Building Agents with Commonsense Knowledge

FILIP ILIEVSKI
Information Sciences Institute
@earthling91
ilievski@isi.edu

2050 A.I. agents

Today's AI lacks Common sense

Figure 1: An image from PASCAL and a high scoring car detection from DPM [8]. Why did the detector fail?

Common sense is the common knowledge about the world that is possessed by every schoolchild and the methods for making obvious inferences from this knowledge.

Davis, E. (2014). Representations of commonsense knowledge.

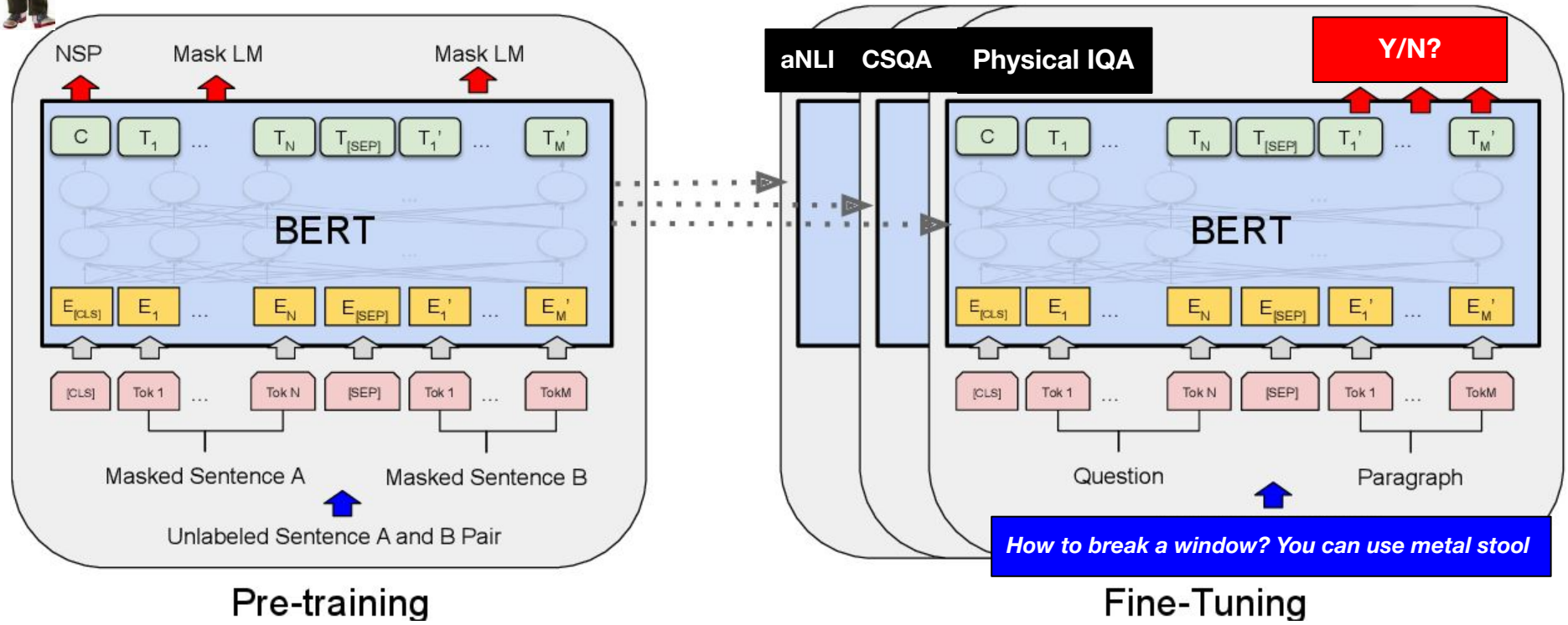
Commonsense knowledge includes the basic facts about events (including actions) and their effects, facts about knowledge and how it is obtained, facts about beliefs and desires. It also includes the basic facts about material objects and their properties.

McCarthy, J. (1989). Artificial intelligence, logic and formalizing common sense.

Research goal:

How do we build a natural-language agent that performs well on all known aspects of common sense?

Transformers as commonsense agents?



You need to break a window. Which object would you rather use?

- **a) a metal stool**
- b) a giant bear
- c) a bottle of water

Physical IQA

Human Performance	Accuracy: 0.9490

Download

Rank	Submission	Accuracy
1	UNICORN *Anonymous*	0.9013
2	UnifiedQA (T5-11B) - finetuned *AI2*	0.8950
3	UnifiedQA (T5-3B) - finetuned *AI2*	0.8526

Have Transformers solved common sense?

In the school play, Robin played a hero in the struggle to the death with the angry villain. How would others feel as a result?

- a) sorry for the villain
- **b) hopeful that Robin will succeed**
- c) like Robin should lose the fight

Social IQA

Human Performance	Accuracy: 0.8810

Download

Rank	Submission	Accuracy
1	UNICORN *Anonymous*	0.8315
2	UnifiedQA-11B (finetuned) *AI2*	0.8145
3	UGAmix *UNC-NLP*	0.8004

How it is done today

+ SocialIQA training data =

SocialIQA test data

+ CommonsenseQA training data =

CSQA test data

+ NumerSense training data =

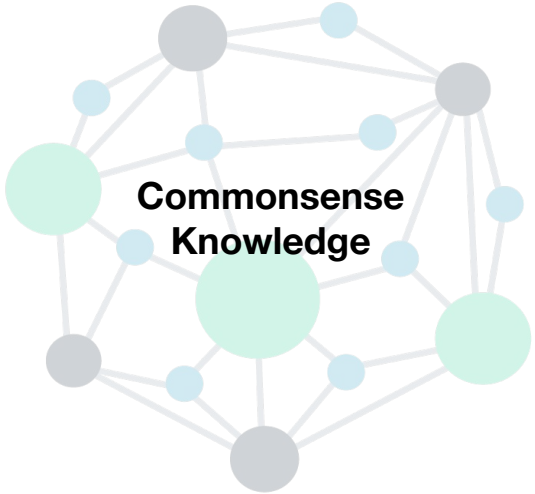
NumerSense test data

+ PhysicalIQA training data =

PhysicalIQA test data

Idea: Use existing commonsense knowledge sources

+

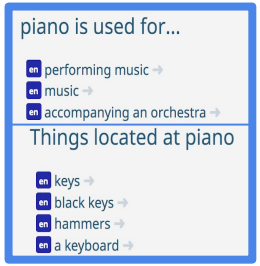


~~SocialIQA training data~~ ~~PhysicalIQA training data~~ ~~NumerSense training data~~ ~~CSQA training data~~

On stage, a woman takes a seat at the piano. She

1. sits on a bench as her sister plays with the doll.
2. smiles with someone as the music plays.
3. is in the crowd, watching the dancers.
4. nervously sets her fingers on the keys.

