Differentiable Reasoning Over a Virtual Knowledge Base

ICLR 2020

Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, William W. Cohen

Presented by Michal P

Open Domain Multi-Hop QA

Query: When was the Grateful Dead and Bob Dylan album released?

Corpus: Bob Dylan (born Robert Allen Zimmerman; May 24, 1941) is an American singer-songwriter, author, and visual artist who has been a major figure in popular culture for more than 50 years. Much of his most celebrated work dates from the 1960s, when songs such as "Blowin' in the Wind" (1963) and "The Times They Are a-Changin'" (1964) became anthems for the civil rights and anti-war movements. His lyrics during this period incorporated a range of political, social, philosophical, and literary influences, defied pop music conventions and appealed to the burgeoning counterculture. Following his self-titled debut album in 1962, which mainly comprised traditional folk songs, Dylan made his breakthrough as a songwriter with the release of The Freewheelin' Bob Dylan the following year. The album featured "Blowin' in the Wind" and the thematically complex "A Hard Rain's a-Gonna Fall". For many of these songs, he adapted the tunes and phraseology of older folk songs. He went on to release the politically charged The Times They Are a-Changin' and the more lyrically abstract and introspective Another Side of Bob Dylan in 1964. In 1965 and 1966, Dylan drew controversy when he adopted electrically amplified rock instrumentation, and in the space of 15 months recorded three of the most important and influential rock albums of the 1960s: Bringing It All Back Home (1965), Highway 61 Revisited (1965) and Blonde on Blonde (1966). Commenting on the six-minute single "Like a Rolling Stone" (1965), Rolling Stone wrote: "No other pop song has so thoroughly challenged and transformed the commercial laws and artistic conventions of its time, for all time."[3] ....

Answer: 1989

$$Y = X.\text{follow}(R) = \{x' : \exists x \in X \ \text{ s.t. } \ R(x, x') \ \text{ holds}\}$$

DrKit

Current solutions (entity/relation based): Freebase, WikiData

Alternative approach: treat corpus as virtual KB (vKB) (recall images in MAC networks)

Goals:

- End to end differentiable
- Efficient
- Multi-hop

Approach: vKB, TFIDF + Max Inner Product Search (MIPS)

DrKit: Multi-hop QA (Level 1)

1. Find entities z in query q
2. Expand z to all mentions m in vKB
3. Filter m relevant to q
4. Aggregate m and obtain new entities z'
5. Repeat!

1. Query: When was the Grateful Dead and Bob Dylan album released?
2. Expand: In 1986 and 1987, Dylan toured with Tom Petty and the Heartbreakers, sharing vocals with Petty on several songs each night. Dylan also toured with the Grateful Dead in 1987, resulting in a live album Dylan & The Dead. This received negative reviews; AllMusic said it was "Quite possibly the worst album by either Bob Dylan or the Grateful Dead."[222] Dylan then initiated what came to be called the Never Ending Tour on June 7, 1988, performing with a back-up band featuring guitarist G. E. Smith. Dylan would continue to tour with a small, changing band for the next 30 years........
3. Filter (top-K)...
4. Aggregate (new entities):
   a. Dylan & The Dead
   b. Never Ending Tour
   c. ...
5. Repeat with new entities

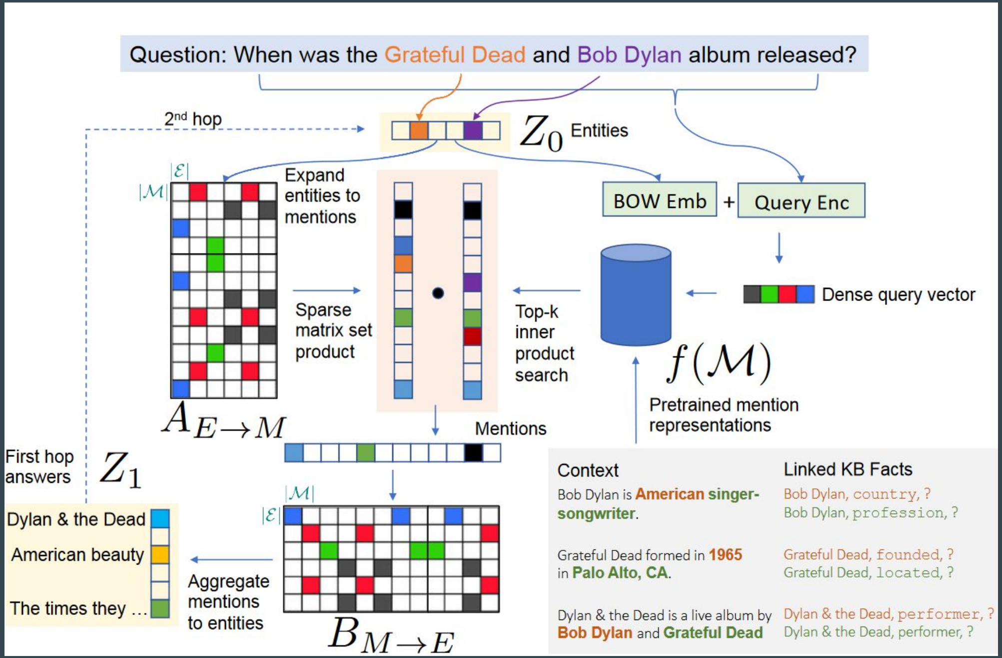
DrKit: Multi-hop QA (Level 2)

