Knowledge and NLP

Seung-won Hwang

SNU CSE

October 23, 2022

KB

Structured databases of knowledge usually containing

- Entities (nodes in a graph)
- Relations (edges between nodes)

How can we learn to create/expand knowledge bases with neural networks? (NLP for KB)

How can we learn from the information in knowledge bases to improve neural representations? (KB for NLP)

How can we use structured knowledge to answer questions (KB QA)

WordNet

WordNet is a large database of words including parts of speech, semantic relations

Nouns: is-a relation (hatchback-car), part-of (wheel-car), type-instance distinction

Verb relations: ordered by specificity (communicate $\rightarrow$ talk $\rightarrow$ whisper)

Adjective relations: antonymy (wet-dry)

`http://wordnetweb.princeton.edu/perl/webwn`

Cyc

A manually curated database attempting to encode all common sense knowledge, 30 years in the making

e.g., A birthday is the calendar date when an animal was born. Humans often celebrate the anniversary of birth with parties.

Factual knowledge: Casey Hart's birthday is August 2, 1986.

Automated

Extraction of structured data from Wikipedia

Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.943583°W

Carnegie Mellon University (**Carnegie Mellon** or **CMU** /ˈkɑːrnɪɡi ˈmɛlən/ or /kɑːrˈneɪɡi ˈmɛlən/) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

The university's 140-acre (57 ha) main campus is 3 miles (5 km) from Downtown Pittsburgh. Carnegie Mellon has seven colleges and independent schools: the College of Engineering, College of Fine Arts, Dietrich College of Humanities and Social Sciences, Mellon College of Science, Tepper School of Business, H. John Heinz III College of Information Systems and Public Policy, and the School of Computer Science. The university also has campuses in Qatar and Silicon Valley, with degree-granting programs in six continents.

Carnegie Mellon is ranked 25th in the United States and 77th in the world by *U.S. News & World Report*.[9] It is home to the world's first degree-granting Robotics and Drama programs,[10] as well as one of the first Computer Science departments.[11] The university was ranked 89th for R&D in 2015 having spent $242 million.[12]

Carnegie Mellon counts 13,650 students from 114 countries, over 100,000 living alumni, and over 5,000 faculty and staff. Past and present faculty and alumni include 20 Nobel Prize Laureates,[13] 12 Turing Award winners, 22 Members of the American Academy of Arts & Sciences,[14] 19 Fellows of the American Association for the Advancement of Science, 72 Members of the National Academies, 114 Emmy Award winners, 44 Tony Award laureates, and 7 Academy Award winners.[15]

Carnegie Mellon University

Former names	Carnegie Technical Schools (1900–1912) Carnegie Institute of Technology (1912–1967) Carnegie-Mellon University (1968–1988) [1] Carnegie Mellon University (1988–present)
Motto	"My heart is in the work" (Andrew Carnegie)
Type	Private university
Established	1900 by Andrew Carnegie

WikiData

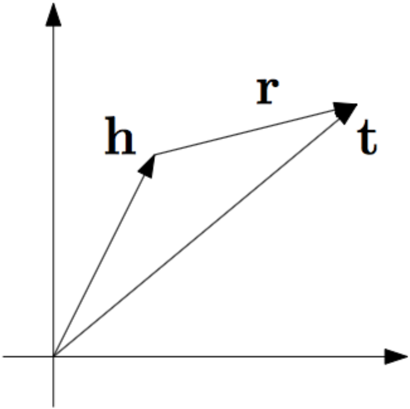
Similar to DBPedia, but crowd-sourced

Curated database of entities, linked, and extremely large scale, multilingual

Learning KG Embedding

Motivation: express triples as additive transformation (Recall: Word2Vec Analogies)

Method: minimize the distance of existing triples with a margin-based loss that minimizes $d(h+r,t) - d(h'+r,t')$



How do you evaluate?