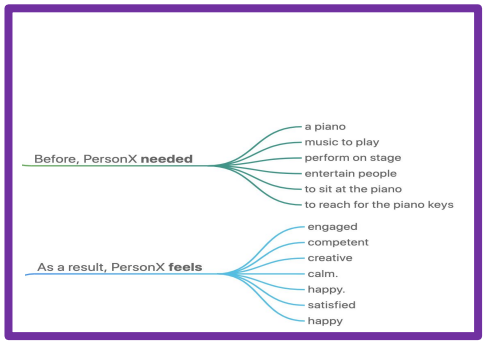
(Zellers et al., 2018)



ConceptNet: pianos have keys, are used to perform music

- S: (n) **piano**, pianoforte, forte-piano (a keyboard instrument that is played by depressing keys that cause hammers to strike tuned strings and produce sounds)

WordNet: pianos are played by pressing keys



ATOMIC: to play piano, a person needs to sit at it, on stage and reach for the keys; feelings

On stage, a woman takes a seat at the piano. She

1. sits on a bench as her sister plays with the doll.
2. smiles with someone as the music plays.
3. is in the crowd, watching the dancers.
4. nervously sets her fingers on the keys.

FrameNet: performer entertains audience

Audience [Aud]	The Audience experiences the Performance.
Medium [Medium]	Medium is the physical entity or channel used by the Performer to transmit the Performance to the Audience.
Performance [Perance]	The Performer generates the Performance which the Audience perceives.
Performer [Perfer]	The Performer provides an experience for the Audience.

Visual Genome: person can play a piano while sitting, his hands are on the keyboard

- man plays piano
- keys ON piano
- woman watches man
- pillow ON couch
- light ON wall
- window IN room
- person playing piano
- guy ON bench
- hands ON keyboard

How to enhance a natural-language agent with commonsense knowledge?

+

Commonsense knowledge sources are heterogeneous

Representation

- symbolic
- natural language
- neural

Acquisition method

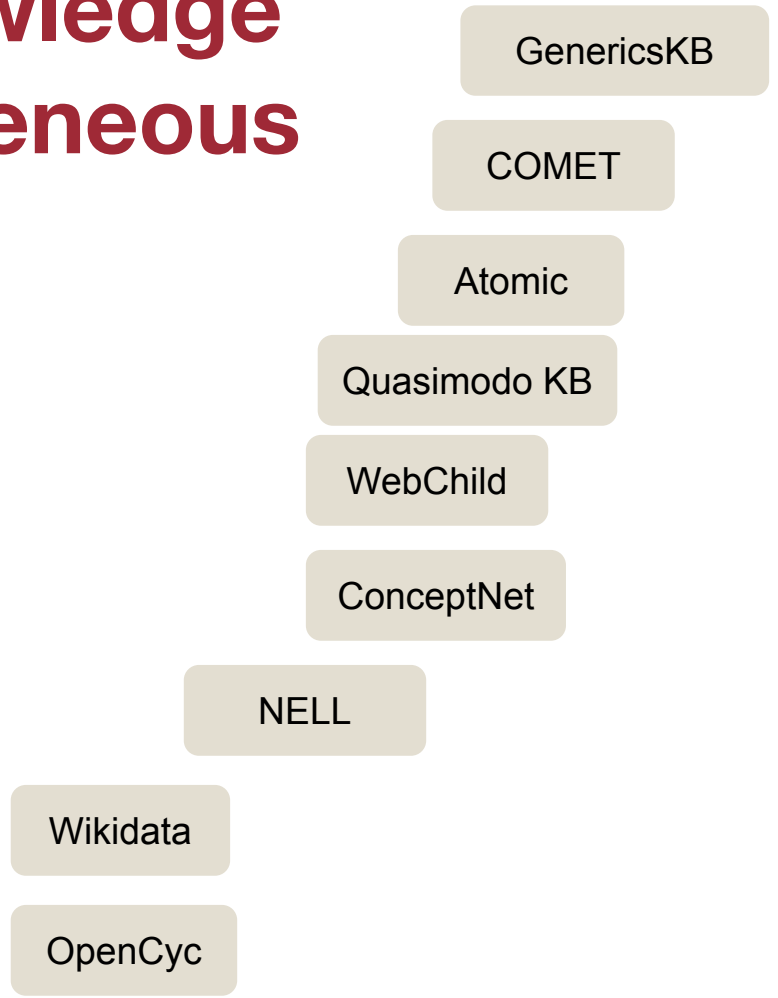
- expert input
- crowdsourcing
- information extraction, machine learning

Knowledge type

- entities and actions
- inferential/rules

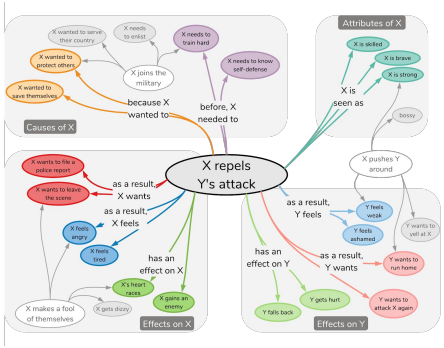
Topic

- general
- social

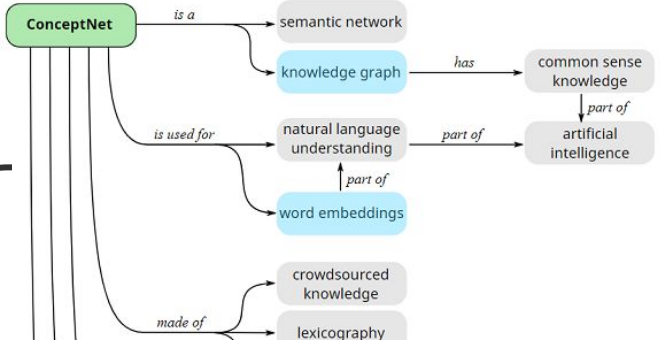


The Commonsense Knowledge Graph (CSKG)

FrameNet (Baker et al., 1998)



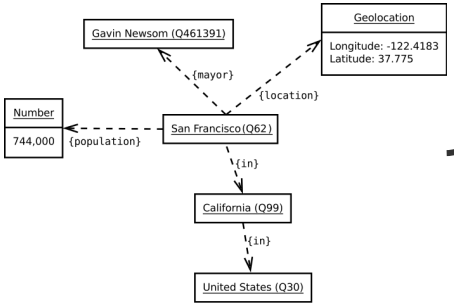
ATOMIC (Sap et al. 2019)



ConceptNet (Speer, Chin and Havasi 2017)

