Language Models as Knowledge Bases?

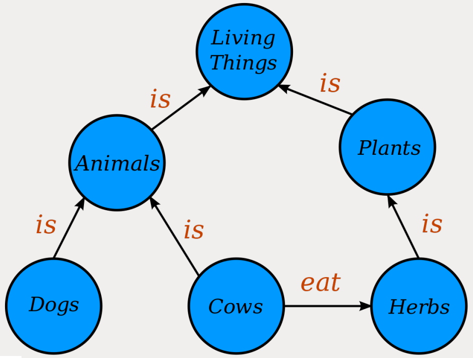
Fabio Petroni, Tim Rockt¨aschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu,
Alexander H. Miller, Sebastian Riedel

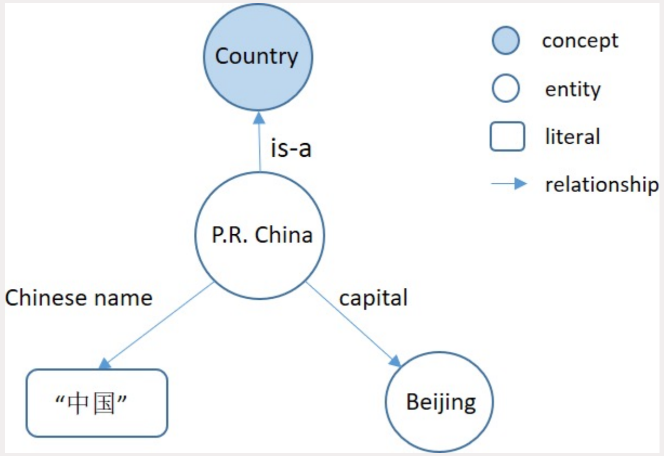
Jon Vincent Medenilla
ECE 594
March 24, 2022

Knowledge Bases

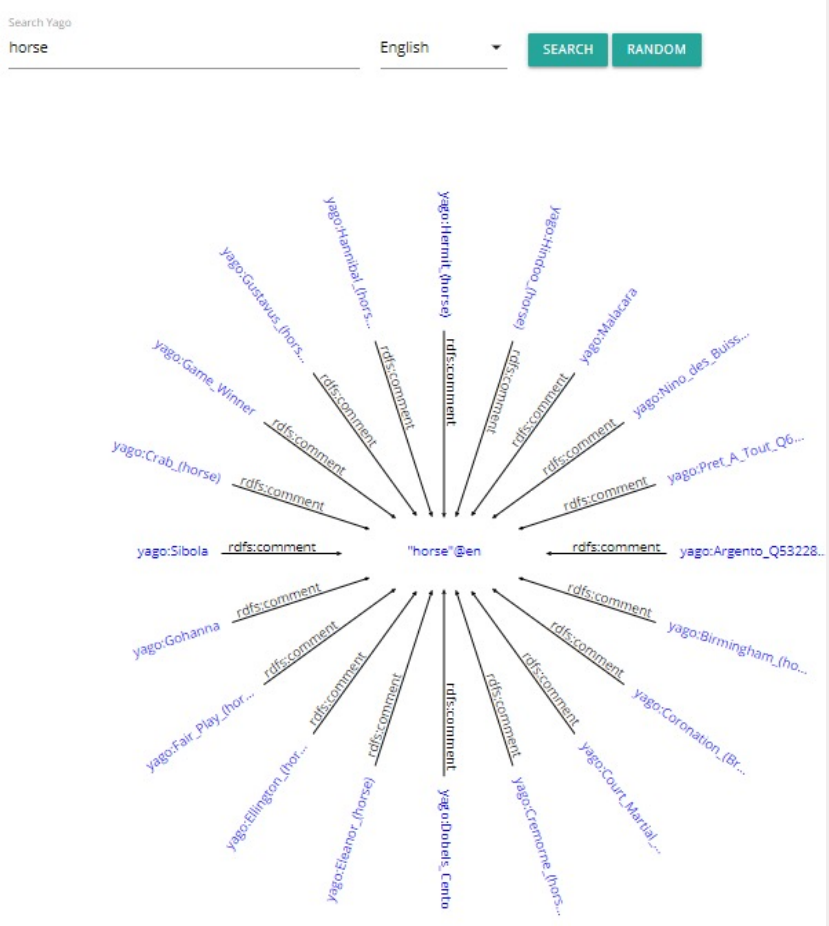
- A knowledge base allows for rapid search, retrieval, and reuse
- Stores information as answers to questions or solutions to problems
- Can be fed into a language model

