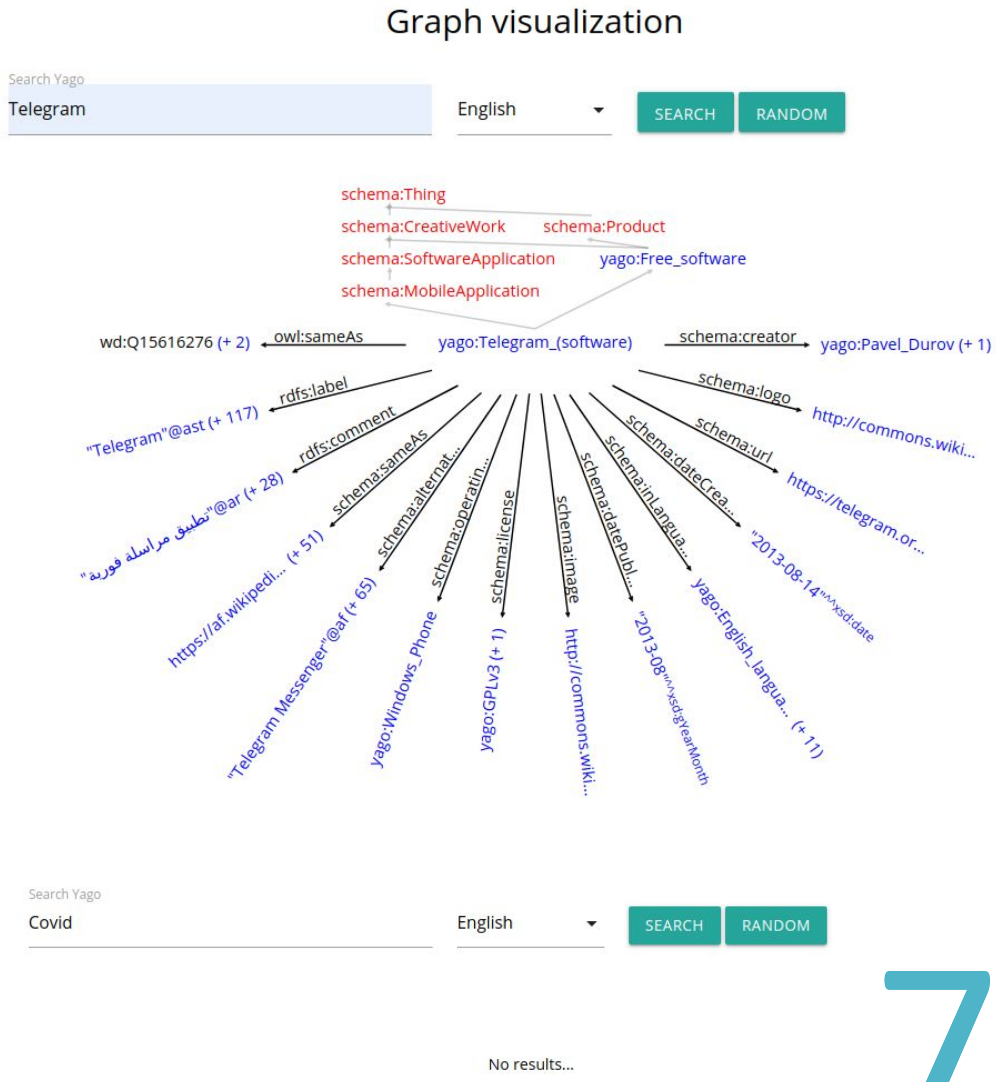
- http://dbpedia. org
- Mainly in English, can be extended to other languages.
- General knowledge
- RDF
- Monthly updated
- Can be freely downloaded
- Or online access through API
- With the dominance of Wikidata, DBpedia has lost a lot of ground, but it serves to connect different resources and includes Wikipedia abstracts.

About: Coronavirus disease 2019

An Entity of Type : disease, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

YAGO

- https://yago-knowledge.org/
- Almost every language (by community).
- General knowledge
- RDF, TSV
- Latest release March 2020, YAGO4
- Can be freely downloaded
- Or online access through API
- YAGO is currently a “cleaner” Wikidata, ensuring schema.org taxonomy.



YAGO

Table 2. Size statistics for YAGO 4 in the flavors Full, Wikipedia (W), and English Wikipedia (E), Wikidata and DBpedia (per DBpedia SPARQL server on 2020-03-04).