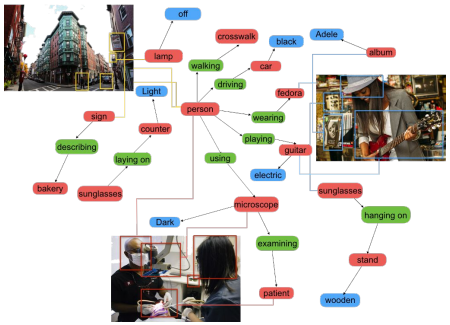
CSKG

CSKG (Ilievski et al. 2021)



Wikidata (Vrandecic and Krotzsch 2014)

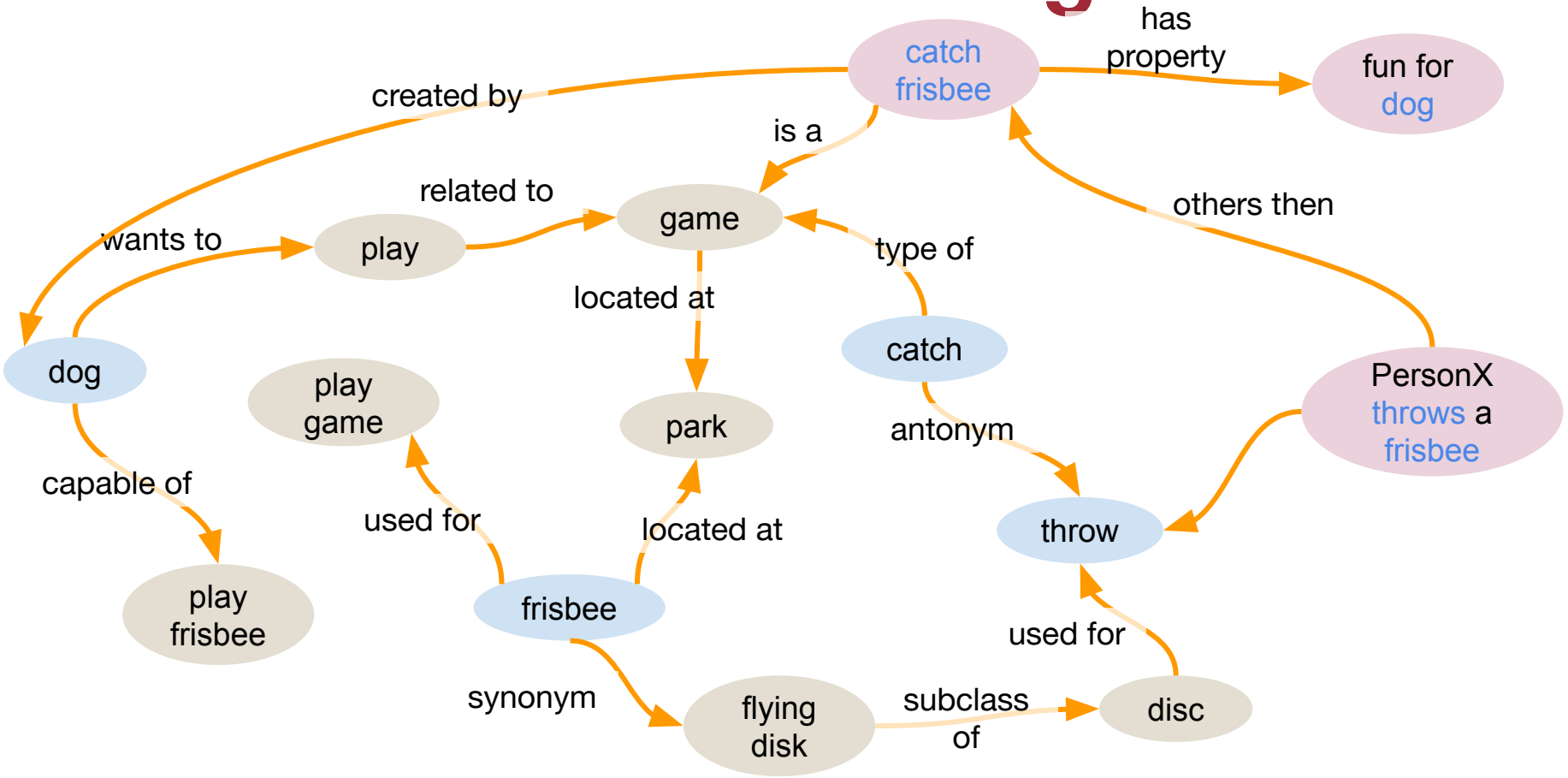
WordNet (Miller 1995)



Visual Genome (Krishna et al. 2017)

F Ilievski, P Szekely, B Zhang (2021). CSKG: The CommonSense Knowledge Graph. ESWC

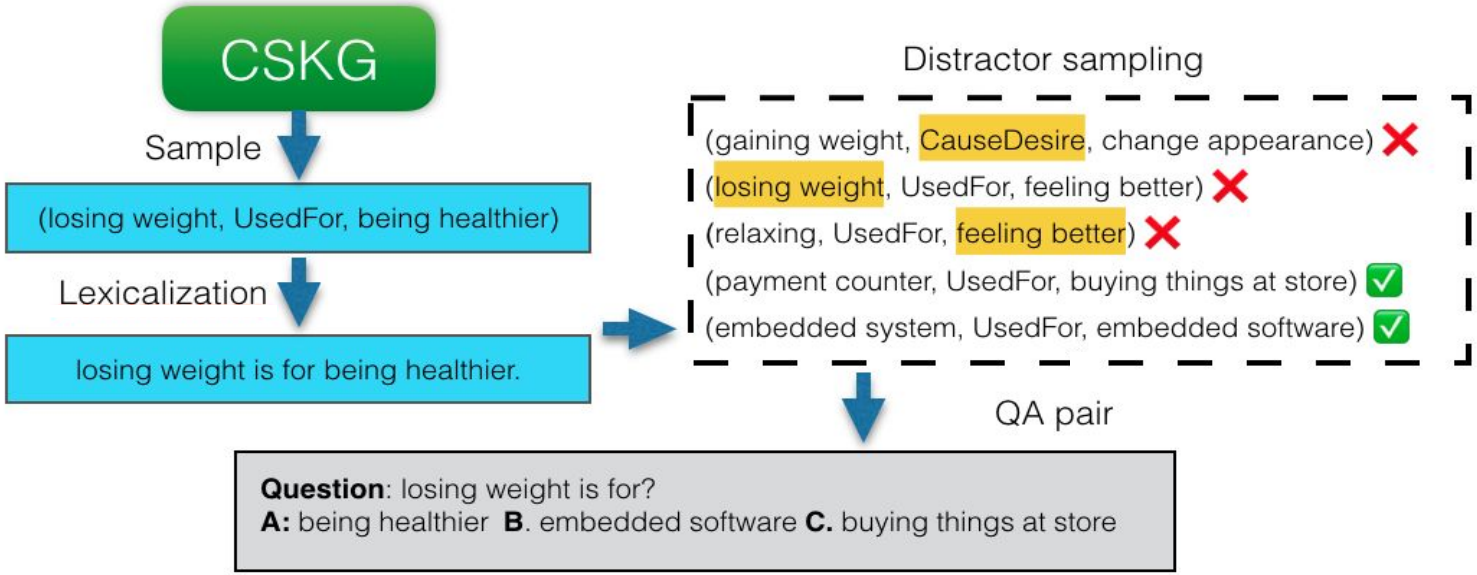
Consolidated knowledge



Generating questions with CSKG

Pretrain LMs with synthetic MCQA sets generated from CSKG

Answer commonsense questions on unseen datasets



K Ma, F Ilievski, J Francis, Y Bisk, E Nyberg, A Oltramari (2021).
Knowledge-driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering. In AAAI

Pre-training LMs with CSKG questions

CSKG

Question: losing weight is for?
A: being healthier **B**. embedded software **C.** buying things at store

SocialIQA test data

CSQA test data

PhysicalIQA test data

Winogrande test data

aNLI test data

Experimental setup

Language models
- *GPT-2*
- *RoBERTa*

CSKG

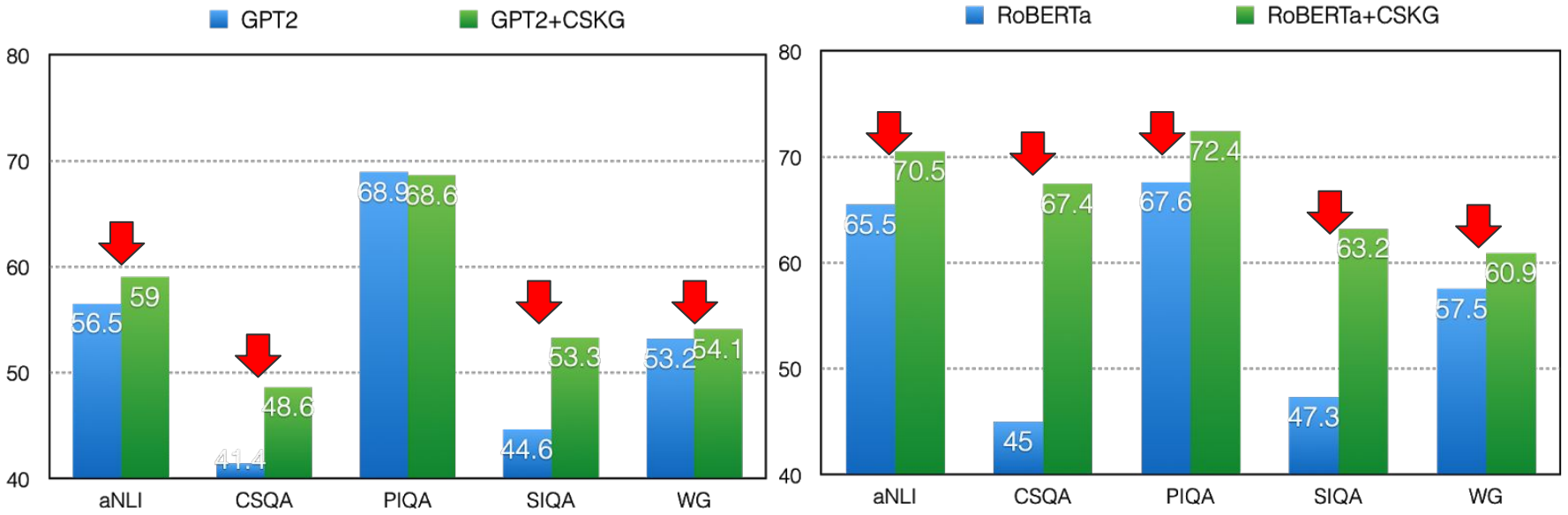
KG subsets
- *ATOMIC*
- *CWWV (ConceptNet, WordNet, Wikidata, Visual Genome)*
- *full CSKG*

\+ **Question**: losing weight is for?
A: being healthier **B**. embedded software **C.** buying things at store =

- SocialIQA test data
- CSQA test data
- PhysicalIQA test data
- Winogrande test data
- aNLI test data

5 different tasks

Pretraining on CSKG MCQA sets helps accuracy



More knowledge is generally better

Model	KG	aNLI	CSQA	PIQA	SIQA	WG
Majority	-	50.8	20.9	50.5	33.6	50.4
GPT2-L	-	56.5	41.4	68.9	44.6	53.2
RoBERTa-L	-	65.5	45.0	67.6	47.3	57.5
Self-talk (Shwartz et al. 2020)	-	-	32.4	70.2	46.2	54.7
COMET-CGA (Bosselut and Choi 2019)	ATOMIC	-	-	-	49.6	-
SMLM (Banerjee and Baral 2020)	ATOMIC	-	-	-	48.5	-
GPT2-L (MR)	ATOMIC	59.2(±0.3)	48.0(±0.9)	67.5(±0.7)	53.5(±0.4)	54.7(±0.6)
GPT2-L (MR)	CWWV	58.3(±0.4)	46.2(±1.0)	68.6(±0.7)	48.0(±0.7)	52.8(±0.9)
GPT2-L (MR)	CSKG	59.0(±0.5)	48.6(±1.0)	68.6(±0.9)	53.3(±0.5)	54.1(±0.5)
RoBERTa-L (MR)	ATOMIC	**70.8(±1.2)**	64.2(±0.7)	72.1(±0.5)	63.1(±1.5)	59.6(±0.3)
RoBERTa-L (MR)	CWWV	70.0(±0.3)	**67.9(±0.8)**	72.0(±0.7)	54.8(±1.2)	59.4(±0.5)
RoBERTa-L (MR)	CSKG	70.5(±0.2)	67.4(±0.8)	**72.4(±0.4)**	**63.2(±0.7)**	**60.9(±0.8)**
RoBERTa-L (supervised)	-	85.6	78.5	79.2	76.6	79.3
Human	-	91.4	88.9	94.9	86.9	94.1

The impact of more knowledge depends on the KG-task alignment

Model	KG	aNLI	CSQA	PIQA	SIQA	WG
Majority	-	50.8	20.9	50.5	33.6	50.4
GPT2-L	-	56.5	41.4	68.9	44.6	53.2
RoBERTa-L	-	65.5	45.0	67.6	47.3	57.5
Self-talk (Shwartz et al. 2020)	-	-	32.4	70.2	46.2	54.7
COMET-CGA (Bosselut and Choi 2019)	ATOMIC	-	-	-	49.6	-
SMLM (Banerjee and Baral 2020)	ATOMIC	-	-	-	48.5	-
GPT2-L (MR)	ATOMIC	59.2(±0.3)	48.0(±0.9)	67.5(±0.7)	53.5(±0.4)	54.7(±0.6)
GPT2-L (MR)	CWWV	58.3(±0.4)	46.2(±1.0)	68.6(±0.7)	48.0(±0.7)	52.8(±0.9)
GPT2-L (MR)	CSKG	59.0(±0.5)	48.6(±1.0)	68.6(±0.9)	53.3(±0.5)	54.1(±0.5)
RoBERTa-L (MR)	ATOMIC	**70.8(±1.2)**	64.2(±0.7)	72.1(±0.5)	63.1(±1.5)	59.6(±0.3)
RoBERTa-L (MR)	CWWV	70.0(±0.3)	**67.9(±0.8)**	72.0(±0.7)	54.8(±1.2)	59.4(±0.5)
RoBERTa-L (MR)	CSKG	70.5(±0.2)	67.4(±0.8)	**72.4(±0.4)**	**63.2(±0.7)**	**60.9(±0.8)**
RoBERTa-L (supervised)	-	85.6	78.5	79.2	76.6	79.3
Human	-	91.4	88.9	94.9	86.9	94.1

Takeaways

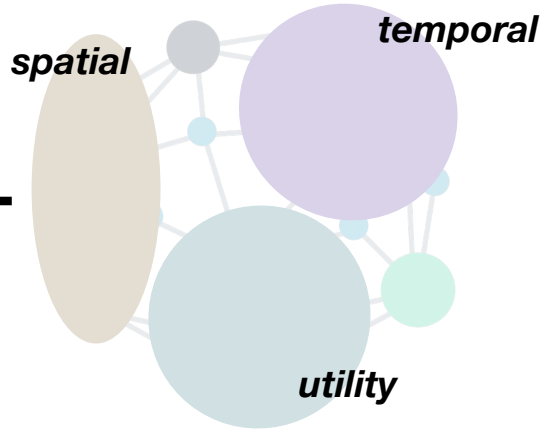
A single pre-trained model can learn to perform all the tasks

Pretraining with CSKG knowledge brings large and consistent improvement over vanilla LMs

More knowledge helps, IF well-aligned with the task

What is the role of different types of commonsense knowledge?

+



Approach

1. Organize the CSKG knowledge along dimensions (knowledge types)

2. Align knowledge dimensions with tasks

F Ilievski, A Oltramari, K Ma, B Zhang, DL McGuinness, P Szekely (2021). Dimensions of commonsense knowledge. Under review

Approach

1. ***Organize the CSKG knowledge along dimensions (knowledge types)***
    a. ***Survey relation types in relevant sources***
    b. ***Manually cluster them into dimensions***

2. **Align knowledge dimensions with tasks**

F Ilievski, A Oltramari, K Ma, B Zhang, DL McGuinness, P Szekely (2021). Dimensions of commonsense knowledge. Under review

Overview of sources

Category	Source	Relations
Commonsense KGs	ConceptNet*	34
	ATOMIC	9
	GLUCOSE	10
	WebChild	4 (groups)
	Quasimodo	78,636
	SenticNet	4
	HasPartKB	1
Common KGs	Wikidata	6.7k
	YAGO4	116
	DOLCE*	1
	SUMO*	1,614
Lexical resources	WordNet	10
	Roget	2
	FrameNet	8 (f2f)
	MetaNet	14 (f2f)
	VerbNet	36 (roles)
Visual sources	Visual Genome	42,374
	Flickr30k	1
Corpora & LMs	GenericsKB	n/a
	GPT-2	n/a

F Ilievski, A Oltramari, K Ma, B Zhang, DL McGuinness, P Szekely (2021). Dimensions of commonsense knowledge. Under review

Annotation of dimensions (1/2)

Dimension	ATOMIC	ConceptNet	WebChild	Other	Wikidata
lexical		FormOf DerivedFrom EtymologicallyDerivedFrom		lexical_unit (FN) lemma (WN)	label
similarity		Synonym SimilarTo DefinedAs	hassimilar	reframing_mapping (FN) metaphor (FN) Synonym (RG) synonym (WN)	said to be the same as
distinctness		Antonym DistinctFrom		Antonym (RG) antonym (WN) excludes (FN)	different from opposite of
taxonomic		IsA InstanceOf MannerOf	hasHypernymy	perspective_on (FN) inheritance (FN) hypernym (WN)	subClassOf instanceOf description
part-whole		PartOf HasA MadeOf AtLocation*	physicalPartOf memberOf substanceOf	HasPart (HP) meronym (WN) holonym (WN)	has part member of material used
spatial		AtLocation* LocatedNear	location spatial		location anatomical location
creation		CreatedBy			creator

Annotation of dimensions (2/2)

Dimension	ATOMIC	ConceptNet	WebChild	Other	Wikidata
utility		ReceivesAction UsedFor CapableOf ¬NotCapableOf	hassynsetmember activity participant	using (FN)	used by use uses
desire/goal	xIntent xWant oWant	CausesDesire MotivatedByGoal Desires ¬NotDesires ObstructedBy			
quality	 xAttr	 HasProperty ¬NotHasProperty SymbolOf	shape size color taste_property temperature	frame_element (FN)	 color has quality
comparative			*6.3k relations*		
temporal	xNeed xEffect oEffect xReact oReact	HasFirstSubevent HasLastSubevent HasSubevent HasPrerequisite Causes Entails	 time emotion prev next	subframe (FN) precedes (FN) inchoative_of (FN) causative_of (FN)	 has cause has effect
relational -other		RelatedTo HasContext EtymologicallyRelatedTo	 thing agent	 see_also (FN) requires (FN)	field of this occupation depicts health specialty

Dimensions of commonsense knowledge in bottom-up sources

lexical

similarity

distinctness

taxonomic

part-whole

spatial

creation

utility

desire/goal

quality

comparative

temporal

relational-other

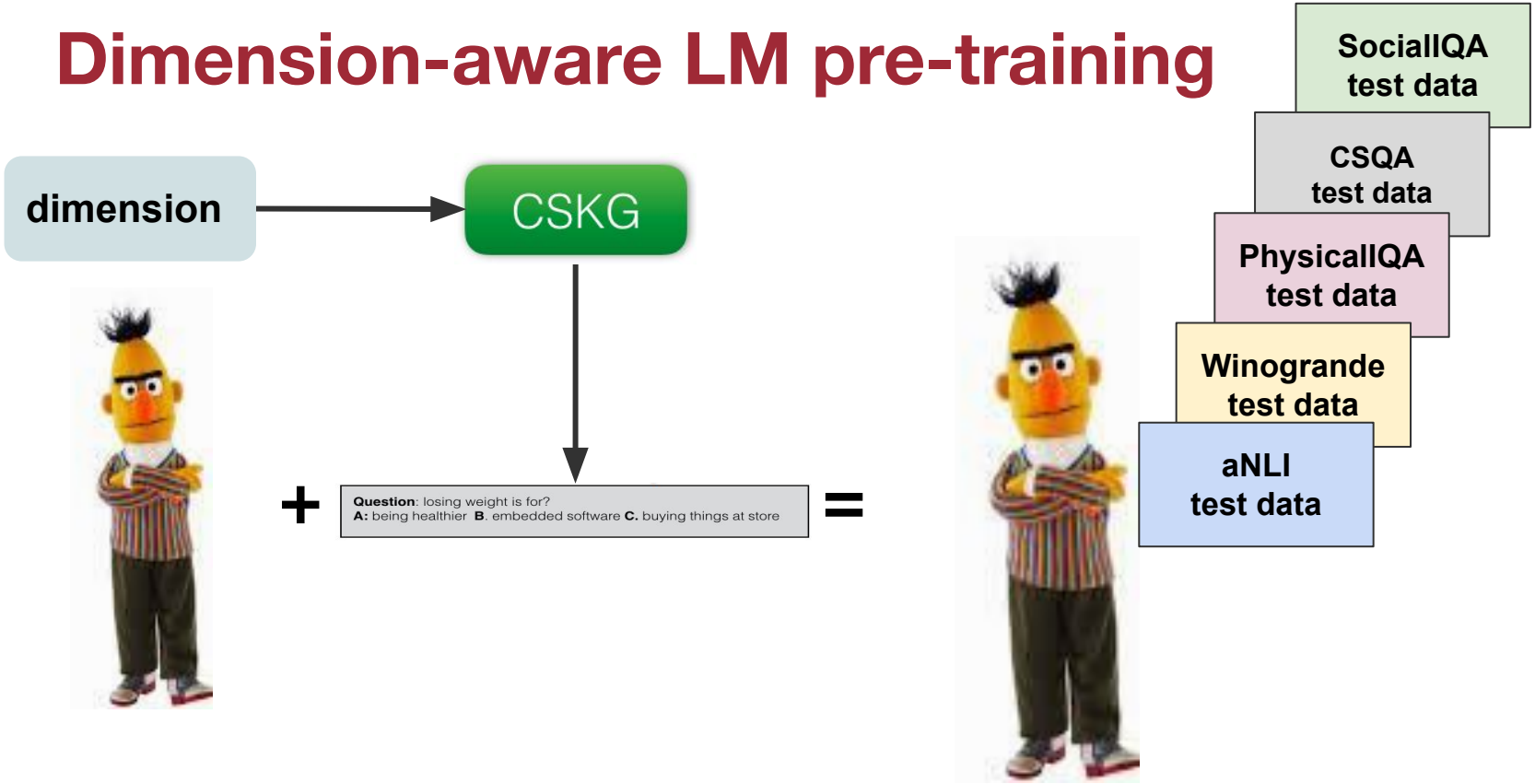
Approach

1. Organize the CSKG knowledge along dimensions (knowledge types)
    a. Survey relation types in relevant sources
    b. Manually cluster them into dimensions

2. **Align knowledge dimensions with tasks**
    a. **Split CSKG along its dimensions**
    b. **Generate pre-training data per dimension**
    c. **Measure the impact of each dimension against each task**

F Ilievski, A Oltramari, K Ma, B Zhang, DL McGuinness, P Szekely (2021). Dimensions of commonsense knowledge. Under review

Dimension-aware LM pre-training



One dimension at a time to measure their impact on a task

Dimensions	Train	Dev
part-whole	87,765	4,620
taxonomic	340,609	17,927
lexical	107,861	5,677
distinctness	20,286	1,068
similarity	166,575	8,768
quality	116,593	12,492
utility	63,862	3,362
creation	304	17
temporal	312,628	31,587
relational-other	242,759	12,777
spatial	21,726	1,144
desire/goal	194,906	20,912

Pre-training language models with dimensions

CSQA = Commonsense QA

SIQA = SocialIQA

Dimensions	CSQA	SIQA
Baseline	45.0	47.3
+part-whole	63.0(±1.4)	52.6(±1.9)
+taxonomic	62.6(±1.4)	52.2(±1.6)
+lexical	49.9(±2.9)	49.0(±0.4)
+distinctness	57.2(±0.5)	50.2(±1.5)
+similarity	61.4(±0.8)	53.5(±0.6)
+quality	65.7(±0.5)	60.0(±0.7)
+utility	**67.4(±1.0)**	54.8(±0.7)
+creation	49.9(±1.1)	47.8(±0.2)
+temporal	67.3(±0.3)	**62.6(±0.9)**
+relational-other	58.2(±1.7)	51.3(±1.7)
+spatial	63.3(±0.2)	53.1(±0.3)
+desire/goal	65.0(±1.8)	60.0(±0.6)
+all	66.2(±1.4)	61.0(±0.7)

Novelty per dimension

Can 'vanilla' RoBERTa answer the questions without pretraining?

Dimensions	Dev
part-whole	67.5
taxonomic	57.0
lexical	90.1
distinctness	77.3
similarity	65.6
quality	45.5
utility	67.9
creation	82.4
temporal	47.2
relational-other	37.6
spatial	56.9
desire/goal	48.0

Findings

The knowledge dimensions of CSKG allow for controlled adaptation of LMs

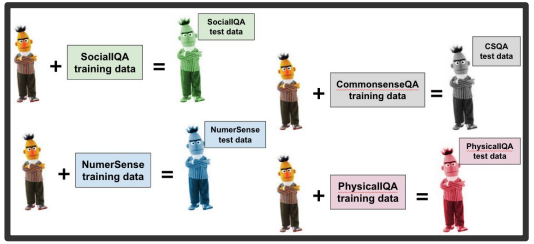
Not all dimensions are as informative to LMs

- **lexical and distinctness knowledge is largely redundant**
- **temporal and desire/goal knowledge is both novel and useful**

Research goal:

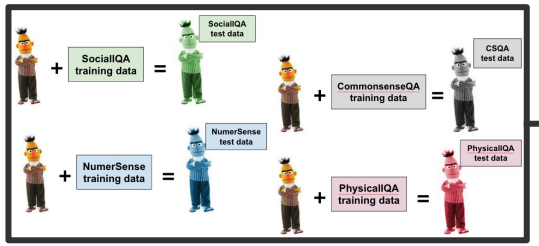
How do we build a natural-language agent that performs well on all known aspects of common sense?

The story so far



Many dataset-specific NL commonsense agents

The story so far



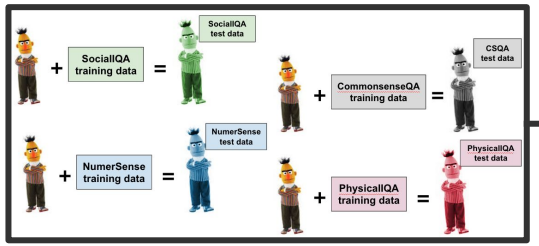
Many dataset-specific NL commonsense agents

Pre-training LMs with commonsense knowledge

KG pre-training method

CSKG

The story so far

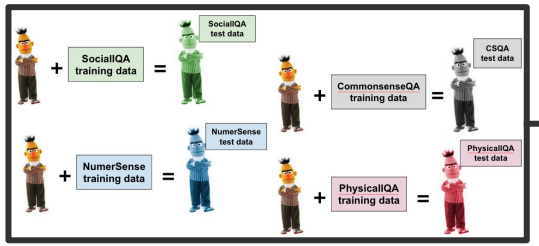


Many dataset-specific NL commonsense agents

Pretraining LMs with all dimensions of commonsense knowledge

KG pre-training method | **CSKG** | **dimensions** | **data maps**

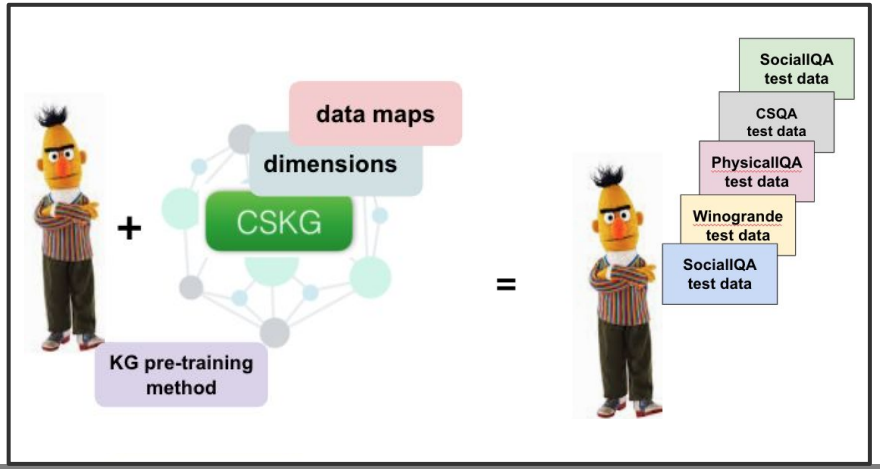
The story so far



Many dataset-specific NL commonsense agents

Pretraining LMs with all dimensions of commonsense knowledge

KG pre-training method | **CSKG** | **dimensions** | **data maps**



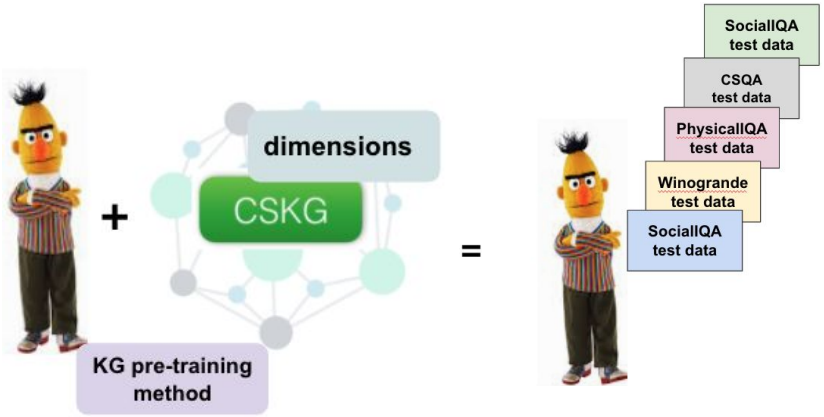
A single NL agent for all commonsense dimensions

What about ‘perform well’?

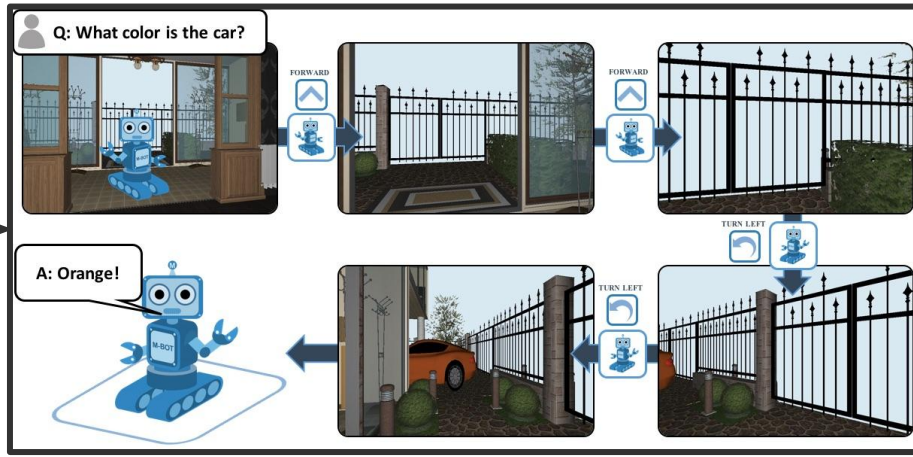
Model	KG	aNLI	CSQA	PIQA	SIQA	WG
Majority	-	50.8	20.9	50.5	33.6	50.4
RoBERTa-L	-	65.5	45.0	67.6	47.3	57.5
RoBERTa-L (MR)	CSKG	70.5(±0.2)	67.4(±0.8)	**72.4(±0.4)**	**63.2(±0.7)**	**60.9(±0.8)**
RoBERTa-L (supervised)	-	85.6	78.5	79.2	76.6	79.3
Human	-	91.4	88.9	94.9	86.9	94.1

Are we there yet!?

Vision: Towards a general embodied agent



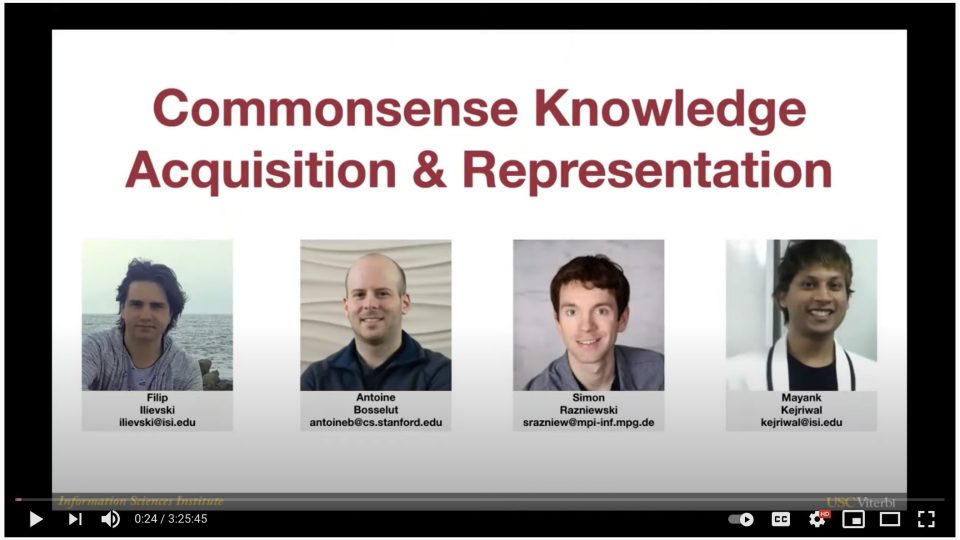
?

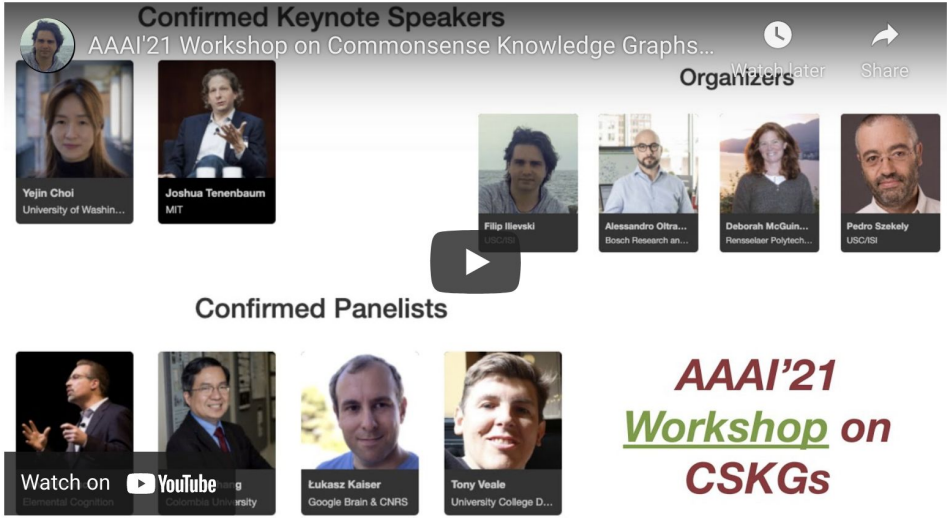


A single NL agent for all commonsense dimensions

A single embodied agent for all commonsense dimensions

Recent events





Upcoming

Special issue on Commonsense Knowledge and Reasoning, Semantic Web Journal

Editors: Filip Ilievski, Antoine Bosselut, Ken Forbus, Simon Razniewski, Vered Schwarz

Deadline: October 20th, 2021

Thanks!

@earthling91
ilievski@isi.edu

References (1)

Davis, E. (2014). Representations of commonsense knowledge. Morgan Kaufmann.

McCarthy, J. (1989). Artificial intelligence, logic and formalizing common sense. In Philosophical logic and artificial intelligence (pp. 161-190). Springer, Dordrecht.

Hayes, P. J. (1979). The naive physics manifesto. Expert systems in the microelectronic age.

Hayes, P. (1985). Naive physics manifesto I: Ontology for liquids. Formal theories of the commmonsense world, 71-107.

Forbus, K. D. (1980, August). Spatial and Qualitative Aspects of Reasoning about Motion. In AAAI (Vol. 80, pp. 170-173).

Davis, E. (1993). The kinematics of cutting solid objects. Annals of Mathematics and Artificial Intelligence, 9(3), 253-305.

Davis, E. (1998). Naive physics perplex. AI magazine, 19(4), 51-51.

References (2)

Gordon, A. S., & Hobbs, J. R. (2017). A formal theory of commonsense psychology: How people think people think. Cambridge University Press.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11), 33-38.

Speer, R., Chin, J., & Havasi, C. (2017, February). Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv preprint arXiv:1808.05326.

References (3)

Ilievski, F., Szekely, P., Zhang, B. (2021). CSKG: The CommonSense Knowledge Graph. ESWC

Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D. L., & Szekely, P. (2021). Dimensions of commonsense knowledge. arXiv preprint arXiv:2101.04640.

Ma, K., Ilievski, F., Francis, J., Bisk, Y., Nyberg, E., & Oltramari, A. (2021). Knowledge-driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering. AAAI

Forbus, K. & Hinrichs, T. (2017). Analogy and Qualitative Representations in the Companion Cognitive Architecture. AI Magazine 38(4):34-42.

Botschen, T., Sorokin, D., & Gurevych, I. (2018, November). Frame-and entity-based knowledge for common-sense argumentative reasoning. In Proceedings of the 5th Workshop on Argument Mining (pp. 90-96).

Becker, M., Hulpuş, I., Opitz, J., Paul, D., Kobbe, J., Stuckenschmidt, H., & Frank, A. (2020). Explaining Arguments with Background Knowledge. Datenbank-Spektrum, 20(2), 131-141.

References (4)

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., ... & Choi, Y. (2019, July). Atomic: An atlas of machine commonsense for if-then reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 3027-3035).

Lin, B. Y., Shen, M., Xing, Y., Zhou, P., & Ren, X. (2019). Commongen: A constrained text generation dataset towards generative commonsense reasoning.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8, 34-48.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). Adversarial NLI: A new benchmark for natural language understanding. arXiv preprint arXiv:1910.14599.

Kalyanpur, A., Breloff, T., Ferrucci, D., Lally, A., & Jantos, J. (2020). Braid: Weaving Symbolic and Statistical Knowledge into Coherent Logical Explanations. arXiv preprint arXiv:2011.13354.

References (5)

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-10).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and brain sciences, 40.

Urbanek, J., Fan, A., Karamcheti, S., Jain, S., Humeau, S., Dinan, E., ... & Weston, J. (2019). Learning to speak and act in a fantasy text adventure game. arXiv preprint arXiv:1903.03094.

Fan, A., Urbanek, J., Ringshia, P., Dinan, E., Qian, E., Karamcheti, S., ... & Weston, J. (2020, April). Generating interactive worlds with text. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 02, pp. 1693-1700).

Marcus, G. (2019). Deep Understanding: The Next Challenge for AI. Neurips, slides.