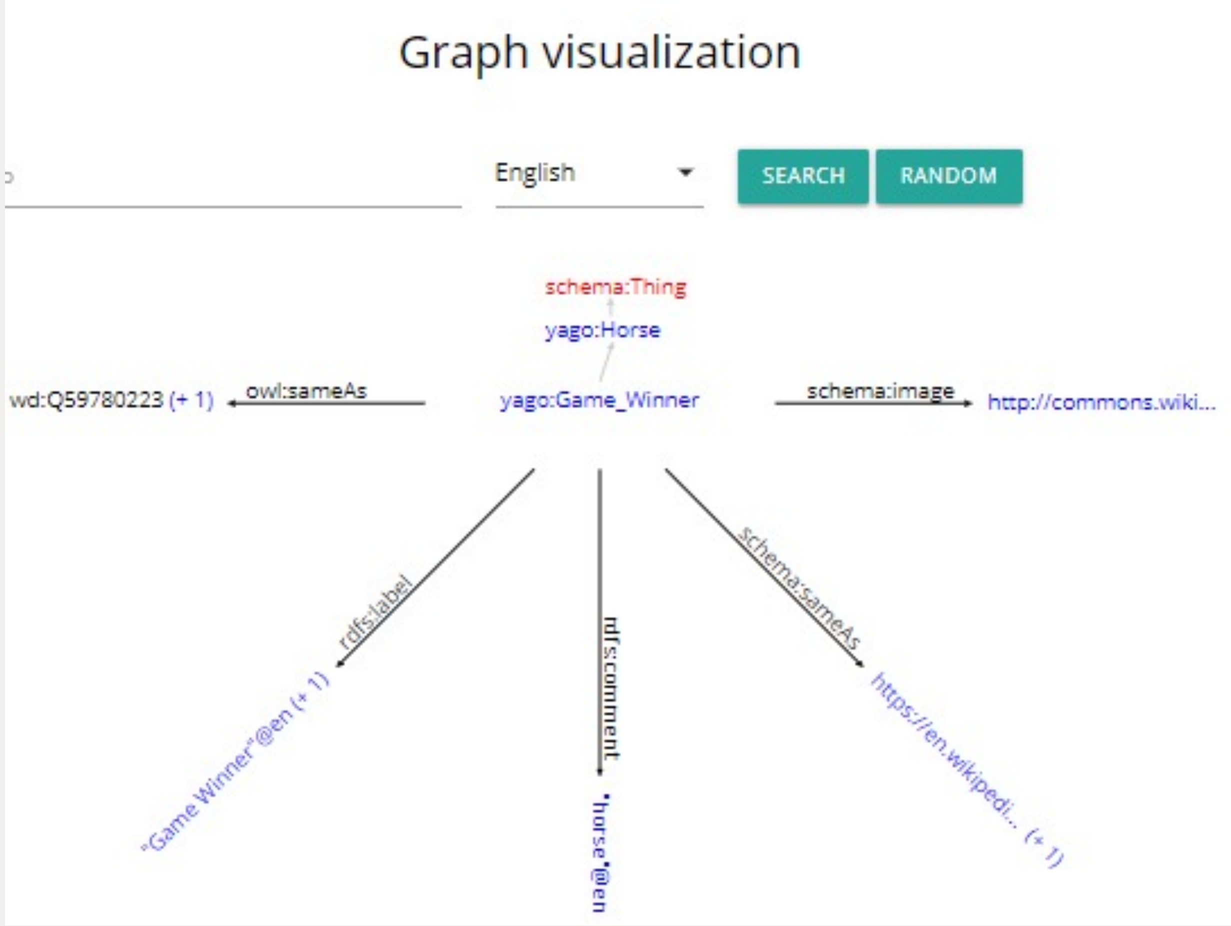
WIKIPEDIA
The Free Encyclopedia





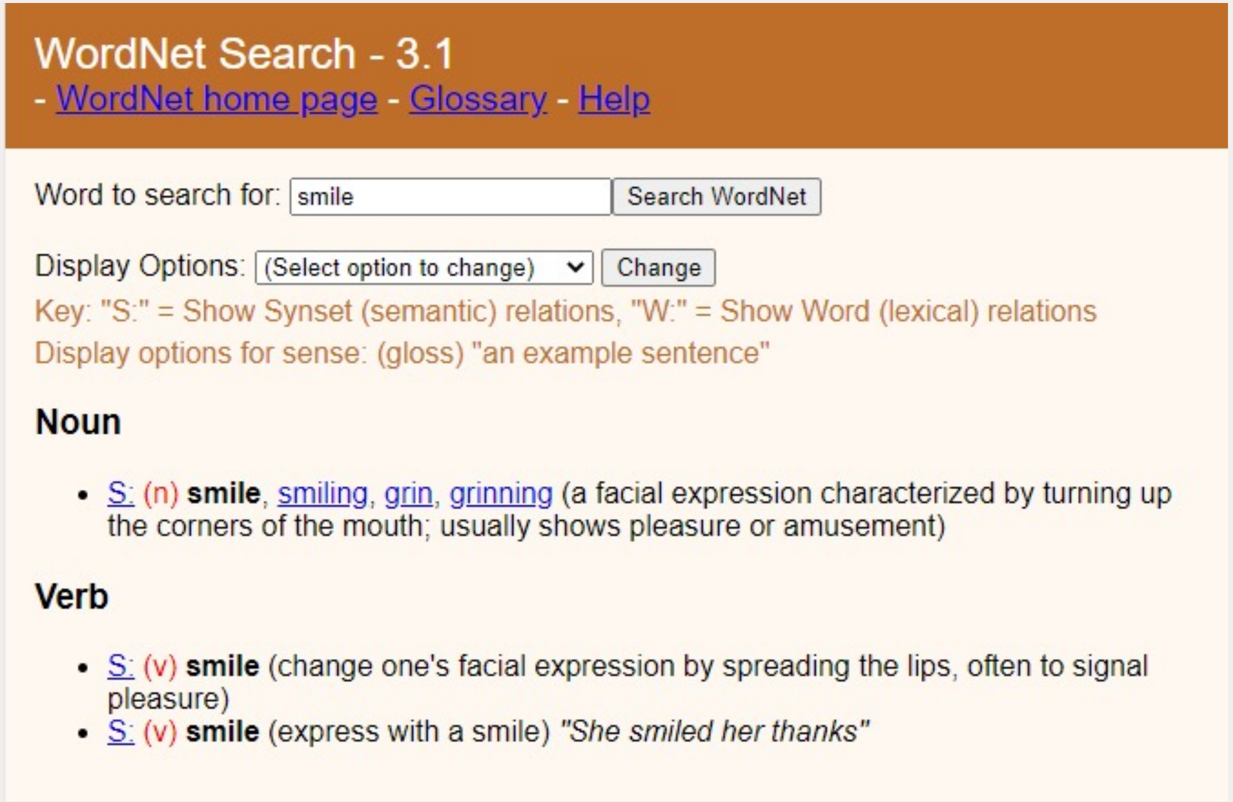
Examples of Knowledge bases





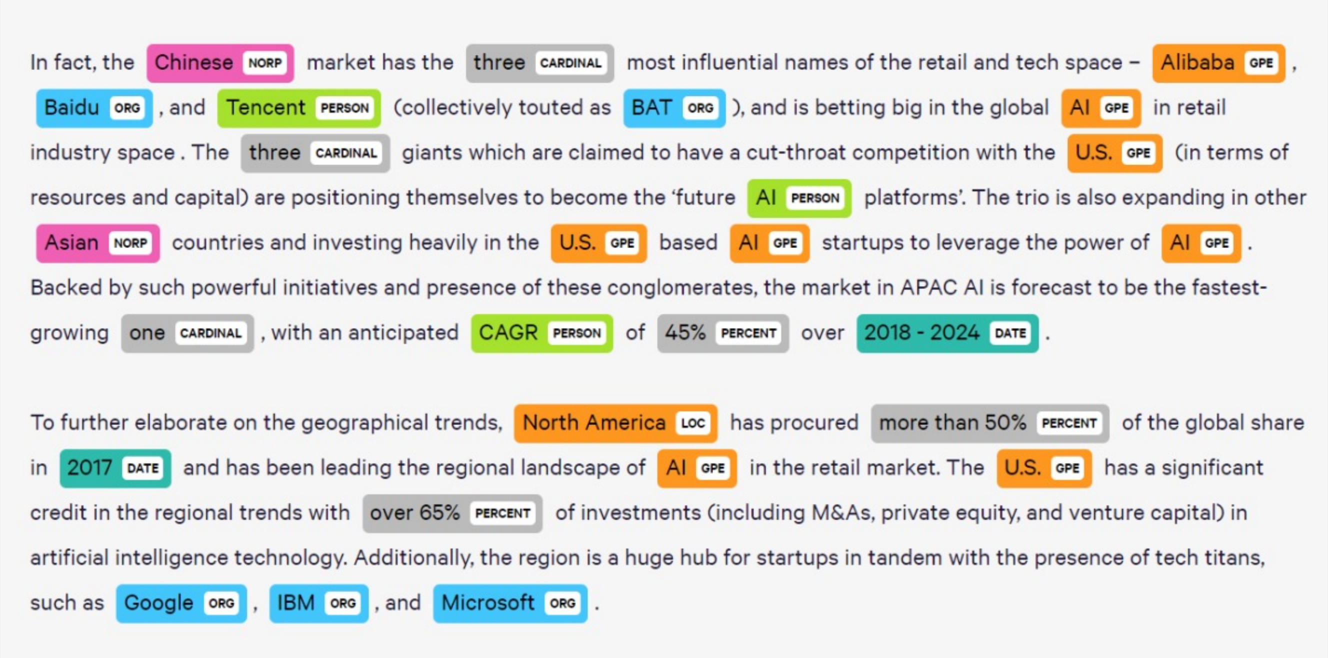
- Concepts like *classes* and *individuals* are modeled as nodes