	Yago Full	Yago W	Yago E	Wikidata	DBpedia
Classes	10124	10124	10124	2.4M	484k
Classes from Wikidata	9883	9883	9883	2.4M	222
Individuals	67M	15M	5M	78M	5M
Labels (*rdfs:label*)	303M	137M	66M	371M	22M
Descriptions (*rdfs:comment*)	1399M	139M	50M	2146M	12M
Aliases (*schema:alternateName*)	68M	21M	14M	71M	0
rdf:type (without transitive closure)	70M	16M	5M	77M	114M
Facts	343M	48M	20M	974M	131M
Avg. # of facts per entity	5.1	3.2	4	12.5	26
sameAs to Wikidata	67M	15M	5M	N.A.	816k
sameAs to DBpedia	5M	5M	5M	0	N.A.
sameAs to Freebase	1M	1M	1M	1M	157k
sameAs to Wikipedia	43M	43M	26M	66M	13M
Fact annotations	2.5M	2.2M	1.7M	220M	0
Dump size	60GB	7GB	3GB	127GB	99GB

Thomas Pellissier Tanon, Gerhard Weikum, Fabian M. Suchanek: "YAGO 4: A Reason-able Knowledge Base"

Resource paper at the Extended Semantic Web Conference (ESWC), 2020

Others

- Babelnet
- Cyc (and OpenCyc)
- Freebase
- Google KG
- NELL…
- Even Wikipedia!

Did you say BabelNet?

But why are you telling us about all these KG Pere?

Dataset Creation

- Commonsense QA:
  - Crowdsourced **QA dataset** by using **ConceptNet** entities and relations.
  - 12,247 multiple-choice (5 options) questions (only in English).
  - SOTA (83.3% acc):

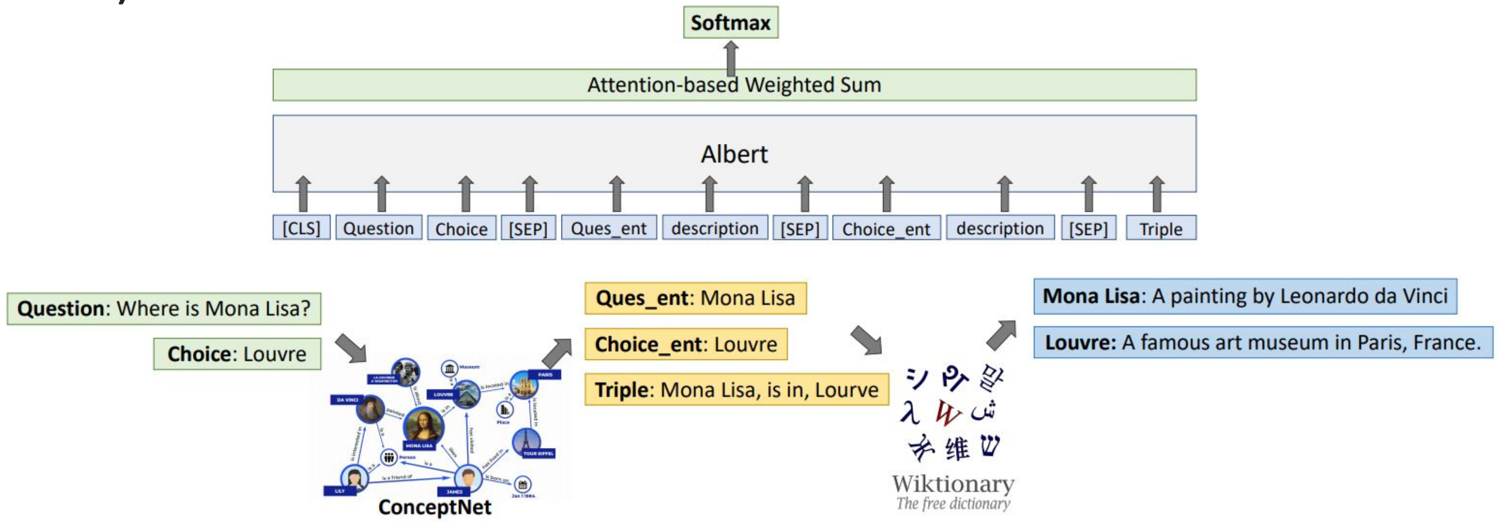


Figure 1: The input to Albert includes the question, choice, entity names, description text and triple. An attention-based weighted sum and a softmax layer processes the output from Albert to produce the prediction.