Even w/ extremely large scale, knowledge bases are by nature incomplete

e.g. in FreeBase 71% of humans were missing "date of birth" (West et al. 2014)

Can we perform "relation extraction" to extract information for knowledge bases?

Distant Supervision: Snowball

Given an entity-relation-entity triple, extract all text that matches this and use it to train

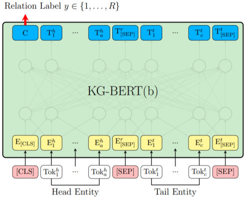
works-for: (BillG) works for (MS) vs (BillG) is a shareholder of (MS).

Creates a large corpus of (noisily) labeled text to train a system

Alternative: KG-BERT

Fine-tuning BERT to predict likely relation between (h,t)

Fine-tuning BERT to predict plausibility of triple (h,r,t)



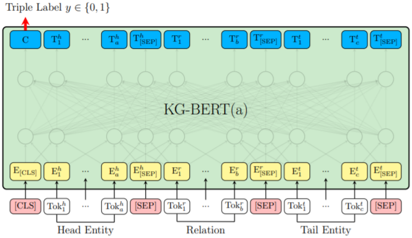


Figure 1: Illustrations of fine-tuning KG-BERT for predicting the plausibility of a triple.

Retrofitting Embedding using Knowledge

Similar to joint learning, but done through post-hoc transformation of embeddings

Advantage of being usable with any pre-trained embeddings

Double objective of making transformed embeddings close to neighbors, and close to original embedding

Can also force antonyms away from each-other (Mrksic et al. 2016)

Injecting Knowledge to LM

Provide LMs with topical knowledge in the form of copiable graphs

Each (Wiki) text is given relevant KB taken from Wikidata

Examine all possible decoding “paths” and maximize the marginal probability

During pretraining, finetuning, decoding...

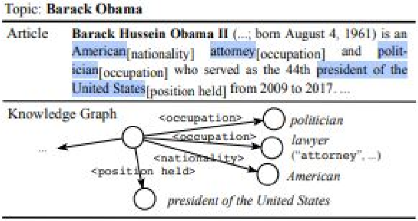


Figure 1: Overview of our task of language modeling conditioned on structured knowledge. For a given topic, we want to learn an LM that leverages the knowledge graph through relations when modeling the text.

LM as a KB

Traditional QA/MRC models usually refer to external resources to answer questions, e.g., Wikipedia articles or KGs.

LMs pre-trained on a large text corpus already capture those knowledge

LAMA benchmark "prompts" for 41 relations: "[X] was born in [Y]."

Accuracy: ELMo 7.1%, Transformer-XL 18.3%, BERT-base 31.1%

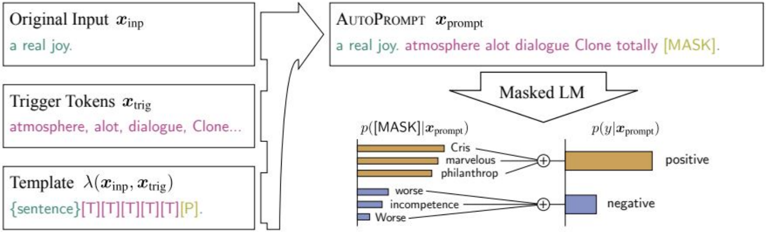
Prompt Engineering

Query LMs with different prompts might lead to different predictions.

Ensemble multiple mined/paraphrased prompts further increase the accuracy: 31.1% $\rightarrow$ 39.6% (room for research)

AutoPrompt (Shin et al. 2020): Search tokens in the prompts (i.e., trigger tokens [T]) guided by gradients that maximize the probability of correct answers.

Further increase the accuracy: 39.6% $\rightarrow$ 43.3%



Prompt Engineering: P-tuning (Liu et al. 2021)

Optimizing embeddings (continuous) is easier than searching tokens (discrete).

Further increase the accuracy: 43.3% $\rightarrow$ 48.3%