TextGraphs 2020 Shared Task on Explanation Regeneration for Multi-hop Inference

Peter Jansen, School of Information, University of Arizona
Dmitry Ustalov, Yandex

cognitiveai.org

Overview: Shared Task on Explanation Regeneration

Introduction

What is multi-hop inference?

What is explanation regeneration?

Why is it hard?

What explanation corpus was used?

What approaches did the participating teams take?

How well did they do? (automatic measures)

How well did they really do? (additional measures)

Where we are, and where we're going

My long term interest is in building **inference algorithms** capable of **answering questions** and producing **human-readable explanations** by **aggregating information from multiple sources** and knowledge bases.

My long term interest is in building **inference algorithms** capable of **answering questions** and producing **human-readable explanations** by **aggregating information from multiple sources** and knowledge bases.

Three observations in the context of science exams and multi-hop inference:

Standardized elementary science exam questins require an average of 4 to 6 facts (range 1-16) to answer and explain their reasoning.
(Jansen et al., COLING 2016; Jansen et a., LREC 2018 Explanation Corpus)

Assembling long chains of facts to answer questions is challenging. Inference algorithms typically struggle to aggregate more than 2 pieces of information together due to "semantic drift".
(Fried et al., TACL 2015; Khashabi et al., IJCAI 2016; Jansen et al., CL 2017)

Shared Task: Can we make significant progress in combining very large amounts of knowledge (up to 16 facts) and apply this towards explanation-centered inference?
(This shared task)

Aggregation and Inference for Science Exams

Q: Which of the following is an example of an organism taking in nutrients?

(A) A dog burying a bone
(B) A girl eating an apple
(C) An insect crawling on a leaf
(D) A boy planting tomatoes

Science Exam Example Question

Q: Which of the following is an example of an organism taking in nutrients?

(A) A dog burying a bone
(B) A girl eating an apple
(C) An insect crawling on a leaf
(D) A boy planting tomatoes

Rarely will we be able to **retrieve a single passage** in a corpus that directly answers a given question:

"A girl eating an apple is an example of an organism taking in nutrients..."

Aggregation and Inference for Science Exams

Q: Which of the following is an example of an organism taking in nutrients?

(A) A dog burying a bone
(B) A girl eating an apple
(C) An insect crawling on a leaf
(D) A boy planting tomatoes

Girl

"a girl means a human girl"

"humans are living organisms"

Simple Wiktionary

AND

Eating

"eating is when an organism takes in nutrients in the form of food"