Xu, Yichong, et al. "Fusing Context Into Knowledge Graph for Commonsense Reasoning." arXiv preprint arXiv:2012.04808 (2020).

Dataset Creation

- T-REX:
  - Provide context to Wikidata relations using DBpedia abstracts.
  - Provide Relation and Entity extraction data.
  - Noisy (97.8% accuracy claimed):

	AllEnt	SPO	NoSub
Accuracy	0.88	0.96	0.98
Inter-Annotator	0.85	0.93	0.96

Table 4: Accuracy of each alignment methodology.

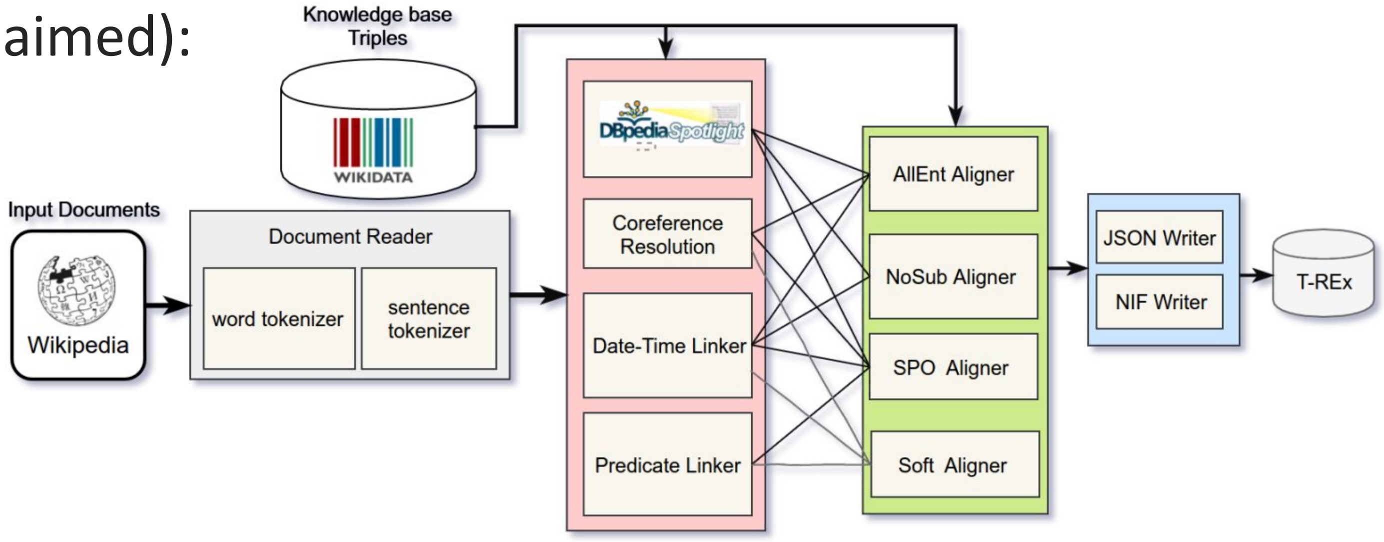


Figure 1: Overview of the alignment pipeline and its components

  - Potentially multilingual.

Hady ElSahar, et al. 2018. T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In LREC.

Language Models “Augmentation”

T-REX is used to pre-train K-Adapter model:

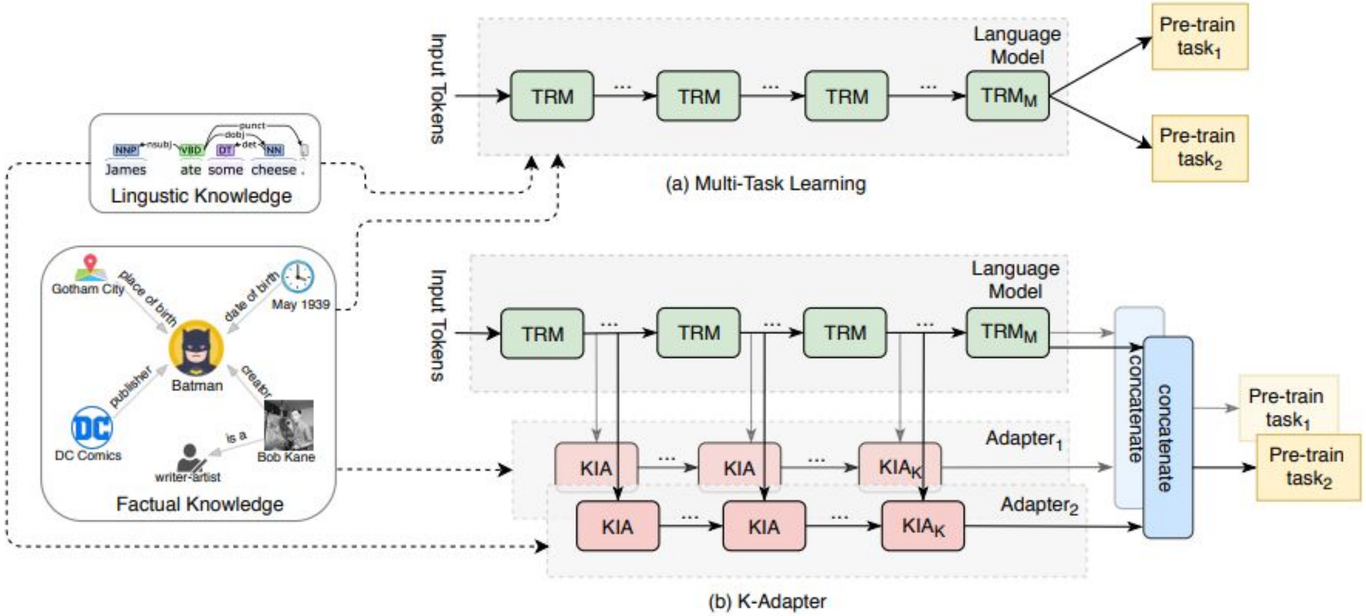


Figure 1: (a) Pre-trained language models inject multiple kinds of knowledge with multi-task learning. Model parameters need to be retrained when injecting new kinds of knowledge, which may result in the catastrophic forgetting (b) Our K-ADAPTER injects multiple kinds of knowledge by training adapters independently on different pre-train tasks, which supports continual knowledge infusion. When we inject new kinds of knowledge, the existing knowledge-specific adapters will not be affected. KIA represents the adapter layer and TRM represents the transformer layer, both of which are shown in Figure 2.

Ruize Wang et al. 2020. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. CoRR, cs.CL/2002.01808v3.

Language Models “Augmentation”

KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation

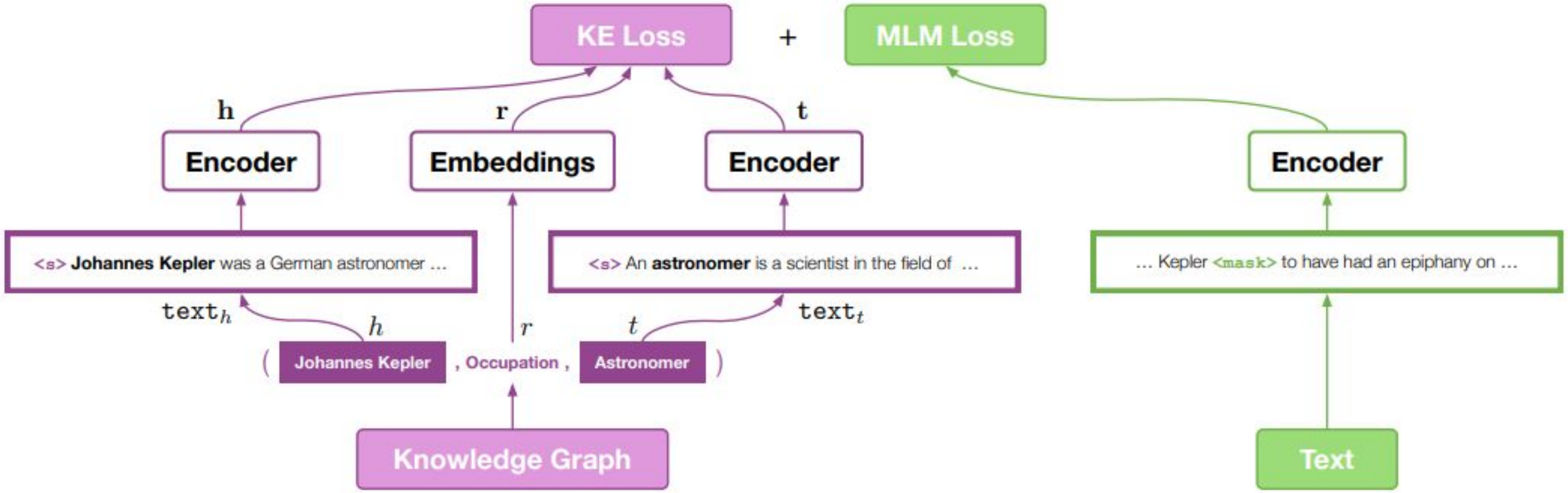


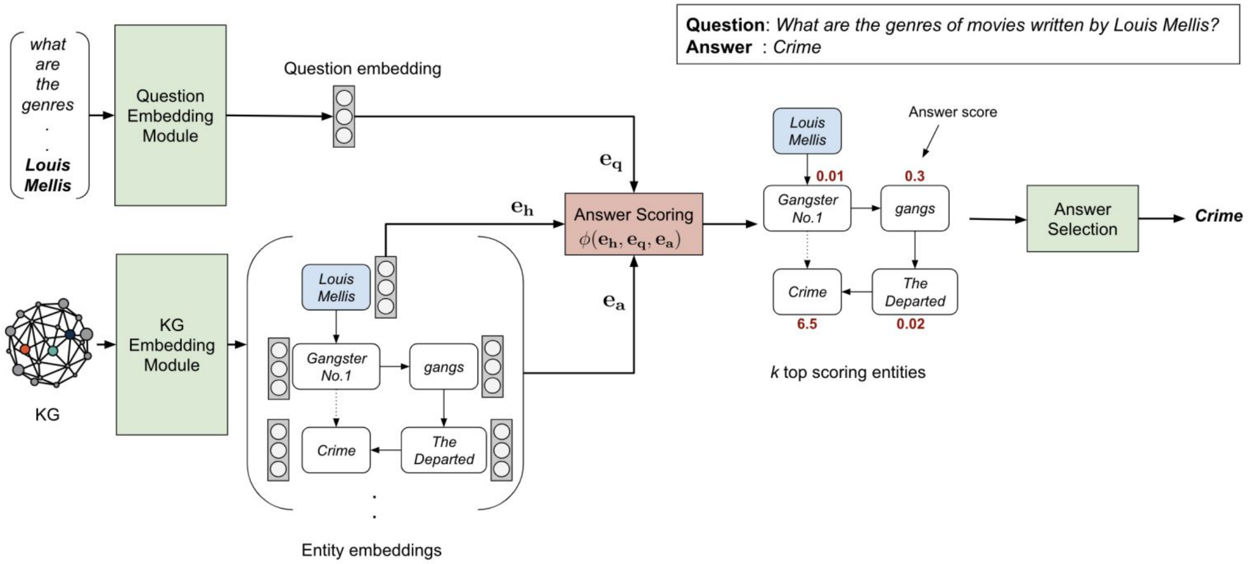
Figure 2: The KEPLER framework. We encode entity descriptions as entity embeddings and jointly train the knowledge embedding (KE) and masked language modeling (MLM) objectives on the same PLM.

Xiaozhi Wang et al. 2020. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. TACL, 2020.

Downstream Tasks

- Question Answering:

Apoorv Saxena et al. Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings. ACL, 2020.



François Gardères, Maryam Ziaeefard, Baptiste Abeloos and Freddy Lecue. ConceptBert: Concept-Aware Representation for Visual Question Answering. Findings of EMNLP, 2020.

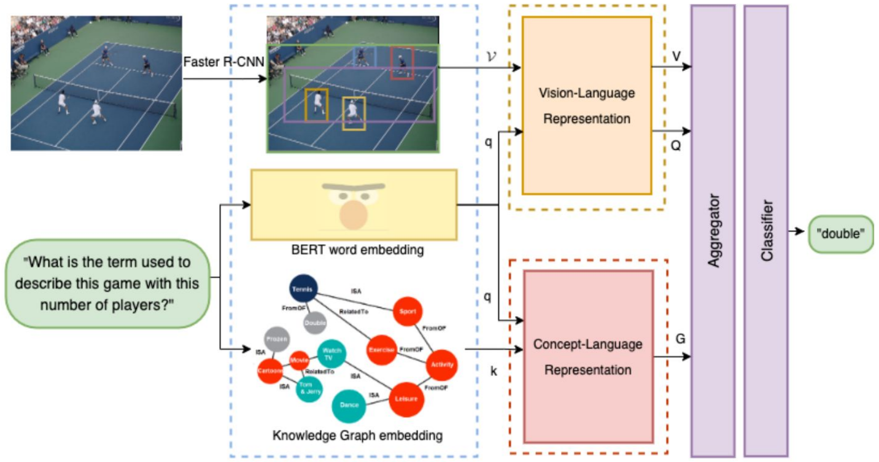


Figure 1: Model architecture of the proposed ConceptBert.