- *Relations* as edges of graphs

- *Classes* – concepts like documents, events, or subjects
- *Individuals – instances of a class or an object*
- *Relations – capture relationships between classes and individuals*
    - *is-type-of, is-instance-of, and has-attribute*



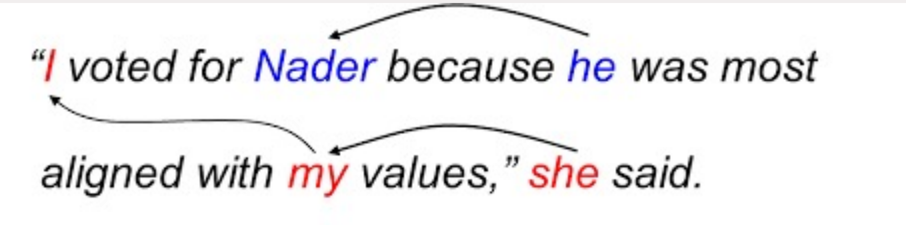
How knowledge bases are used in NLP models:

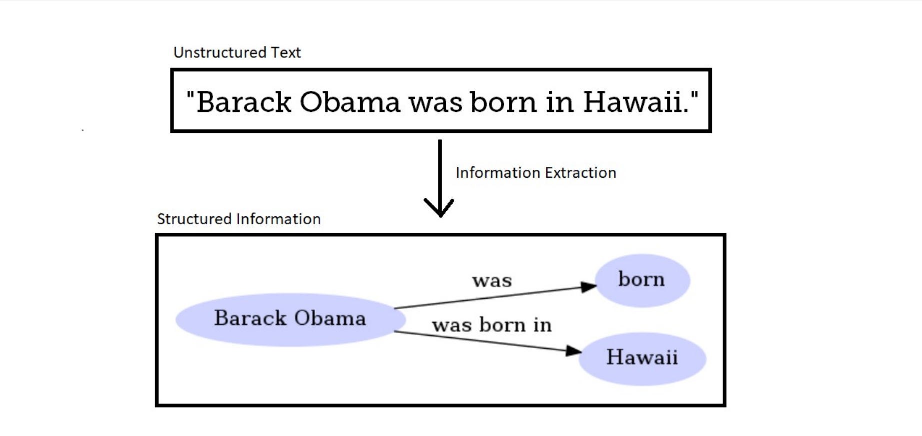
- Entity extraction – replace or augment entity occurrences in text