4th Grade Study Guide

AND

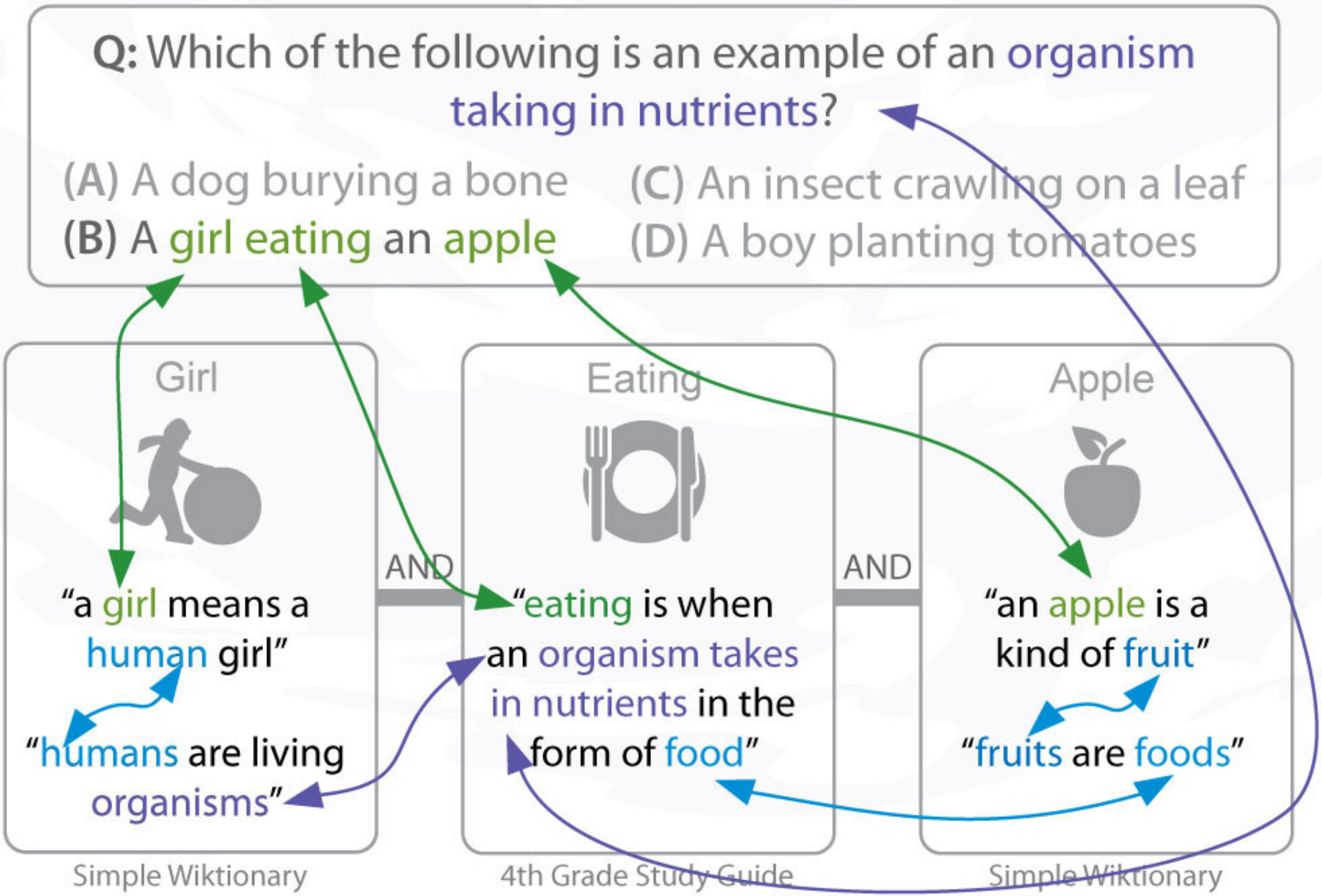
Apple

"an apple is a kind of fruit"

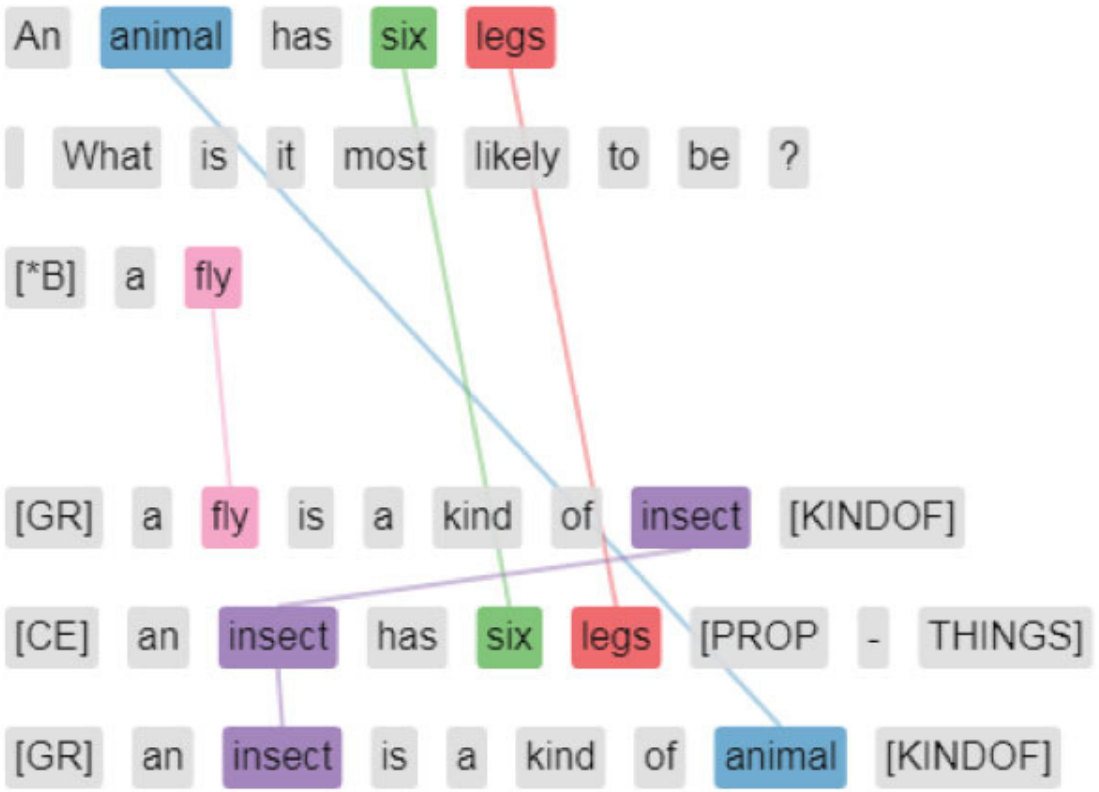
"fruits are foods"

