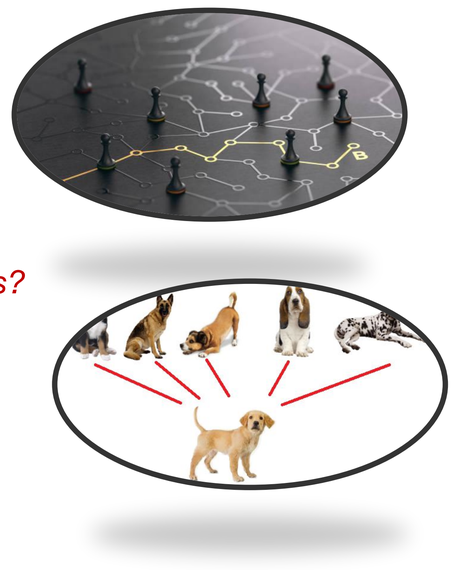
A Study of Zero-shot Adaptation with Commonsense Knowledge

Jiarui Zhang, Filip Ilievski, Kaixin Ma, Jonathan Francis, Alessandro Oltramari

University of Southern California, Information Sciences Institute
Faculty of Sciences, VU Amsterdam
Language Technologies Institute, Carnegie Mellon University
Human-Machine Collaboration, Bosch Research Pittsburgh

Problem Statement

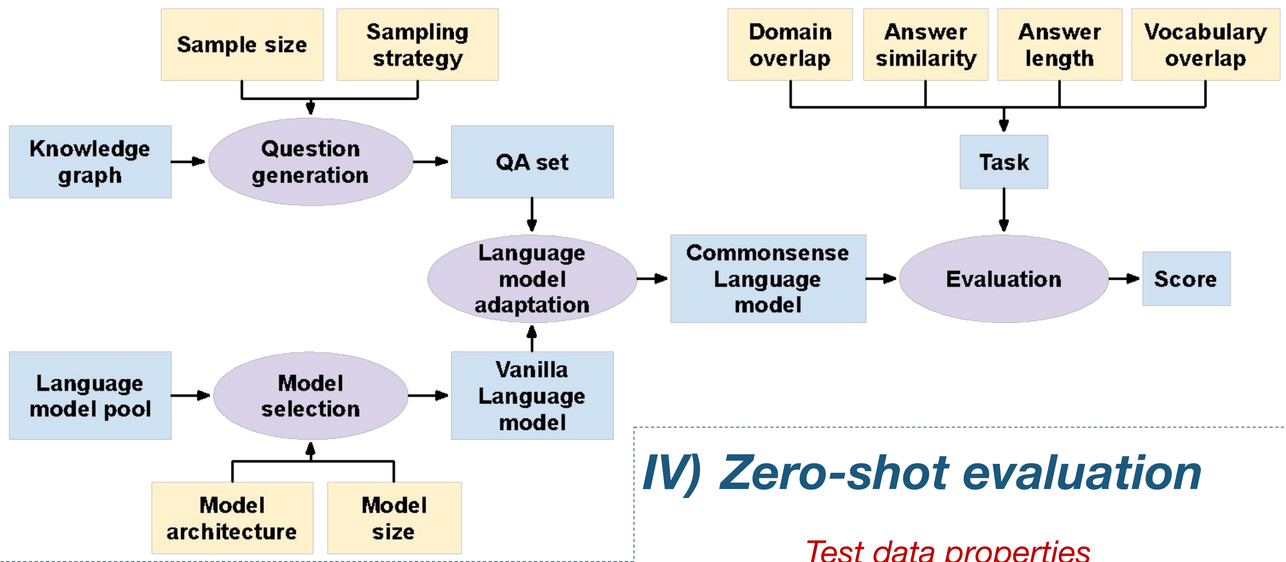
Zero-shot training of in-house LMs with structured knowledge has proven effective (Ma et al., 2021)



Many open questions:

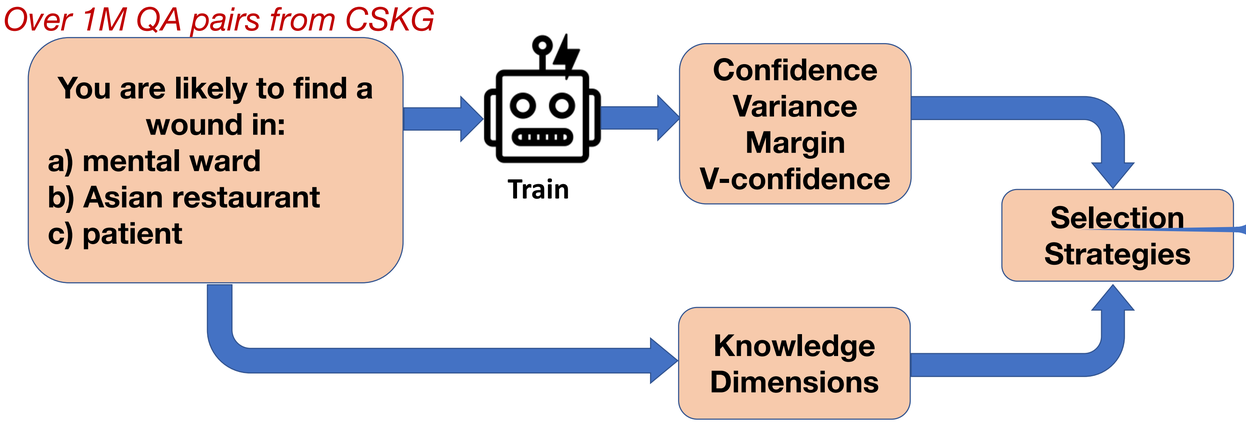
- *Overall impact to models of knowledge training?*
- *The optimal training knowledge data size for LMs?*
- *The best training data sample strategy?*
- *LMs' ability of generalizing the knowledge?*
- *The connection between model's accuracy and properties of the task?*

Research Framework

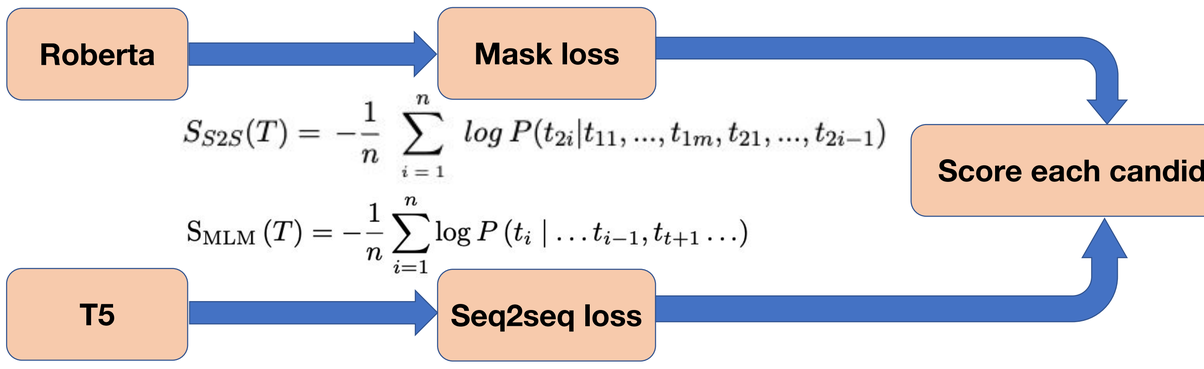


I) Question Generation & Selection

Over 1M QA pairs from CSKG



III) Language Model Selection & Adaptation



$$S_{S2S}(T) = -\frac{1}{n}\sum_{i=1}^{n} log\, P(t_{2i}|t_{11},...,t_{1m},t_{21},...,t_{2i-1})$$

$$S_{MLM}(T) = -\frac{1}{n}\sum_{i=1}^{n} \log P(t_i \mid \ldots t_{i-1}, t_{t+1} \ldots)$$

$$\mathcal{L} = \frac{1}{m}\sum_{\substack{i=1 \\ i\neq y}}^{m} \max(0, \eta - S_y + S_i)$$

IV) Zero-shot evaluation

Test data properties

Length $AL(q) = \sum_{i=1}^{n} |T_{A_i}|$

Similarity $AS(q) = \frac{|T_{A_i} \cap T_{A_j}|}{|T_{A_i} \cup T_{A_j}|}$

Vocabulary $VO(q) = \frac{1}{m}\sum_{k=1}^{m} \frac{1}{f(t_k)}$

Benchmarks

High domain overlap:

CommonsenseQA [Talmor et al., 2019] (CSQA)
SocialIQA [Sap et al., 2019b] (SIQA)

Low domain overlap:

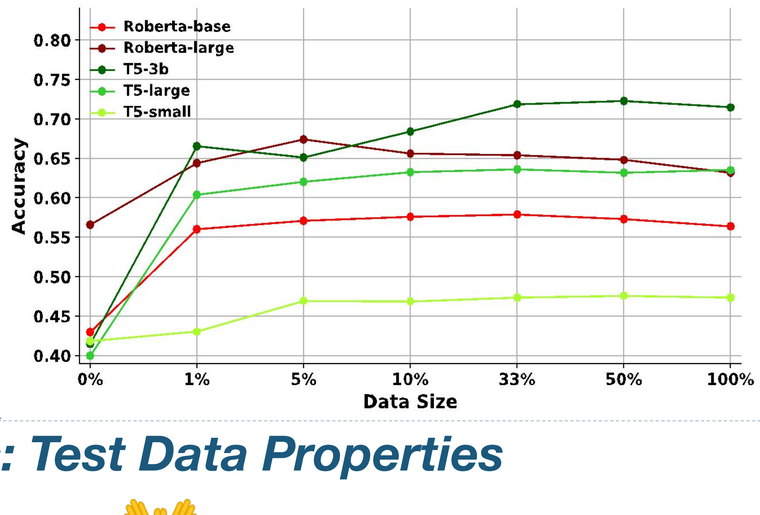
Abductive NLI [Bhagavatula et al., 2019] (ANLI)
PhysicalIQA [Bisk et al., 2020] (PIQA)
WinoGrande [Sakaguchi et al., 2019] (WG)

Results: Best Model Comparison

Model	LDO aNLI	LDO WG	LDO PIQA	HDO SIQA	HDO CSQA	Avg(LDO)	Avg(HDO)	Avg
Majority [Ma et al., 2021a]	50.8	50.4	50.5	33.6	20.9	50.6	27.25	41.2
RoBERTa-large [Liu et al., 2019b]	65.5	57.5	67.6	47.3	45.0	63.5	46.1	56.6
COMET [Bosselut et al., 2019]	-	-	-	50.1	-	-	*50.1	*50.1
Self-Talk [Shwartz et al., 2020]	-	54.7	70.2	46.2	32.4	*62.5	39.3	50.9
SMLM [Banerjee and Baral, 2020]	65.3	-	-	48.5	38.8	*65.3	43.7	50.9
Ma et al. [Ma et al., 2021a]	70.5	60.9	72.4	63.2	67.4	67.9	65.3	66.8
Dou & Peng [Dou and Peng, 2022]	-	-	-	59.9	67.4	-	63.6	63.6
RoBERTa-base (ours)	59.9	53.1	65.7	54.6	53.6	59.6	54.1	57.4
RoBERTa-large (ours)	71.5	60.0	72.6	63.6	66.4	68.0	65.0	66.8
T5-small (ours)	50.6	51.6	56.2	42.3	36.4	52.8	39.4	47.4
T5-large (ours)	66.1	58.7	70.8	57.5	63.1	65.2	60.3	63.2
T5-3b (ours)	**76.6**	**71.0**	**76.7**	**65.3**	**69.9**	**74.7**	**67.6**	**71.9**
RoBERTa-large (supervised)	85.6	79.3	79.2	76.6	78.5	81.4	77.5	79.8
T5-3b (supervised)	87.5	84.4	76.3	78.6	81.5	82.7	80.1	81.7