In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

- Coreference resolution: 

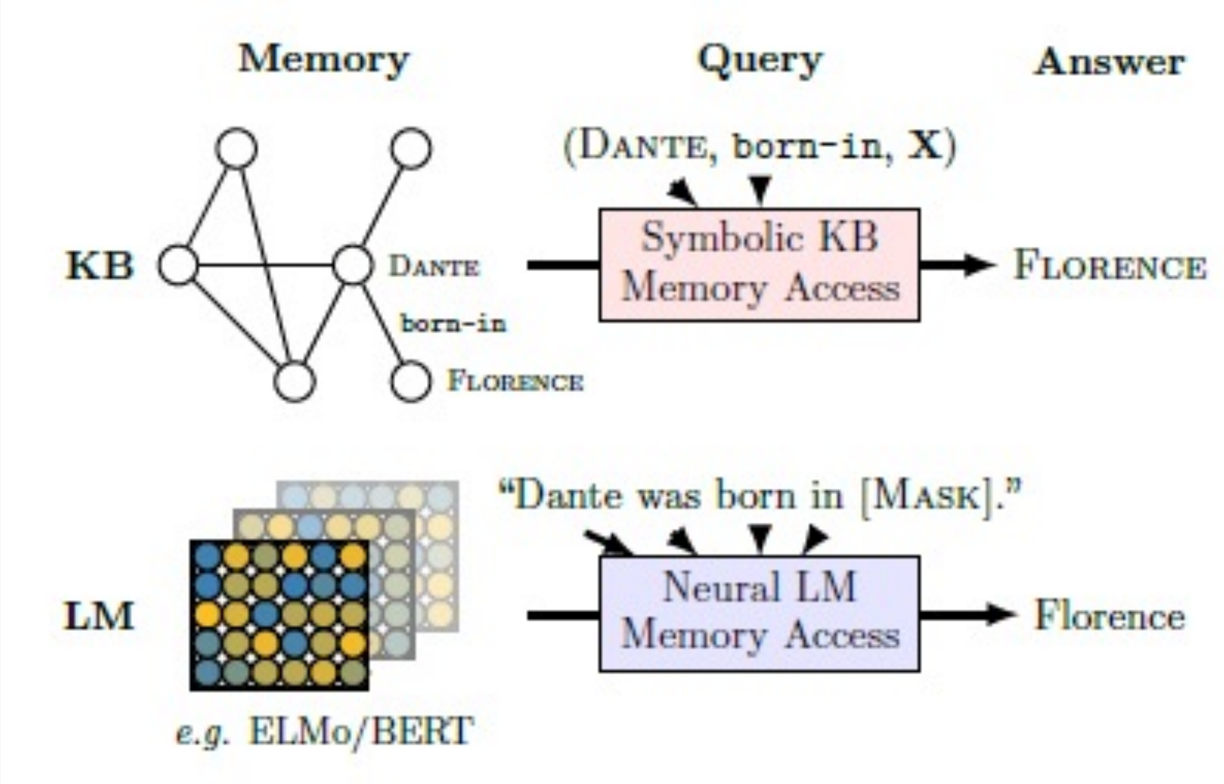
- Entity Linking: 

Proposed Solution:

- Ask the model to fill in masked tokens
- "Alex was born in [MASK]"
- Pre-trained high-capacity models such as ELMo and BERT store vast amounts of linguistic knowledge useful for downstream tasks

The Pros:

- Requires no schema engineering
- No need for human annotations
- Supports a more diverse/open set of inquiries



Questions this paper addresses:

- How much relational knowledge do they store?
- How does this differ for different types of knowledge such as facts about entities, common sense, and general question answering?
- How does their performance without fine-tuning compare to symbolic knowledge bases automatically extracted from text?