Simple Wiktionary

Aggregation and Inference for Science Exams



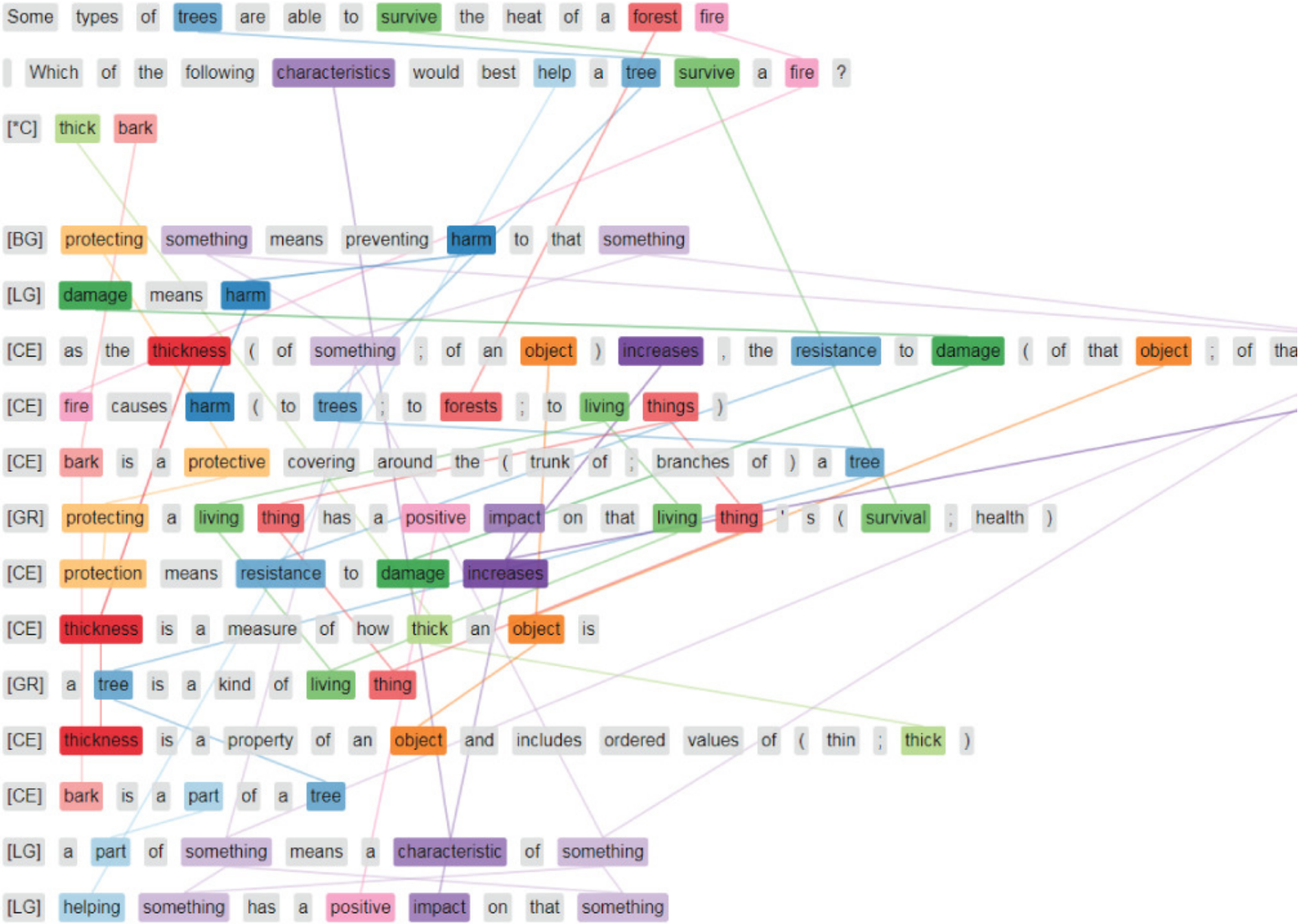
What is Explanation Regeneration?