$$\Pr(z_t|q) = \sum_{z_{t-1} \in \mathcal{E}} \Pr(z_t|q, z_{t-1}) \Pr(z_{t-1}|q)$$

$$\Pr(z_t|q) = \sum_{m \in \mathcal{M}} \sum_{z_{t-1} \in \mathcal{E}} \Pr(z_t|m) \Pr(m|q, z_{t-1}) \Pr(z_{t-1}|q)$$

$$\Pr(m|q, z_{t-1}) \propto \underbrace{\mathbb{1}\{G(z_{t-1}) \cdot F(m) > \epsilon\}}_{\text{expansion to co-occurring mentions}} \times \underbrace{s_t(m, z_{t-1}, q)}_{\text{relevance filtering}}$$

DrKit: Multi-hop QA (Level 3)



DrKit: Multi-hop QA (Level 3)

$$Z_t = \text{softmax}\left(\left[Z_{t-1}^T A_{E\to M} \odot \mathbb{T}_K(s_t(m, z_{t-1}, q))\right] B_{M\to E}\right)$$

$A_{E\text{->}M}$ (matrix): pre-compute TFIDF for all entities + mentions, sparse

$Z_{t\text{-}1}$ (vector): probabilities of z from previous iteration, sparse

T_K (vector): top-K relevant mentions, sparse

$B_{M\text{->}E}$ (matrix): entity to which mention points, sparse

$$Z_t = \mathrm{softmax}\left(\left[Z_{t-1}^T A_{E\to M} \odot \mathbb{T}_K(s_t(m, z_{t-1}, q))\right] B_{M\to E}\right)$$

Efficient Implementation $A_{E\text{->}M}$

TFIDF vectors F(m) and $G(z_{t-1})$ constructed from unigrams and bigrams

Hashed to vocab of 16M buckets

Limit number of retrieved mentions per entity to μ

$$Z_t = \text{softmax}\left(\left[Z_{t-1}^T A_{E\to M} \odot \mathbb{T}_K(s_t(m, z_{t-1}, q))\right] B_{M\to E}\right)$$

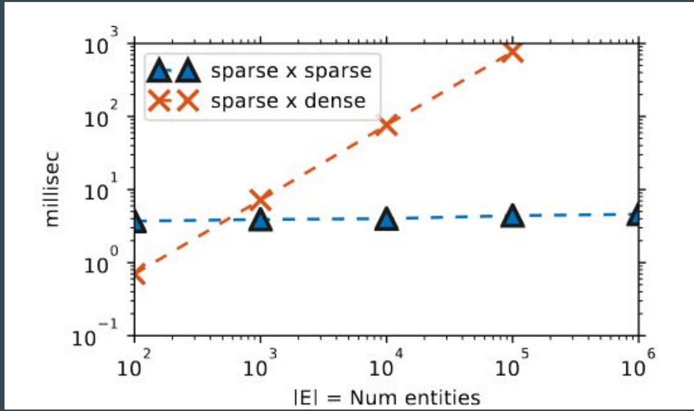
Efficient Implementation $Z_{t-1}A_{E\text{->}M}$

Z_{t-1}: K nonzero, $A_{M\text{->}E}$: has μ nonzero -> Ω(Kμ)

Note: independent of size of matrix (number of entities/relations)

Solution: Two ragged list-of-lists (K sparse vectors with μ nonzero elements each)

$$\begin{aligned}&[[idx_{0,0},\ idx_{0,1},\ idx_{0,2}],\\ &\ [idx_{1,0}],\\ &\ [idx_{2,0},\ idx_{2,1}]]\\ &[[val_{0,0},\ val_{0,1},\ val_{0,2}],\\ &\ [val_{1,0}],\\ &\ [val_{2,0}, val_{2,1}]]\end{aligned}$$



$$Z_t = \text{softmax}\left(\left[Z_{t-1}^T A_{E\to M} \odot \mathbb{T}_K(s_t(m, z_{t-1}, q))\right] B_{M\to E}\right)$$