LAMA (Language Model Analysis) Probe

- consisting of a set of knowledge sources, each comprised of a set of facts (subject, relation, object)
- Success depends on predicting masked objects such as “Dante was born in ___”
- tested for a variety of types of knowledge: relations between entities stored in Wikidata, common sense relations between concepts from ConceptNet, and knowledge necessary to answer natural language questions in SQuAD.
- Key Steps:
  - Query each model for a missing token
  - Evaluate each model based on how highly they rank the ground truth token against every word in a fixed candidate vocabulary

Knowledge Sources Used:

- Google-RE – contains ~60K facts manually extracted from Wikipedia
  - Only utilized 3 relations: “place of birth”, “date of birth” and “place of death”
  - manually defined a template for each considered relation, e.g., “[Adam] was born in [Illinois]” for “place of birth”
- T-Rex – is a subset of Wikidata triples
  - Much larger than Google-RE with broader relations
  - Facts were automatically aligned to Wikipedia (can be noisy)
- SQuAD
  - Question-answering dataset
  - a subset of 305 context-insensitive questions with single token answers
  - rewriting “Who developed the theory of relativity?” as “The theory of relativity was developed by __”.
- ConceptNet
  - Multilingual knowledge base, initially built on top of Open Mind Common Sense sentences
  - English parts that have single-token objects covering 16 relations

Language Models evaluated:

- **Unidirectional Language Models:**
  - Given a string of input tokens w = [w1,w2,…,wn], assign probability p(w)

$$p(\mathbf{w}) = \prod_t p(w_t \mid w_{t-1}, \ldots, w_1).$$

  - Using neural language models:

$$p(w_t \mid w_{t-1}, \ldots, w_1) = \mathrm{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b})$$

    - ht = output vector at position t
    - W = learned parameter matrix

Fairseq-fconv

- Multiple layers of gated convolutions
- Trained on the WikiText-103 corpus

Transformer-XL

- Large-scale LM based on the Transformer
- Takes into account a longer history
- Used relative instead of absolute positional encoding
- Trained on the WikiText-103 corpus

- **Bidirectional Language Models:**
  - ELMO:
    - Given a string of input tokens w = [w1,w2,…,wn] and position 1 <= i <= N, estimate

$$p(w_i) = p(w_i \mid w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_N)$$

    - ELMo: Forward and backward LSTM, resulting in $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$
      - Trained on the Google Billion Word dataset
    - ELMo 5.5B
      - Trained on English Wikipedia and monolingual news crawl from WMT 2008-2012

- BERT:
  - Transformer architecture
  - Trained on the BookCorpus and English Wikipedia
  - **language modelling** (15% of tokens were masked and BERT was trained to predict them from context) and **next sentence prediction** (if a chosen next sentence was probable or not given the first sentence)
  - BERT-base (12 encoders with 12 bidirectional self-attention heads)
  - BERT-large (24 encoders with 16 bidirectional self-attention heads)

Methodology

$$p(\mathbf{w}) = \prod_t p(w_t \mid w_{t-1}, \ldots, w_1).$$

$$p(w_t \mid w_{t-1}, \ldots, w_1) = \mathrm{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b})$$

W = ['compare', 'language', 'models', 'to', 'canonical', 'ways']

$$p(\text{'ways'}) = \prod p(\text{'ways'} \mid \text{'canonical'}, \ldots, \text{'compare'}]$$
$$= \mathrm{softmax}(W h_{ways} + b)$$

Unidirectional:

h_{t-1} = output vector at 'canonical'

Bidirectional:

ELMo : ($t = 2 \Rightarrow$ 'models')

$\vec{h}_{t-1}$ = output vector at 'language'

$\vec{h}_{t+1}$ = output vector at 'to'

- ELMo: averaged forward and backward probabilities from the corresponding softmax layers
- BERT: masked the token at position *t*, fed output to vector corresponding to masked token (ht) into softmax layer

Baselines

- **Freq**
  - subject and relation pair, this baseline ranks words based on how frequently they appear as objects for the given relation in the test data
- **Relation Extraction (RE)**
  - extracts relation triples from a given sentence using an LSTM-based encoder and an attention mechanism
  - constructs a knowledge graph of triples
  - At test time, they queried this graph by finding the subject entity and then rank all objects in in the correct relation based on the confidence scores by the RE
- **DrQA**
  - a popular system for open-domain question answering
  - Two-step pipeline:
    - First, a TF/IDF information retrieval step is used to find relevant articles from a large store of documents (e.g. Wikipedia)
    - Secondly, on the retrieved top k articles, a neural reading comprehension model then extracts answers

Metrics

- Rank-based metrics
- For multiple valid objects for Subject-Relation pair, removed all other valid objects from the candidates when ranking at test time other than the ones they were testing
- Mean precision at k (P@k)
    - For a given fact, this value is 1 if the object is ranked among the top k results, 0 otherwise

Considerations in LAMA

- Manually Define Templates:
  - Manually defined a template that queries for the object slot for each relation
  - For example, for a relation ID “works-for”, and the user asks for “is-working-for”, the accuracy would be 0
  - e.g., “[S] was born in [O]” for “place of birth”.
- Single Token
- Object Slots
  - Only in triples (subject, relation, object)
- Intersection of Vocabularies
  - ELMO uses ~800K tokens compared to BERT’s ~30K tokens
  - Intersection of 2 vocabularies yielding ~21K tokens

Results

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	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.

Discussion of Results

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	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**

- From earlier example, “Adam was born in [MASK]”
- BERT-Large (last column) outperformed all models by a substantial margin
- REn – naïve entity linking, i.e. exact string matching
- REo – uses an oracle for entity-linking, i.e. any given (s, r, o) in sentence x, if any other (s’, r, o’) has been extracted in the same sentence, s will be linked to s’, and o to o’

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
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

- More facts and relations than Google-RE
- BERT-Large performed better on 1-to-1 relations, i.e. “capital-of”
- N-1: Multiple valid subjects-relations-> 1 correct object
- N-M relations: multiple objects for a subject-relation pair. i.e. “Brian owns [car, laptop, iPhone,etc]”

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
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

- BERT-Large achieved best performance for ConceptNet
  - Able to retrieve commonsense knowledge at a similar level to factual knowledge

	Relation	Query	Answer	Generation
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]**, fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7], die [-1.7], **laugh [-2.0]**, vomit [-2.6], scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2], cancer [-2.0], **infection [-2.6]**, plague [-3.3], fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8], nests [-3.1], **feathers [-3.2]**, died [-3.7], eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4], human [-3.3], **alive [-3.3]**, young [-3.6], free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], **fish [-2.8]**, recreation [-3.1]

Corpus	Relation	Statistics #Facts	#Rel	Baselines Freq	DrQA	KB RE_n	RE_o	LM Fs	Txl	Eb	E5B	Bb	Bl
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

- Open domain cloze-style (fill in the blanks)
- Huge performance gap between BERT-Large and supervised DrQA
- Note: BERT and ELMo were both unsupervised and not fine-tuned for this task
- In terms of P@10 (Top-10 best answers), gap is remarkably small (57.1 for Bl and 63.5 for DrQA)

Conclusions

- For an unsupervised, not fine-tuned, pre-trained model BERT-Large, it is possible to recall knowledge better than its competitors, comparable to that of a knowledge base extracted with an off-the-shelf relation extractor and an oracle-based entity linker from a corpus known to express the relevant knowledge
- factual knowledge can be recovered surprisingly well from pretrained language models, however, for some relations (particularly N-to-M relations) performance is very poor
- This paper focused on the as-is knowledge inherent in the weights of existing pre-trained models which are often used as starting points for most research works
- Language models trained on ever-growing corpora might become a viable alternative to traditional knowledge bases extracted from text in the future

Limitations

- Only used Single-Token objects as prediction targets
- Chose only query objects in triples
- Still spent time manually defining templates for each relation

Questions/Thoughts?