Given a
Question
and
Correct Answer

Select a series of
interconnected facts
from a knowledge base
that explain
why the answer is correct.

Evaluation: Compare the facts
a model selects with
gold explanations


Some types of trees are able to survive the heat of a forest fire
Which of the following characteristics would best help a tree survive a fire ?
[*C] thick bark
[BG] protecting something means preventing harm to that something
[LG] damage means harm
[CE] as the thickness (of something ; of an object) increases , the resistance to damage (of that object ; of tha
[CE] fire causes harm (to trees ; to forests ; to living things)
[CE] bark is a protective covering around the (trunk of ; branches of) a tree
[GR] protecting a living thing has a positive impact on that living thing ' s (survival ; health)
[CE] protection means resistance to damage increases
[CE] thickness is a measure of how thick an object is
[GR] a tree is a kind of living thing
[CE] thickness is a property of an object and includes ordered values of (thin ; thick)
[CE] bark is a part of a tree
[LG] a part of something means a characteristic of something
[LG] helping something has a positive impact on that something

Explanatory Depth in WorldTree

WorldTree V2 Corpus: 4,400 Standardized Science Questions paired with detailed, manually authored, and lexically-connected explanation graphs. (Xie et al., LREC 2020; Jansen et al., LREC 2018)

Question Which characteristic would best help a tree survive the heat of a forest fire?
Answers [A] large leaves [B] shallow roots *[*C] thick bark* [D] thin trunks

Domain Expert (e.g. teacher)
Bark is a protective covering around the trunk and branches of a tree.

Domain Novice (e.g. student)
As an object's thickness increases, it's resistance to damage will also increase.

Young Child (e.g. 5-year old)
Protecting something means preventing harm.
Fire causes harm to trees, forests, and other living things.
Thickness is a measure of how thick an object is.
A tree is a kind of living thing.

First Principles
Protecting a living thing has a positive impact on it's survival and health

Target Depth

Explanations Grounded in Semi-structured Tables

Explanations are represented as one or more rows in 62 semi-structured tables, providing both ***coarse (sentence-level)*** and ***fine-grained (table column)*** *explanation graph* structure.

Question: Which event involves a consumer and a producer in a food chain ?