Efficient Implementation T_K

$$s_t(m, z_{t-1}, q) \propto \exp\left\{f(m) \cdot g_t(q, z_{t-1})\right\}$$

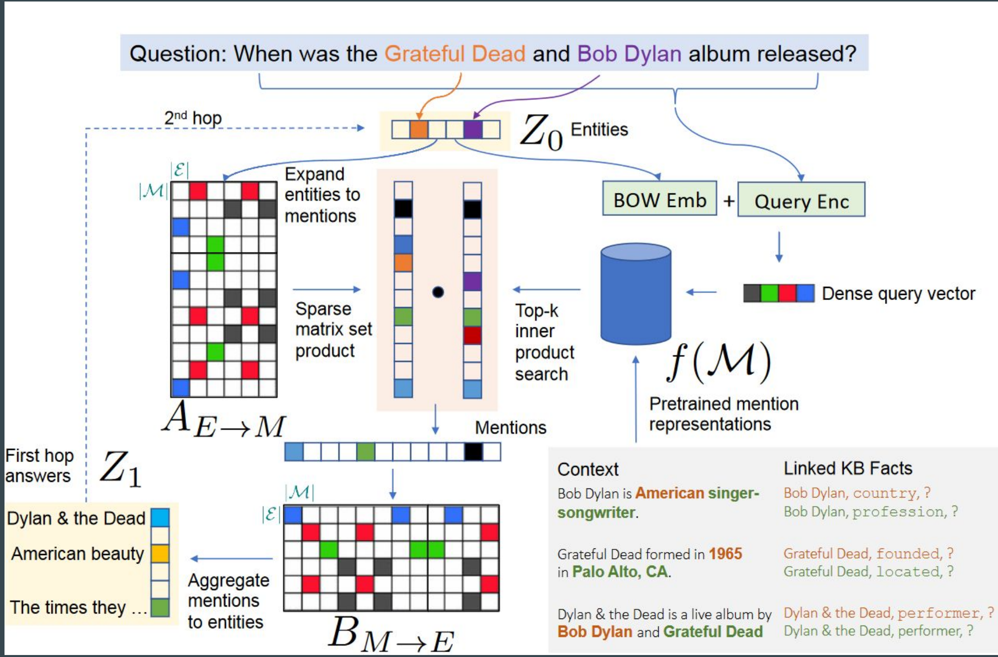
f and g are dense vectors

Matrix multiplication usually expensive

Solution: use Max Inner Product Search (MIPS) to simultaneously take product and take top-K

Review

$$Z_t = \text{softmax}\left(\left[Z_{t-1}^T A_{E\to M} \odot \mathbb{T}_K(s_t(m, z_{t-1}, q))\right] B_{M\to E}\right)$$



Pretraining dense embeddings

End to end training of f(m) is tough: every grad. update -> recompute embeddings of all mentions

BERT out-of-box insufficient

Previous methods: fine-tune BERT via SQuAD

Pretraining

Given:

- KB with facts (e_1, R, e_2)
- Corpus of entity-linked text passages $\{d_k\}$

Automatically identify tuples in corpus

Answer slot-filling queries in RC task

Slot-Filling Task

KB: {... , (Jerry Garcia, birth place, California), ...}

Find passage d mentioning Jerry Garcia and California

Construct query (Jerry Garcia, birth place, ?)

Learn to extract e_2 from d

Also add negative instances:

- Shared-entity negatives: no correct e_2
- Shared-relation negatives: different pair e_1’ and e_2’ have same R
- Random negatives: queries shuffled among passages d

Slot-Filling Task

KB: Wikidata

Corpus: Wikipedia

Identify entity mentions: SLING

Restrict d to Wikipedia article of subject to reduce noise

950K pairs, 550K articles

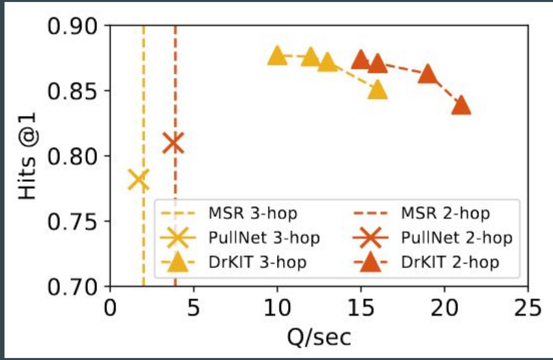
Experiments: MetaQA

Example query: [Joe Thomas] appears in which movies

Example KB entry: Kismet | directed_by | William Dieterle

Paired with corpus of Wikipedia articles

MetaQA			
Model	**1hop**	**2hop**	**3hop**
DrQA (ots)	0.553	0.325	0.197
KVMem†	0.762	0.070	0.195
GraftNet†	0.825	0.362	0.402
PullNet†	0.844	0.810	0.782
DrKIT (e2e)	0.844	0.860	**0.876**
DrKIT (strong sup.)	**0.845**	**0.871**	0.871