Downstream Tasks

- Recommendation systems:

Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. Multi-task feature learning for knowledge graph enhanced recommendation. In WWW'19, pages 2000–2010. ACM, 2019.

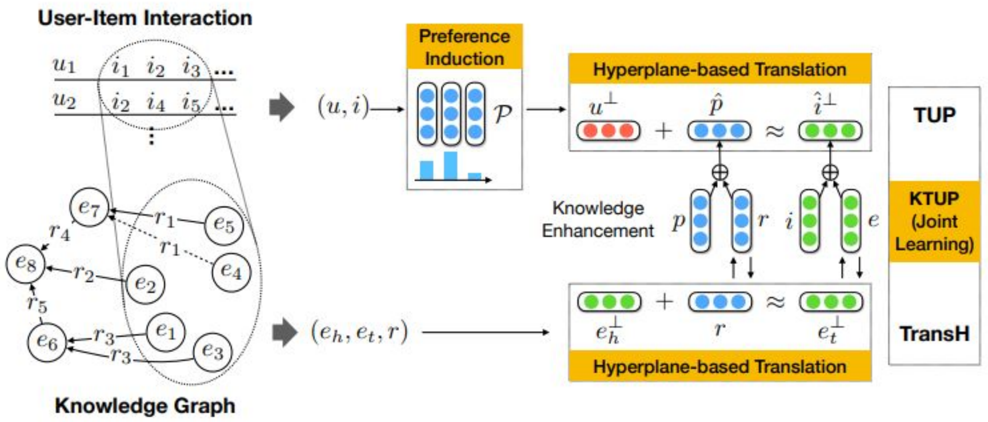


Figure 3: Framwork of KTUP. At the top is TUP for item recommendation including two components: preference induction and hyperplane-based translation. KTUP jointly learns TUP and TransH to enhance the item and preference modeling by transfering knowledge of entities as well as relations.

- Fake News detection:

Shantanu Chandra, Pushkar Mishra, Helen Yannakoudakis, Madhav Nimishakavi, Marzieh Saeidi and Ekaterina Shutova. 2020. Graph-based modeling of online communities for fake news detection. arXiv, 2008.06274

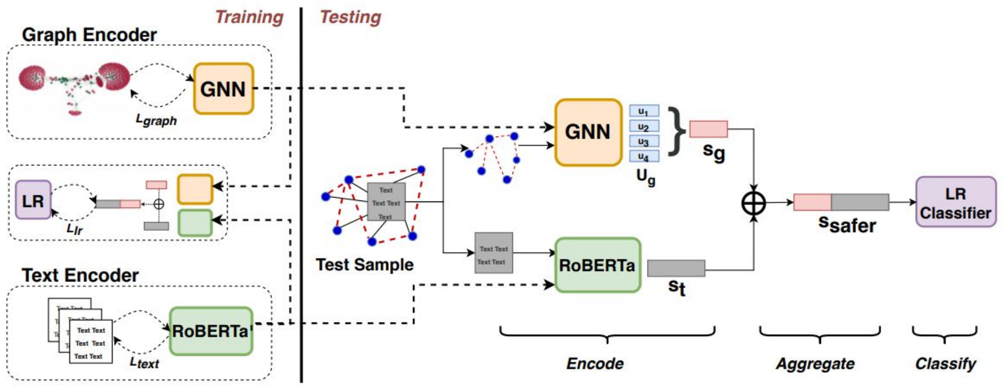


Figure 1: Visual representation of the proposed SAFER framework. Graph and text encoders are trained independently followed by training of a logistic regression (LR) classifier. During inference, the text of the article as well as information about its social network of users are encoded by the trained text and graph encoders respectively. Finally, the social-context and textual features of the article are concatenated for classification using the trained LR classifier.

And others

- Relation Extraction:

Amir DN Cohen, Shachar Rosenman, Yoav Goldberg. Relation Extraction as Two-way Span-Prediction. 2020, arXiv:2010.04829

- Question Generation
- Commonsense Reasoning
- Entity linking
- ...

Thank you for listening