Process Roles Table

	PROCESS NAME		ACTOR		ROLE		ACTION	PATIENT		PURPOSE
In the	food chain	process, a	green plant	has the role of	producer	which	creates	food	for	consumers
In the	food chain	process, an	animal	has the role of	consumer	which	eats	producers	for	food
In the	food chain	process, a	bacteria	has the role of	decomposer	which	recycles	nutrients		
In the	tree reproduction	process, a	squirrel	has the role of	seed disperser	which	relocates	seeds		

Taxonomy Table

	HYPONYM		SCOPE	HYPERNYM
A	deer	is a kind of		animal
	green	is a kind of		color
	shelter	is a kind of	protective	covering
An	electromagnet	is a kind of	electric	magnet

PartOf Table

	PART		WHOLE
A	leaf	is a part of a	green plant
	roots	are a part of a	plant
	pedals	are a part of a	bicycle
A	cell wall	is a part of a	plant cell

Answer Candidates:

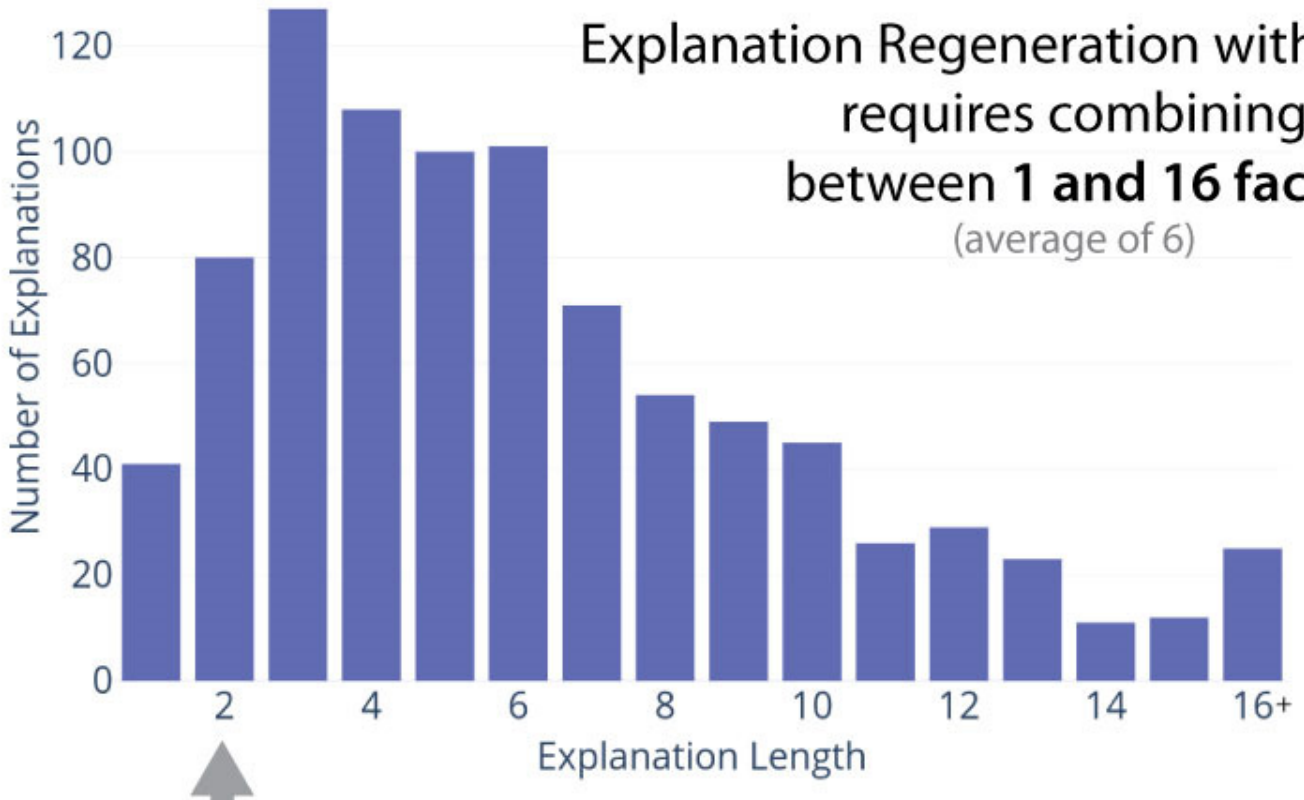
[A] a cat eats a mouse

[*B] a deer eats a leaf

[C] a hawk eats a mouse

[D] a snake eats a rat

Worldtree: How many facts do we have to combine?