Ablations	**1hop**	**2hop**	**3hop**
DrKIT	0.844	0.860	0.876
–Sum over M_{z_t}	0.837	0.823	0.797
$-\lambda = 1$	0.836	0.752	0.799
–w/o TFIDF	0.845	0.548	0.488
–BERT index	0.634	0.610	0.555
Incomplete KB for pretraining			
25% KB	0.839	0.804	0.830
50% KB	0.843	0.834	0.834
(50% KB-only)	*0.680*	*0.521*	*0.597*

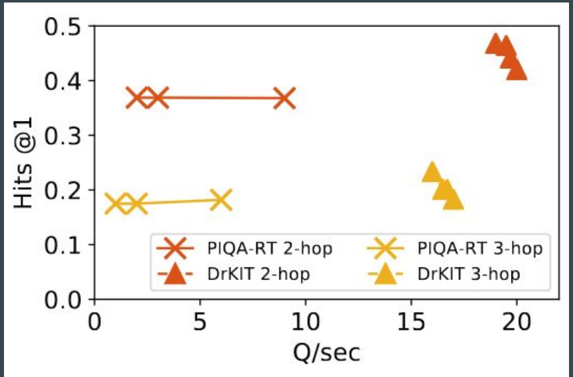
Experiments: WikiData

Larger dataset, unseen documents and entities

Example query: (2000 Hel van het Mergelland, winner, place of birth?)

Example answer: Bert Grabsch→Lutherstadt Wittenberg

WikiData			
Model	**1hop**	**2hop**	**3hop**
DrQA (ots, cascade)	0.287	0.141	0.070
PIQA (ots, cascade)	0.240	0.118	0.064
PIQA (pre, cascade)	0.670	0.369	0.182
DrKIT (pre, cascade)	0.816	0.404	0.198
DrKIT (e2e)	**0.834**	**0.469**	**0.244**
–BERT index	0.643	0.294	0.165



Experiments: HotpotQA

Crowdsourced multi-hop questions over Wikipedia passages

Unknown number of hops

Model	Q/s	Accuracy			
		@2	@5	@10	@20
BM25†	–	0.093	0.191	0.259	0.324
PRF-Task†	–	0.097	0.198	0.267	0.330
BERT re-ranker†	–	0.146	0.271	0.347	0.409
Entity Centric IR†	0.32*	0.230	0.482	0.612	0.674
DrKIT (WikiData)		0.355	0.588	0.671	**0.710**
DrKIT (Hotpot)	**4.26***	**0.385**	0.595	0.663	0.703
DrKIT (Combined)		0.383	**0.603**	**0.672**	**0.710**

Model	EM	F1
Baseline†	0.288	0.381
+EC IR‡	0.354	0.462
+Golden Ret◇	**0.379**	**0.486**
+DrKIT†	0.357	0.466

Experiments: HotpotQA

System	Runtime		Answer		Sup Fact		Joint	
	#Bert	s/Q	EM	F1	EM	F1	EM	F1
Baseline (Yang et al., 2018)	–	–	25.23	34.40	5.07	40.69	2.63	17.85
Golden Ret (Qi et al., 2019)	–	1.4†	37.92	48.58	30.69	64.24	18.04	39.13
Semantic Ret (Nie et al., 2019)	50*	40.0‡	45.32	57.34	38.67	70.83	25.14	47.60
HGN (Fang et al., 2019)	50*	40.0‡	56.71	69.16	**49.97**	76.39	**35.63**	59.86
Rec Ret (Asai et al., 2020)	500*	133.2†	**60.04**	**72.96**	49.08	**76.41**	35.35	**61.18**
DrKIT + BERT	**1.2**◇	**1.3**	42.13	51.72	37.05	59.84	24.69	42.88

Conclusion

Pros:

- End to end differentiable: yes, with pretraining
- Efficient: yes, faster than existing KB and non-KB approaches
- Multi-hop: yes, arguably more flexible than other approaches to multi-hop, due to recursivity (instead of building multi-hop into the length of the architecture)

Cons

- Reliance on entities
- Reliance on mentions (problem for non-redundant corpus)
- Somewhat flimsy probabilistic foundation
- Still unacceptably low performance on certain datasets

…Unanimous accept on OpenReview