Consistent improvement across tasks

Optimal training data size depends on LM size and architecture



Results: Sampling Strategies

Strategy		LDO aNLI	LDO WG	LDO PIQA	HDO SIQA	HDO CSQA	Avg(LDO)	Avg(HDO)	Avg
Random	5%	65.9	56.5	70.5	55.4	61.9	64.3	58.7	62.0
Dimension	temporal	66.6	56.4	**71.2**	54.9	**63.4**	**64.7**	**59.2**	**62.5**
	desire	64.4	57.9	69.6	55.9	62.2	64.0	59.1	62.0
	taxonomic	61.8	54.0	66.8	52.8	57.5	60.9	55.2	58.6
	quality	**66.8**	**58.4**	70.0	**56.4**	59.6	65.1	58.0	62.2
	rel-other	61.0	52.5	65.9	51.7	54.0	59.8	52.9	57.0
Uniform		65.3	57.5	69.2	56.6	62.7	64.0	59.7	62.3
Vanilla-conf	high	65.3	56.8	69.0	55.5	57.5	63.7	56.5	60.8
	low	64.0	56.0	68.1	52.0	59.6	62.7	55.8	59.9
Conf	high	62.9	53.8	66.5	53.9	57.0	61.1	55.5	58.8
	low	41.8	48.5	42.0	24.7	07.7	44.1	16.2	32.9
Varibility	high	64.0	54.6	65.1	51.1	54.5	61.2	52.8	57.9
	low	61.7	54.9	66.8	52.7	55.9	61.1	54.3	58.4
Margin	high	63.8	54.5	67.2	52.8	56.9	61.8	54.9	59.0
	low	41.5	45.0	43.7	24.1	09.1	43.4	16.6	32.7

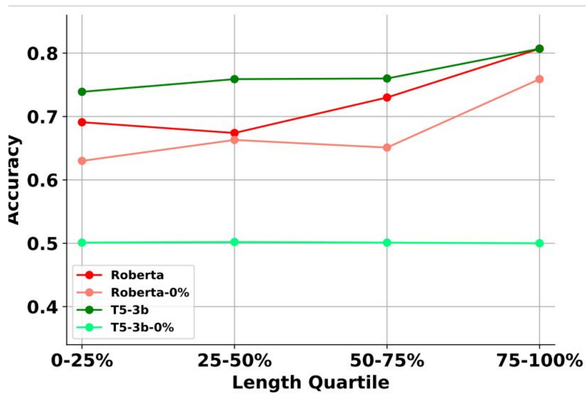
Natural distribution Is the optimal sampling strategy

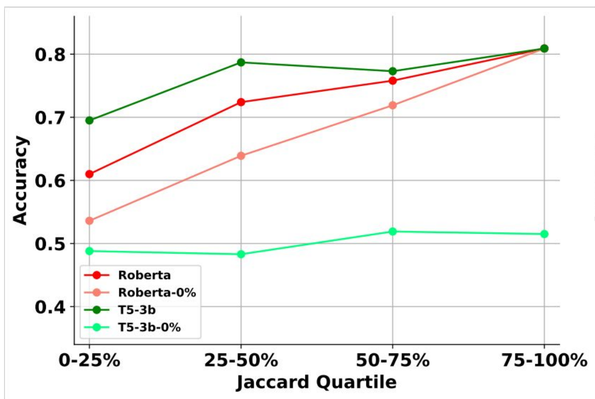
dimension: temporal Q:Jan went out with Quinn's friends and had a great time.What does Jan need to do before this? A1:get dressed(*); A2:cancel her plans; A3:see Quinn's Friends again
dimension: desire Q:Robert has no regret for punching Justin in the nose because _ was the victim of injustice. A1:Robert(*); A2:Justin
dimension: quality Q:What can machines do that humans cannot? A1:fail to work; A2:perform work; A3:answering questions; A4:see work; A5:fly(*)

Dimension-based strategies teach LMs complementary knowledge

Results: Test Data Properties

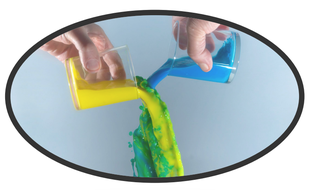
Knowledge training is most effective for questions with short answers and dissimilar answer candidates

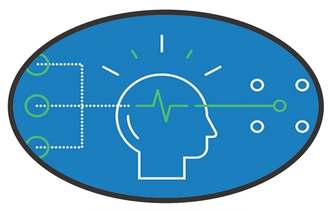




Future Work

Mixture of models





Explainable zero-shot commonsense reasoning

More realistic tasks

Code&data:
https://github.com/saccharomycetes/commonsense-with-KG