Comparison with other multi-hop datasets

HotPotQA (Yang et al., 2018): Combine 2 paragraphs

QASC (Khot et al., 2019): Combine 2 sentences

Why is Multi-Hop Inference Hard?

Semantic Drift: Determining whether two facts can meaningfully combine is hard. When errors happen, inference drifts off context.
(Fried et al., TACL 2015)

Long Inference Chains: Assembling large, detailed explainations requires aggregating up to 16 facts in elementary science. Most methods generally struggle to aggregate more than 2/3 facts.
(Jansen et al., COLING 2016; Jansen et al., LREC 2018)

Evaluation Methods: Recent "multi-hop inference" datasets often do not require multi-hop inference to solve. Papers often do report detailed measurements of multi-hop performance.
(Chen and Durrett, NAACL 2019; Min et al., ACL 2019; Trivedi et al., EMNLP 2020)

Chance Performance: Chance performance on multi-hop inference can either be incredibly easy or impossibly hard, depending on the questions and knowledge resources used.
(Jansen, TextGraphs 2018)

Overview of Team Submissions

Question: A student placed an ice cube on a plate in the sun. Ten minutes later, only water was on the [illegible]. Which process caused the ice cube to change to water?

Answer Candidates: (A) condensation (B) evaporation (C) freezing *(*D) melting*

Gold Explanation from WorldTree Corpus:

Explanatory Role	*Fact (Table Row)*
CENTRAL	melting means changing from a solid into a liquid by adding heat energy
GROUNDING	an ice cube is a kind of solid
GROUNDING	water is a kind of liquid
CENTRAL	water is in the solid state, called ice, for temperatures between -273C and 0 C
LEXGLUE	heat means heat energy
LEXGLUE	adding heat means increasing temperature
CENTRAL	if an object absorbs solar energy then that object will increase in temperature
CENTRAL	if an object is in the sunlight then that object will absorb solar energy
CENTRAL	the sun is a source of (light ; light energy) called sunlight
LEXGLUE	to be in the sun means to be in the sunlight
CENTRAL	melting is a kind of process

Explanation Regeneration Task (Ranking):

Rank	*Gold*	*Fact (Table Row)*
1	*	melting is a kind of process
2		thawing is similar to melting
3		melting is a kind of phase change
4		melting is when solids are heated above their melting point
5		amount of water in a body of water increases by (storms ; rain ; ice melting)
6		an ice cube is a kind of object
7	*	an ice cube is a kind of solid
8		freezing point is similar to melting point
9		melting point is a property of a (substance ; material)
10		glaciers melting has a negative impact on the glaicial environment
...		

Ranks of gold rows: 1, 7, 18, 53, 102, 384, 408, 858, 860, 3778, 3956
Average precision of ranking: 0.149

Question

Answer

Gold Explanation

Model generates a ranked list of KB facts that it beleives are in the gold explanation

Evaluate using Mean Average Precision (MAP)

Performance: Explanation Regeneration

Overall leaderboard performance

2020 Team	Performance	
Baidu PGL	0.603	
LIIR	0.584	
aisys	0.523	
ChiSquareX	0.490	(0.506)*
Red Dragon AI	0.473	(0.561)*
Team IITian	0.452	
AG	0.346	(0.366)*
mler	0.337	
dcandak99	0.325	
Baseline (tf.idf)	0.234	

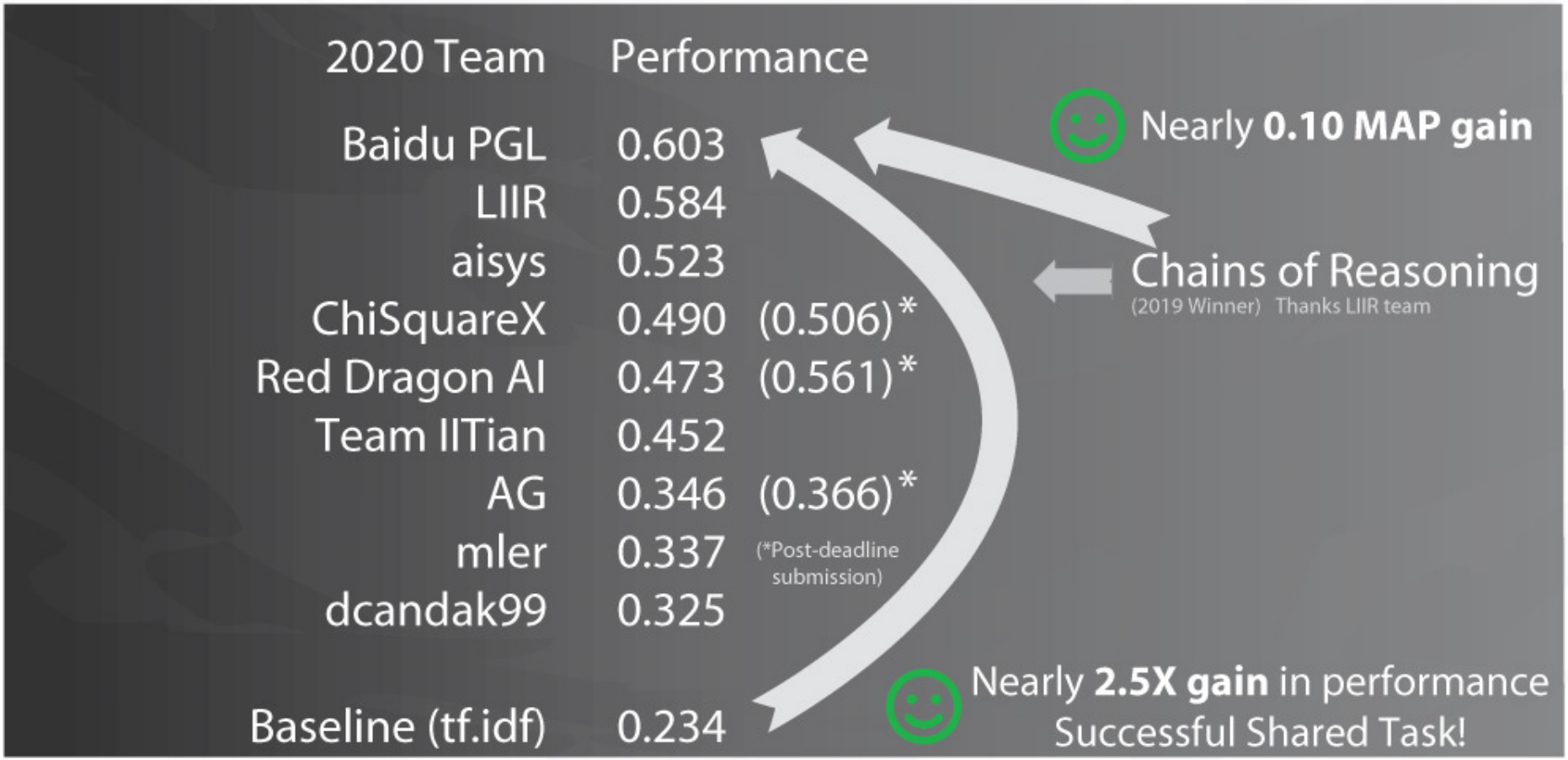
(*Post-deadline submission)

← Chains of Reasoning
(2019 Winner) Thanks LIIR team

Mean Average Precision (MAP)

Performance: Explanation Regeneration

Overall leaderboard performance



Mean Average Precision (MAP)

Shared Task Submissions

Team AG	Integer Linear Programming w/SemanticILP Solver over a shortlist of top-30 BERT facts. *(Gupta et al., TextGraphs 2020)*
Team RDAI	Iterative BM25 (I-BM25) and LSTM-Interleaved Transformer (LIT) *(Chai et al., TextGraphs 2020)*
Team CSX	Reranking using ALBERT, BART, BERT, DistilBERT, ELECTRA, SciBERT, and RoBERTa on top-K tf.idf facts *(Pawate et al., TextGraphs 2020)*
Team LIIR	Auto-regressive ranking problem that iteratively adds facts into a "neighbourhood of visible facts" *(Cartuyvels et al., COLING 2020)*
Team BPGL	ERNIE 2.0 language model used for initial ranking, subquent reranking, followed by GraphSage GNN. *(Li et al., TextGraphs 2020)*

Submission Performance

(Automatic and Manual Evaluations)

Performance: Multi-hop Inference

Question: **Recycling newspapers** is **good** for the **environment** because it:
Answer: helps **conserve resources**

	Explanation Sentences
1-hop, Easy (2 or more shared words with Q/A)	1. **Recycling resources** has a postive impact on the **environment** and the **conservation** of those **resources**
1-hop, Medium (1 shared word with Q/A)	2. A **newspaper** is made of **paper** 3. **Trees** are a kind of **resource** 4. "to be **good** for" means "have a positive impact on"
2+ hop, Hard (no shared words with Q/A) **"Multi-hop Inference"**	5. **Trees** are a source of **paper**

Performance: Multi-hop Inference

Explanation Reconstuction performance broken down by how challenging facts are to find (1-hop to 2+ hop).

	Baseline	Team				
	tf.idf	AG	RDAI	CSX	LIIR	BPGL
1-hop, Easy (2 or more shared words with Q/A)	0.32	0.47	0.65	0.59	0.66	**0.69**

Performance: Multi-hop Inference

Explanation Reconstuction performance broken down by how challenging facts are to find (1-hop to 2+ hop).

	Baseline	Team				
	tf.idf	AG	RDAI	CSX	LIIR	BPGL
1-hop, Easy (2 or more shared words with Q/A)	0.32	0.47	0.65	0.59	0.66	**0.69**
1-hop, Medium (1 shared word with Q/A)	0.07	0.25	0.40	0.35	**0.49**	0.48

Performance: Multi-hop Inference

Explanation Reconstuction performance broken down by how challenging facts are to find (1-hop to 2+ hop).

	Baseline	Team				
	tf.idf	AG	RDAI	CSX	LIIR	BPGL
1-hop, Easy (2 or more shared words with Q/A)	0.32	0.47	0.65	0.59	0.66	**0.69**
1-hop, Medium (1 shared word with Q/A)	0.07	0.25	0.40	0.35	**0.49**	0.48
2+ hop, Hard (no shared words with Q/A) **"Multi-hop Inference"**	0.00	0.19	0.16	*0.07*	**0.31**	0.29

Overview

Shared task a success!
Explanation regeneration performance is
~0.10 MAP better than 2019 Shared Task.

Large relative performance improvements,
but absolute performance still has plenty of room to grow.

Multi-hop inference evaluation is becoming very challenging.

Downstream (leaderboard) evaluation: misses the
full picture, and is easily gamed by strong retrieval modules
(like large language models).
(Chen and Durrett, NAACL 2019; Min et al., ACL 2019; Trivedi et al., EMNLP 2020)

Detailed measures of multi-hop performance: We have them,
but they're rarely used/reported in papers.

Thank You!

Shared Task Participant Kit:
https://github.com/cognitiveailab/tg2020task

Worldtree V2 Explanation Corpus available at:
cognitiveai.org/explanationbank

Thanks to:
Zhengnan Xie
Jaycie Martin
Elizabeth Wainwright
Steven Marmorstein
Peter Clark
National Science Foundation (PJ)

WORLDTREE DEVELOPMENT SUPPORTED BY: