Standardized Tests as benchmarks for Artificial Intelligence?

Oct 31, 2018

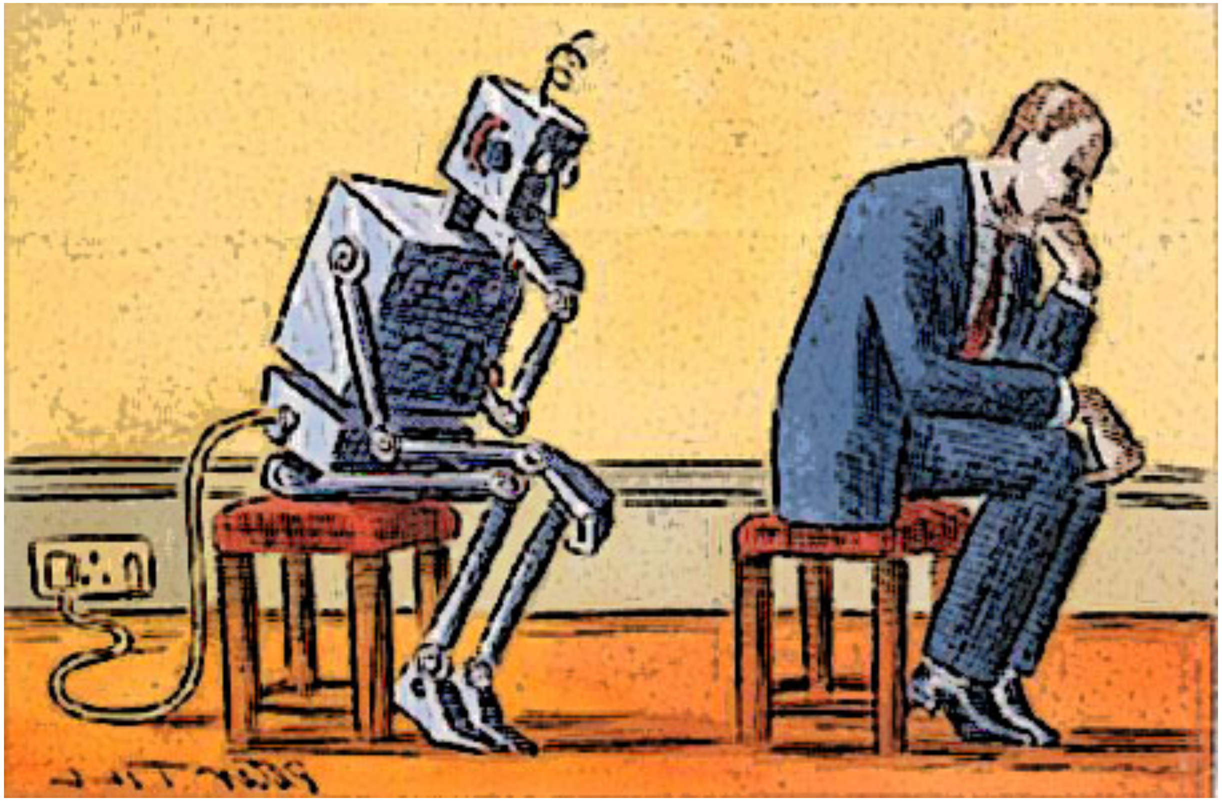
Mrinmaya Sachan[1] Minjoon Seo[2] Hannaneh Hajishirzi[2] Eric P. Xing[1]

[1]Carnegie Mellon University
{mrinmays,epxing}@cs.cmu.edu

[2]University of Washington
{minjoon,hannaneh}@cs.washington.edu

Machine Intelligence

Machine Intelligence



"It would someday be possible for a sufficiently advanced computer to think and to have some form of consciousness"

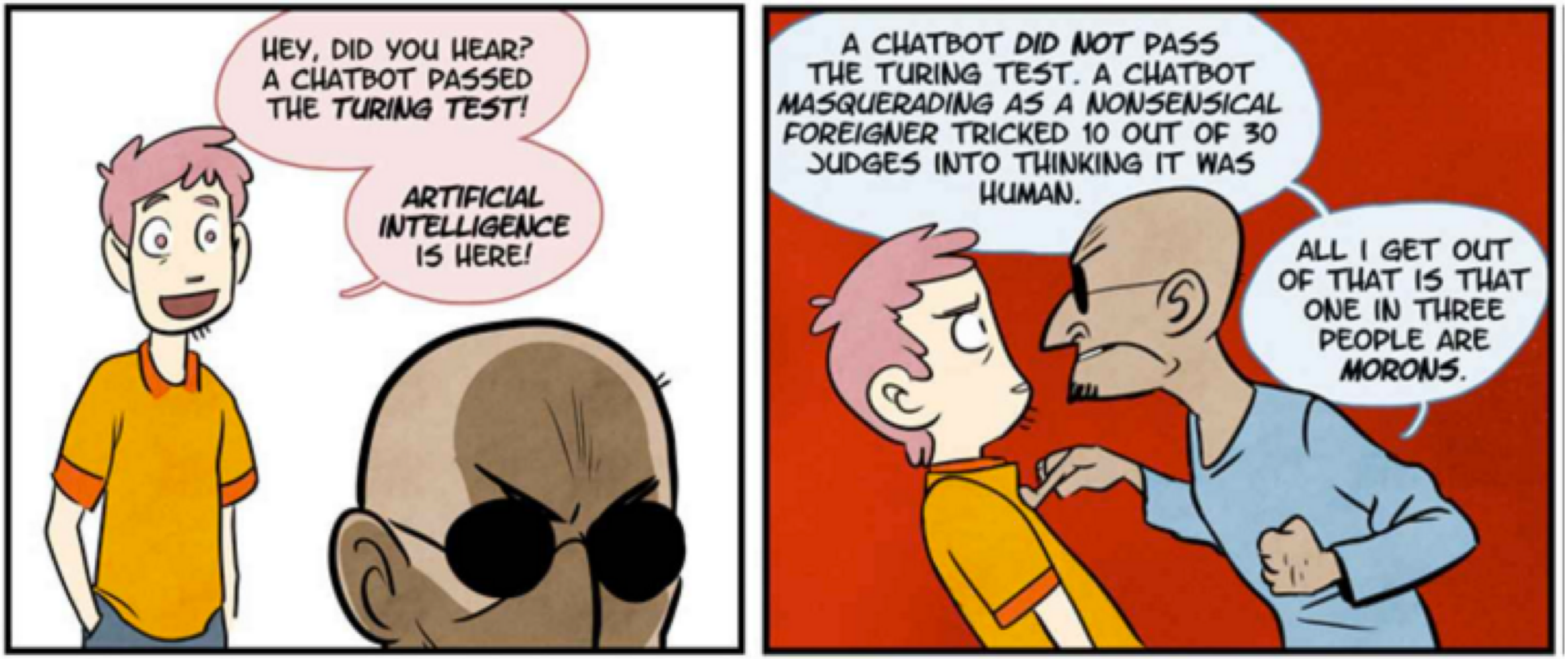
-- Computing Machinery and Intelligence, Mind 1950.

How do we measure progress? What tasks should drive the field?

Turing Test

Oops!

Chatbot



HEY, DID YOU HEAR? A CHATBOT PASSED THE TURING TEST!
ARTIFICIAL INTELLIGENCE IS HERE!
A CHATBOT DID NOT PASS THE TURING TEST. A CHATBOT MASQUERADING AS A NONSENSICAL FOREIGNER TRICKED 10 OUT OF 30 JUDGES INTO THINKING IT WAS HUMAN.
ALL I GET OUT OF THAT IS THAT ONE IN THREE PEOPLE ARE MORONS.

Standardized Tests as drivers for Artificial Intelligence

[Brachman 2005, Levesqeue 2010, Clark 2014]

Standardized Tests

Why Standardized Tests

- *Easily accessible*
- *Easily measurable*
- *Do not cover all aspects*

Standardized Tests as Benchmarks for AI: Limitations

- Aspects of intelligence that are challenging for AI systems are very different from aspects of intelligence that are challenging to humans.
  - Standardized tests do not test knowledge that is obvious for people.

Nevertheless, passing standardized tests still requires better language/visual understanding and reasoning capabilities than those demonstrated by AI systems

- Drivers for progress in AI

(Davis 2018)

Outline

- *Machine Reading for Question Answering:*
  Reading Comprehension
  - *Feature Driven Models*
    - *MCTest*
  - *Deep Learning Models*
    - *WikiQA*
    - *CNN & DailyMail*
    - *SQUAD*
    - *Etc.*

  Beyond Reading Comprehensions
    - Elementary-level Science Exams
    - *Diagram QA*
    - *Textbook QA*

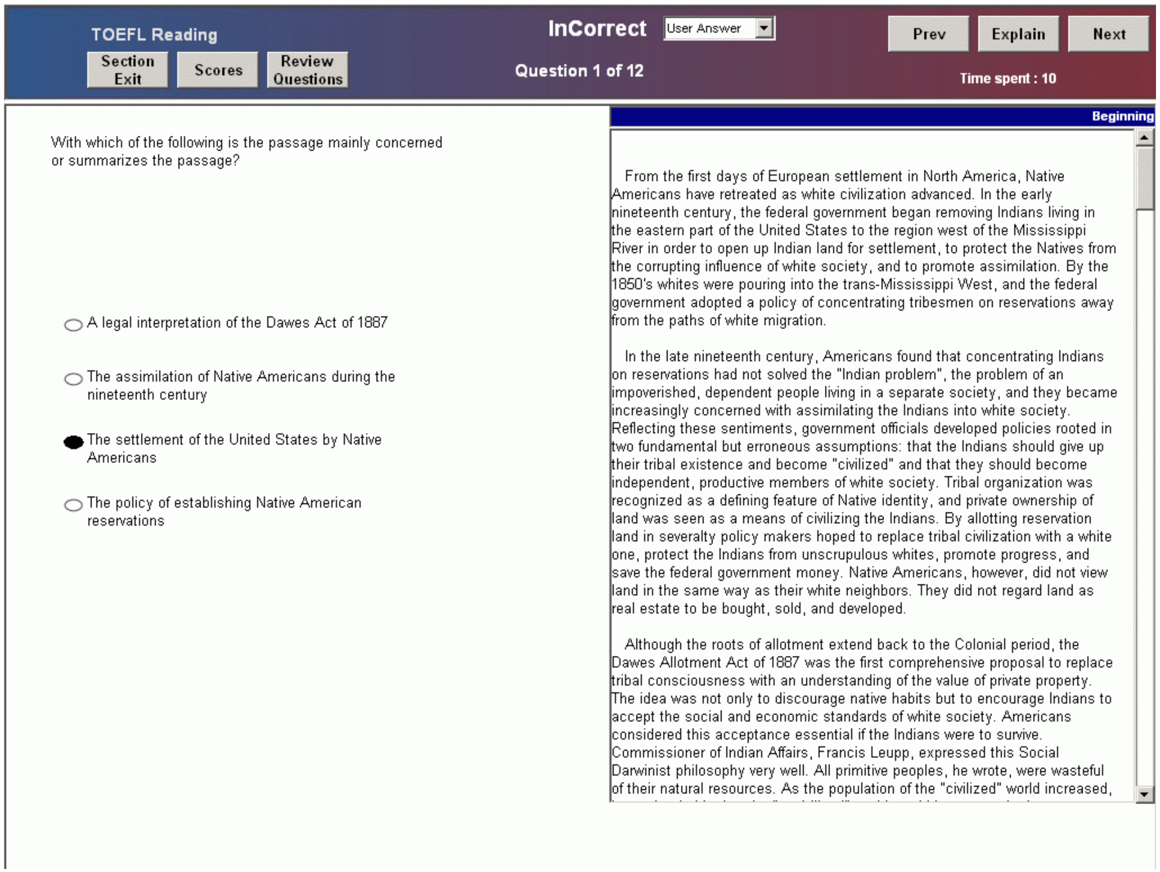
- *Mathematical Question Answering:*
  Advanced Math and Science Problems
  - *Algebra Word Problems*
  - *Geometry Problems*
  - *Newtonian Physics Problems*

Outline

- *Machine Reading for Question Answering:*
  Reading Comprehension
  - *Feature Driven Models*
    - *MCTest*
  - *Deep Learning Models*
    - *WikiQA*
    - *CNN & DailyMail*
    - *SQUAD*
    - *Etc.*

  Beyond Reading Comprehensions
    - Elementary-level Science Exams
    - *Diagram QA*
    - *Textbook QA*
- *Mathematical Question Answering:*
  Advanced Math and Science Problems
  - *Algebra Word Problems*
  - *Geometry Problems*
  - *Newtonian Physics Problems*

Reading Comprehension for Humans



Excerpt from https://learn.eazycoach.com/toefl-decoded/

Reading Comprehension

- Read a piece of text and answer questions
  - Often Multiple-choice
  - Timed!

Note: There are issues with both the aforementioned assumptions if we wish to use standardized tests to contrast machine and human intelligence!

RC for Machines
A Historical NLP Perspective

- Charniak's PhD thesis (1972) - ***Background* model** to answer questions about children's stories.
- Hirschmann et al. 1999 showed that **bag of words pattern matching with some additional linguistic processing** could achieve **40% accuracy** for picking the sentence that best answers "who / what / when / where / why" questions on the **Remedia** dataset. Results on this dataset were later improved upon by Grois and Wilkins (2005); Harabagiu et al. (2003); Wellner et al. (2006).
- Riloff et. al 2000 developed a rule-based system, *Quarc*, which used similar lexical and semantic clues in the question and the story to answer questions about it. On RC tests given to children in grades 3-6, **Quarc** achieved an accuracy of around **40%**.
- Breck et. al 2001 collected 75 stories from Canadian Broadcasting Corporation's web site for children and generated 650 questions where each question was answered by a sentence in the text.
- Leidner et. al 2003 used the **CBC4kids** data and added layers of annotation (such as semantic and POS tags), thus measuring QA performance as a function of question difficulty.

Machine Comprehension

MCTest (Richardson et al. 2013)

- Freely available crowd-sourced set of 500 stories and associated questions:
  - 4 questions per story, 4 answer choices per question
- Open-domain yet restricted to concepts and words that a 7 year old is expected to understand
- *As the stories are fictional, answers are typically in the passage itself.*
  - *This requires the system to deeply "understand" the stories rather than using IR methods or redundancy of the web.*

MCTest

Once upon a time there a little girl named Ana. Ana was a smart girl. Everyone in Ana’s school knew and liked her very much. She had a big dream of becoming spelling bee winner. Ana studied very hard to be the best she could be at spelling. Ana's best friend would help her study every day after school. By the time the spelling bee arrived Ana and her best friend were sure she would win. There were ten students in the spelling bee. This made Ana very nervous, but when she looked out and saw her dad cheering her on she knew she could do it. The spelling bee had five rounds and Ana made it through them all. She was now in the finals. During the final round James, the boy she was in the finals with, was given a really hard word and he spelled it wrong. All Ana had to do was spell this last word and she would be the winner. Ana stepped to the microphone, thought really hard and spelled the word. She waited and finally her teacher said "That is correct". Ana had won the spelling bee. Ana was so happy. She won a trophy. Ana also won a big yellow ribbon. The whole school was also happy, and everyone clapped for her. The whole school went outside. They had a picnic to celebrate Ana winning.

Q1: What made Ana very nervous?

A) The other ten students

B) Her best friend

C) The bright lights

D) The big stage

MCTest

Once upon a time there a little girl named Ana. Ana was a smart girl. Everyone in Ana’s school knew and liked her very much. She had a big dream of becoming spelling bee winner. Ana studied very hard to be the best she could be at spelling. Ana's best friend would help her study every day after school. By the time the spelling bee arrived Ana and her best friend were sure she would win. There were ten students in the spelling bee. This made Ana very nervous, but when she looked out and saw her dad cheering her on she knew she could do it. The spelling bee had five rounds and Ana made it through them all. She was now in the finals. During the final round James, the boy she was in the finals with, was given a really hard word and he spelled it wrong. All Ana had to do was spell this last word and she would be the winner. Ana stepped to the microphone, thought really hard and spelled the word. She waited and finally her teacher said "That is correct". Ana had won the spelling bee. Ana was so happy. She won a trophy. Ana also won a big yellow ribbon. The whole school was also happy, and everyone clapped for her. The whole school went outside. They had a picnic to celebrate Ana winning.

Q1: What made Ana very nervous?

A) The other ten students ✓
B) Her best friend
C) The bright lights
D) The big stage

MCTest

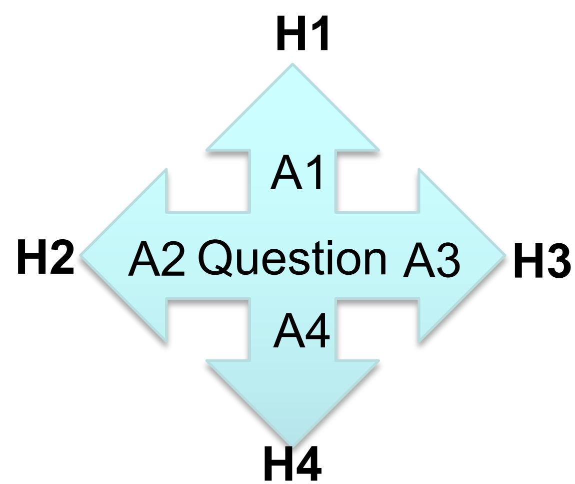
Small Dataset

- Simple Machine Learning Models with hand-engineered features

Core Idea

- Convert each Question/answer-choice pair to a hypothesis statement.

Text (T): ...
Question (Q): ...
(a) A1 (b) A2
(c) A3 (d) A4



- Find which hypothesis statement is best "*entailed*" by the passage

Degree of Entailment

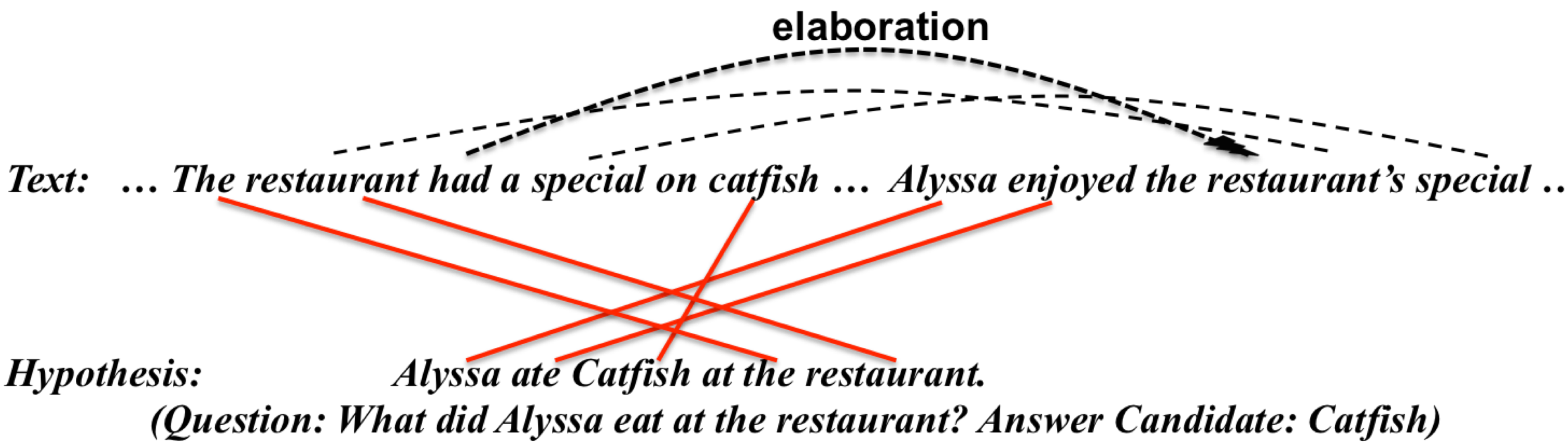
- The models mainly differ in how they measure entailment and feature engineering:
  - Sachan et al. 2015, 2016 - Answer-entailing structures – syntax, semantics and discourse - AMR
  - Wang et al. 2015 – syntax, frames and semantic features.
  - Narasimhan et al. 2015 – Discourse features.

Similar

- Challenging for Deep Models because of small data:
  - Trischler et al. 2016 use **multiple shallow NNs to compare the question and answer candidates to the text** using **several distinct perspectives which mimic features**: word-by-word, sequential and dependency view.

Answer-Entailing Structures

- Alignment based approach (Sachan et al. 2015)
  - Align an answer hypothesis to multiple sentences in the text (not necessarily contiguous)
    - Document Structure – Entity and Event Co-reference, Rhetorical Structure Theory (Mann and Thompson'88)



  - Multi-task Learning

Language Representation

- AMR as a semantic representation
  - Abstract Meaning Representation (Banarescu et al. 2013) captures many aspects of meaning in a single simple data structure:
    - PropBank style semantic roles
    - Within-sentence coreference
    - Named entities
    - Notion of types, modality, negation, quantification, etc.

(Sachan et al. 2016)

Graph Containment Solution

Graph Containment Solution

Hypothesis: Sammy is the name of Katie's dog.
Question: What is the name of Katie's dog. Answer: Sammy

Graph Containment Solution

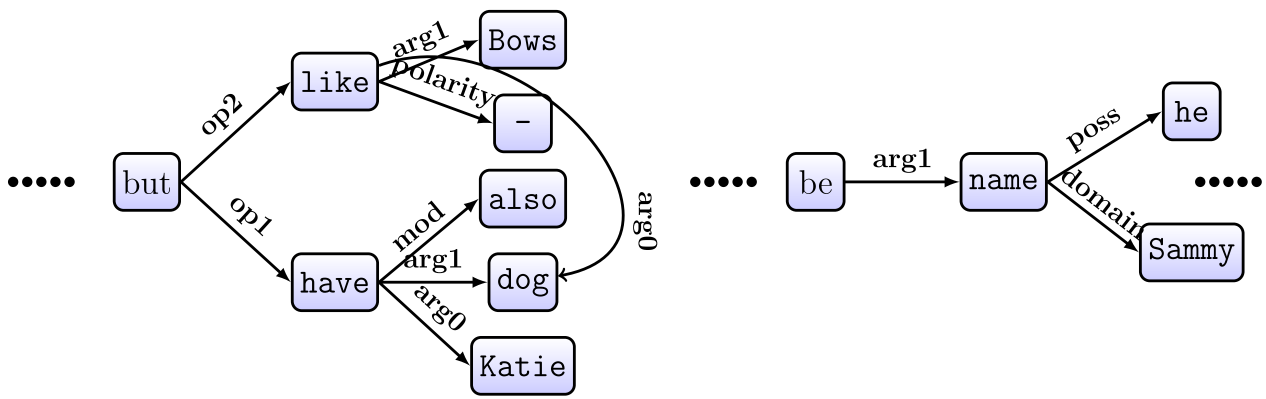
Text: ... Katie also has a dog, but he does not like Bows. ... His name is Sammy. ...

Hypothesis: Sammy is the name of Katie's dog.
Question: What is the name of Katie's dog. Answer: Sammy

Graph Containment Solution

Text: ... Katie also has a dog, but he does not like Bows. ... His name is Sammy. ...

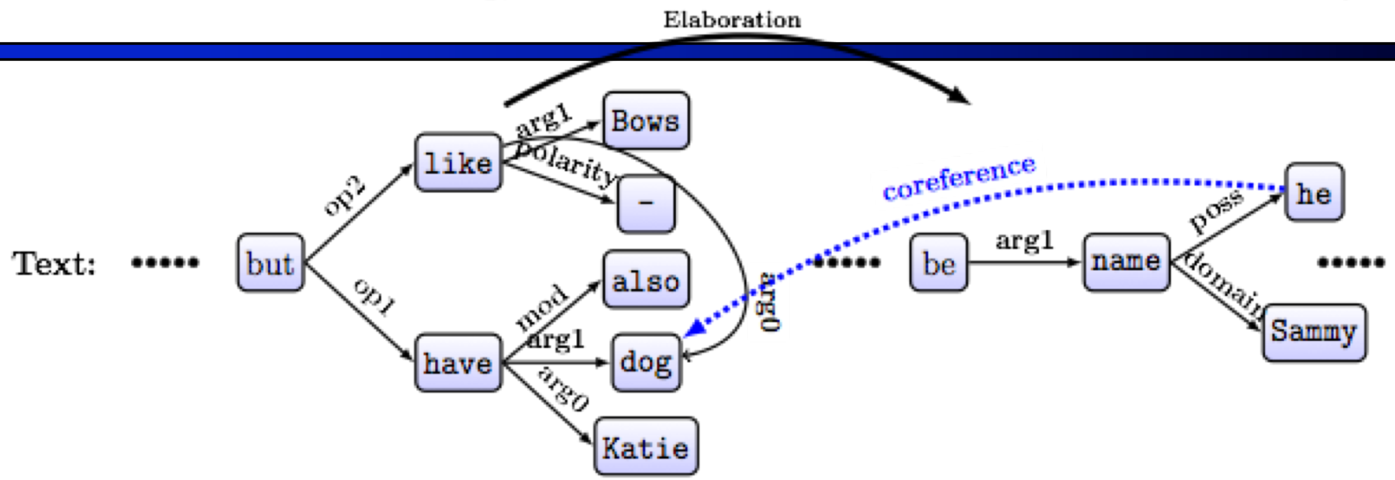
Text:



Hypothesis: Sammy is the name of Katie's dog.
Question: What is the name of Katie's dog. Answer: Sammy

Graph Containment Solution

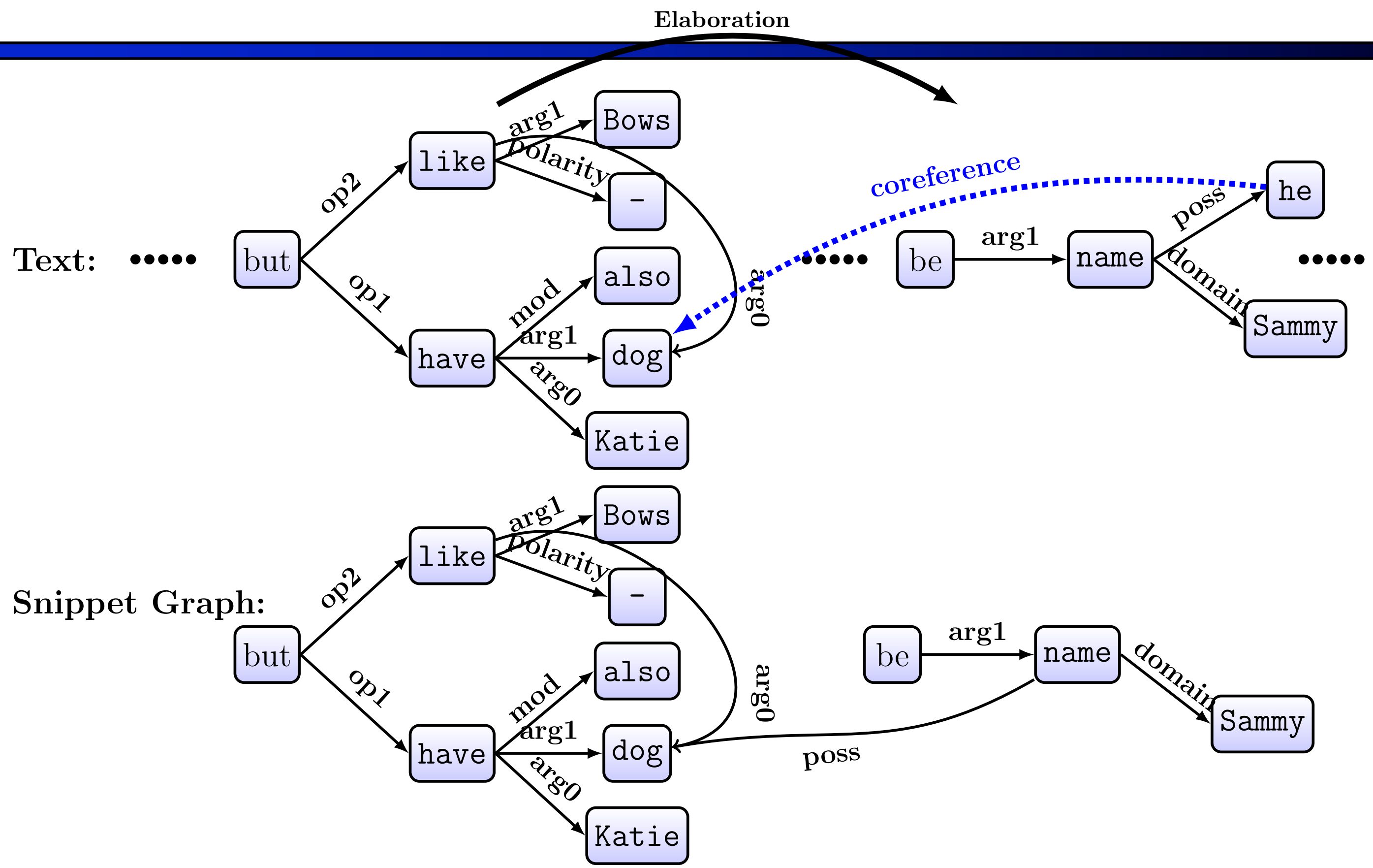
Text: ... Katie also has a dog, but he does not like Bows. ... His name is Sammy. ...



Hypothesis: Sammy is the name of Katie's dog.
Question: What is the name of Katie's dog. Answer: Sammy

Graph Containment Solution

Text: ... Katie also has a dog, but he does not like Bows. ... His name is Sammy. ...

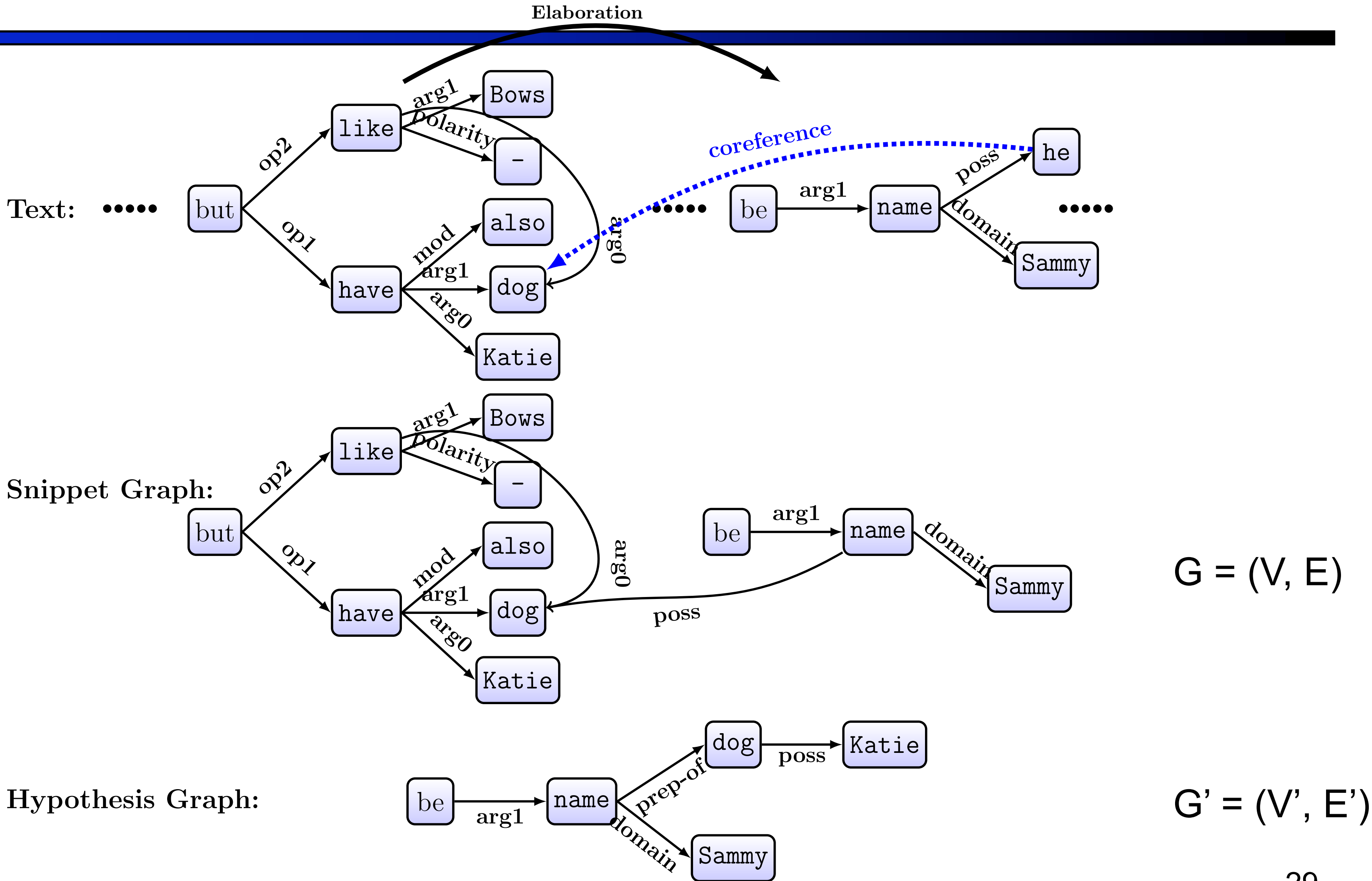


G = (V, E)

Hypothesis: Sammy is the name of Katie's dog.
Question: What is the name of Katie's dog. Answer: Sammy

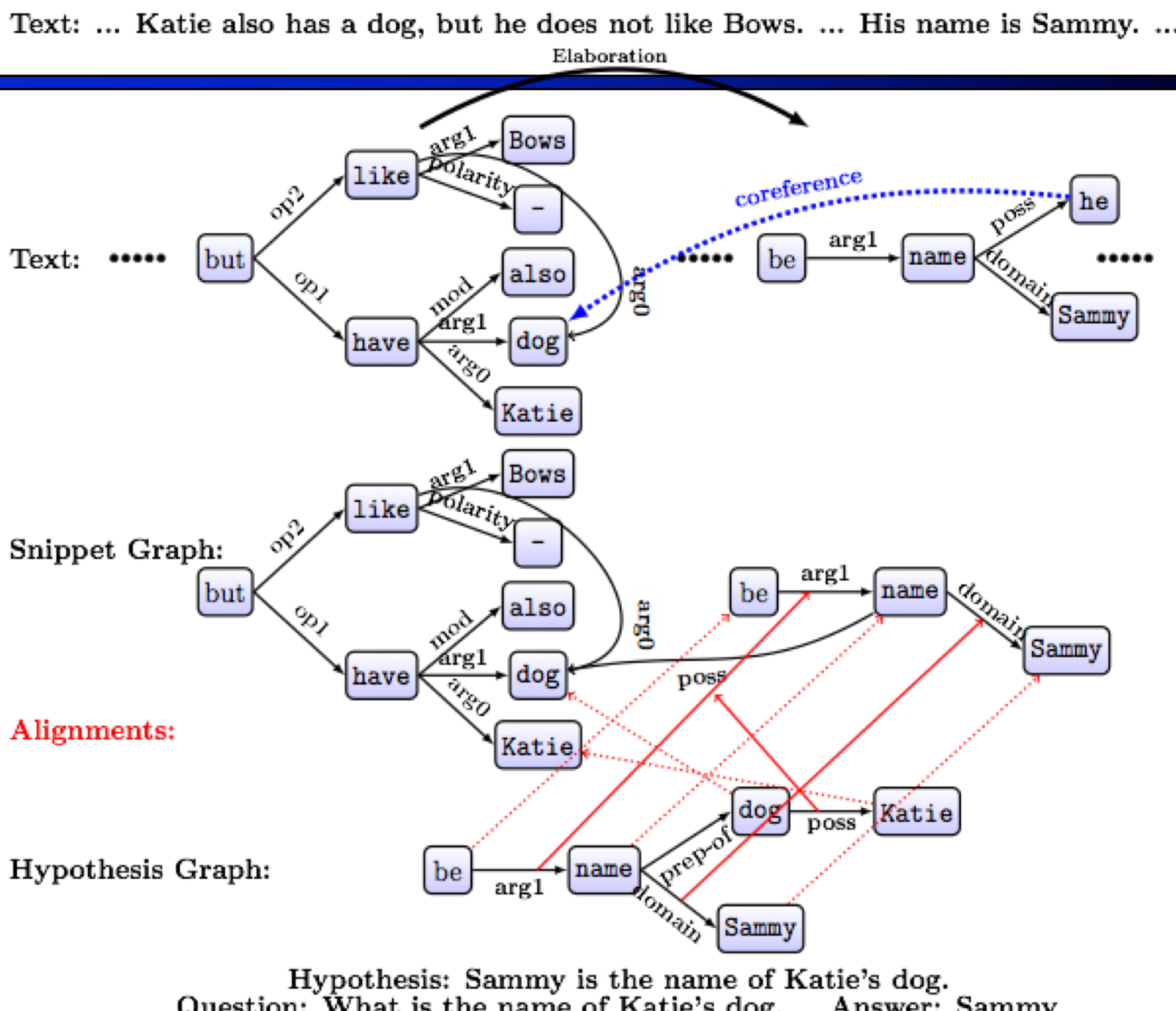
Graph Containment Solution

Text: ... Katie also has a dog, but he does not like Bows. ... His name is Sammy. ...



Hypothesis: Sammy is the name of Katie's dog.
Question: What is the name of Katie's dog. **Answer: Sammy**

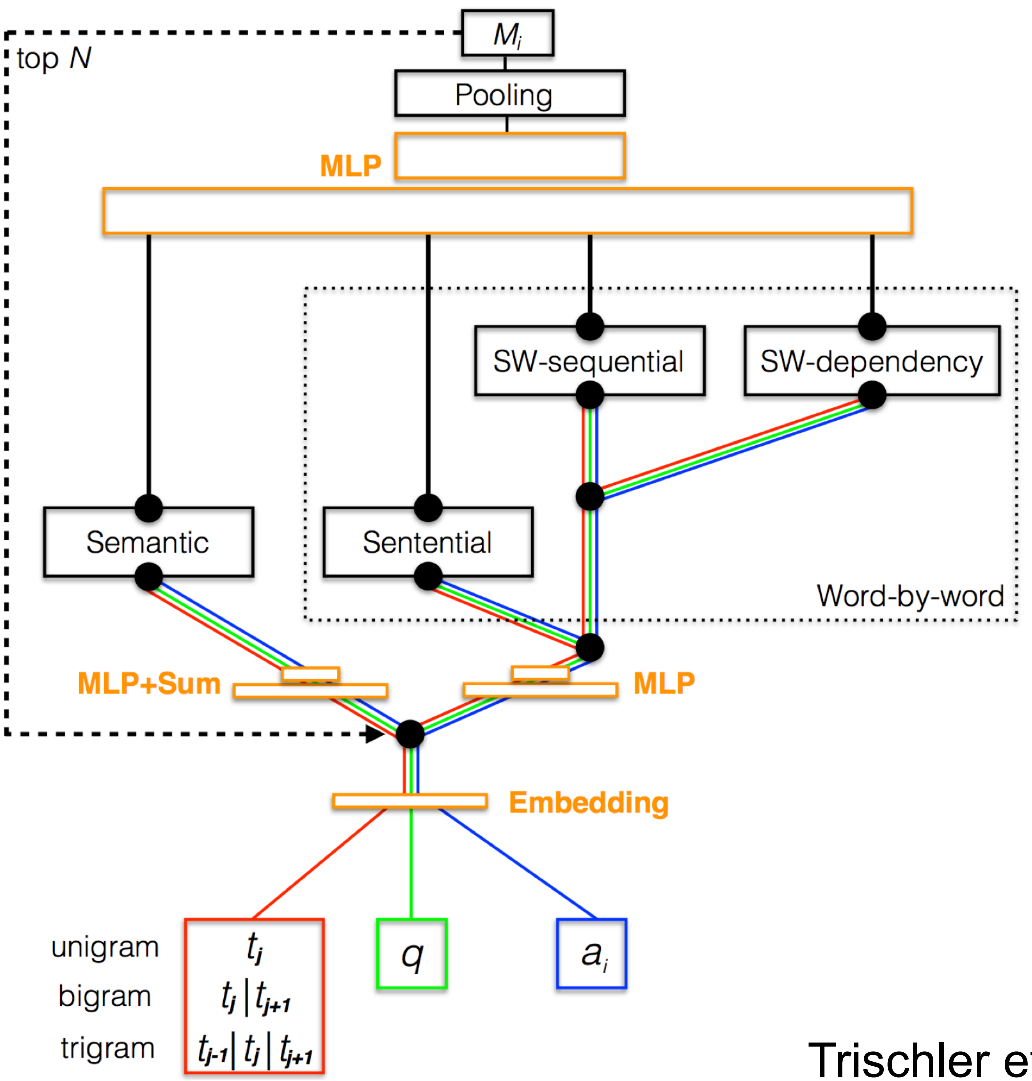
Graph Containment Solution



$G = (V, E)$

$G' = (V', E')$

Parallel Hierarchical Neural Network Model



Trischler et al. ACL 2016

State Of The Art on MCTest

Approach	Accuracy (%)
Sliding Window	54.28
RTE	55.01
Strong Lexical Matching Baseline (Smith et al. 2015)	65.43
Discourse features (Narasimhan et al. 2015)	63.75
Answer-entailing structures (Sachan et al. 2015)	67.83
Syntax, frames and semantic features (Wang et al. 2015)	69.94
Answer-entailing structures - AMR (Sachan et al. 2016)	70.33
Simple neural networks - not deep (Trischler et al. 2016)	71.00

Outline

- *Machine Reading for Question Answering:*
  Reading Comprehension
  - *Feature Driven Models*
    - *MCTest*
  - *Deep Learning Models*
    - *WikiQA*
    - *CNN & DailyMail*
    - *SQUAD*
    - *Etc.*

  Beyond Reading Comprehensions
    - Elementary-level Science Exams
    - *Diagram QA*
    - *Textbook QA*
- *Mathematical Question Answering:*
  Advanced Math and Science Problems
  - *Algebra Word Problems*
  - *Geometry Problems*
  - *Newtonian Physics Problems*

Larger datasets for neural nets

- WikiQA (2015), CNN/Daily Mail (2015), SQuAD (2016), TriviaQA (2017), LAnguage Modeling Broadened to Account for Disclourse Aspects (LAMBADA), QuizBowl questions, NewsQA dataset, MS MARCO …

WikiQA (2015)

- Similar to TREC QA (est. 1999)
- 3k questions
- Answering for real user queries
  - “When was Barack Obama born?”
- Given a question, select the *sentence* (among 10) that best answers the question
- Is this really a “reading comprehension”?
  - Or is it “sentence retrieval”?

CNN & DailyMail QA (2015)

- News article & summary pair
- The task is to predict a masked entity in the summary (cloze test)
- Requires understanding of news article?

Context

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." …

Query

Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer

Oisin Tymon

A Thorough Examination of the CNN/Daily Mail RC Task

(Chen et al., ACL 2016)

- Simple, carefully designed systems can obtain SOTA performance. The dataset is easy!
- Distributed representations are effective at recognizing paraphrases.
- Best systems more have the nature of single-sentence relation extraction systems than larger-discourse-context text understanding.
- Most systems proposed are close to the ceiling of performance for single sentence and unambiguous cases.
- Prospects for getting the final 20% of questions correct are poor, since most of them involve issues in the data preparation.

SQuAD (Rajpurkar et al., 2016)

- Gained a lot of popularity
  - First **massive** and **manually** turked data; deep learning in full effect
  - Factual questions: natural extension to real open-domain QA systems (Chen et al., 2017)
  - Easy to use (small size context, Wikipedia-based, etc.)

SQuAD

Second Epistle to the Corinthians The Second Epistle to the Corinthians, often referred to as Second Corinthians (and written as 2 Corinthians), is the eighth book of the New Testament of the Bible. Paul the Apostle and “Timothy our brother” wrote this epistle to “the church of God which is at Corinth, with all the saints which are in all Achaia”.

Who wrote second Corinthians?

SQuAD

Second Epistle to the Corinthians The Second Epistle to the Corinthians, often referred to as Second Corinthians (and written as 2 Corinthians), is the eighth book of the New Testament of the Bible. <u>Paul the Apostle and “Timothy our brother”</u> wrote this epistle to “the church of God which is at Corinth, with all the saints which are in all Achaia”.

Who wrote second Corinthians?

100,000+

A lot of models!

SQuAD1.1 Leaderboard

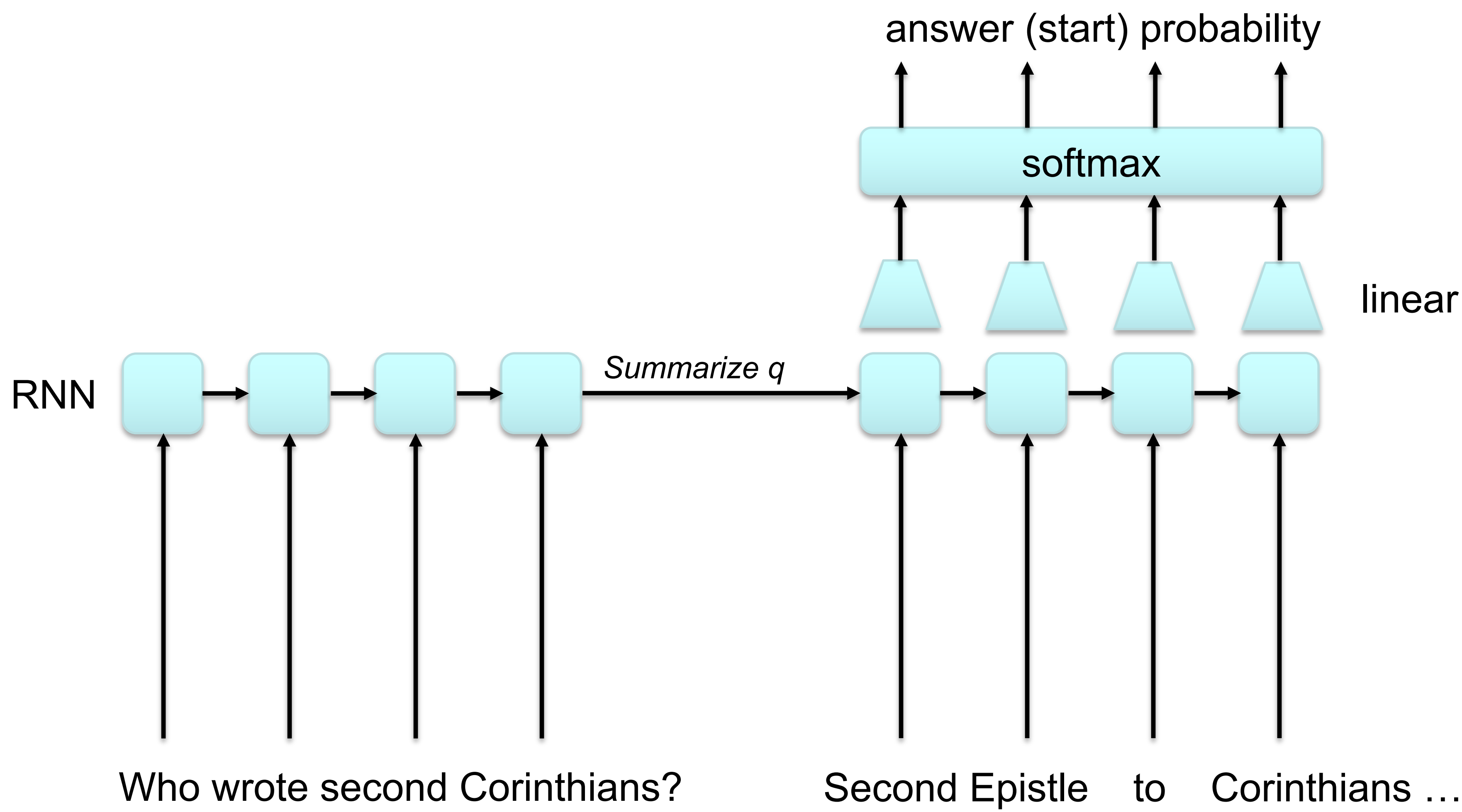
Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google A.I.*	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google A.I.*	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133
5 Jun 20, 2018	MARS (ensemble) *YUANFUDAO research NLP*	83.982	89.796
6 Sep 01, 2018	MARS (single model) *YUANFUDAO research NLP*	83.185	89.547
7 Jan 22, 2018	Hybrid AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.482	89.281

Basic components of models

- Sequential Model – F1 70%
  - RNNs (LSTM, GRU)
- +Cross-Attention – F1 77% (**+7%**)
- +Self-Attention – F1 82% (**+5%**)
- +Transfer Learning – F1 86% (**+4%**)
  - Data Augmentation (back translation via MT)
  - Contextualized Vectors (CoVe, ELMo)
- +Other Tricks – F1 90% (**+4%**)
  - Ensemble, Distillation
  - Sparse features (Chen et al., 2017)
  - Finetuning with RL (Xu et al., 2017)
- Just Attention – F1 93% (**+3%**)
  - (Dublin et al., 2018)

Neural Sequential Model



Hermann et al. Teaching machines to read and comprehend. NIPS 2015.

SQuAD Baselines

Date	Model	F1	EM
May 2016	Feature-based	~50%	~40%
-	Neural Sequential	~70%	~60%

Issues with Sequential Model

- Question needs to be summarized into a *fixed-size vector*

"You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!" *-Ray Mooney*

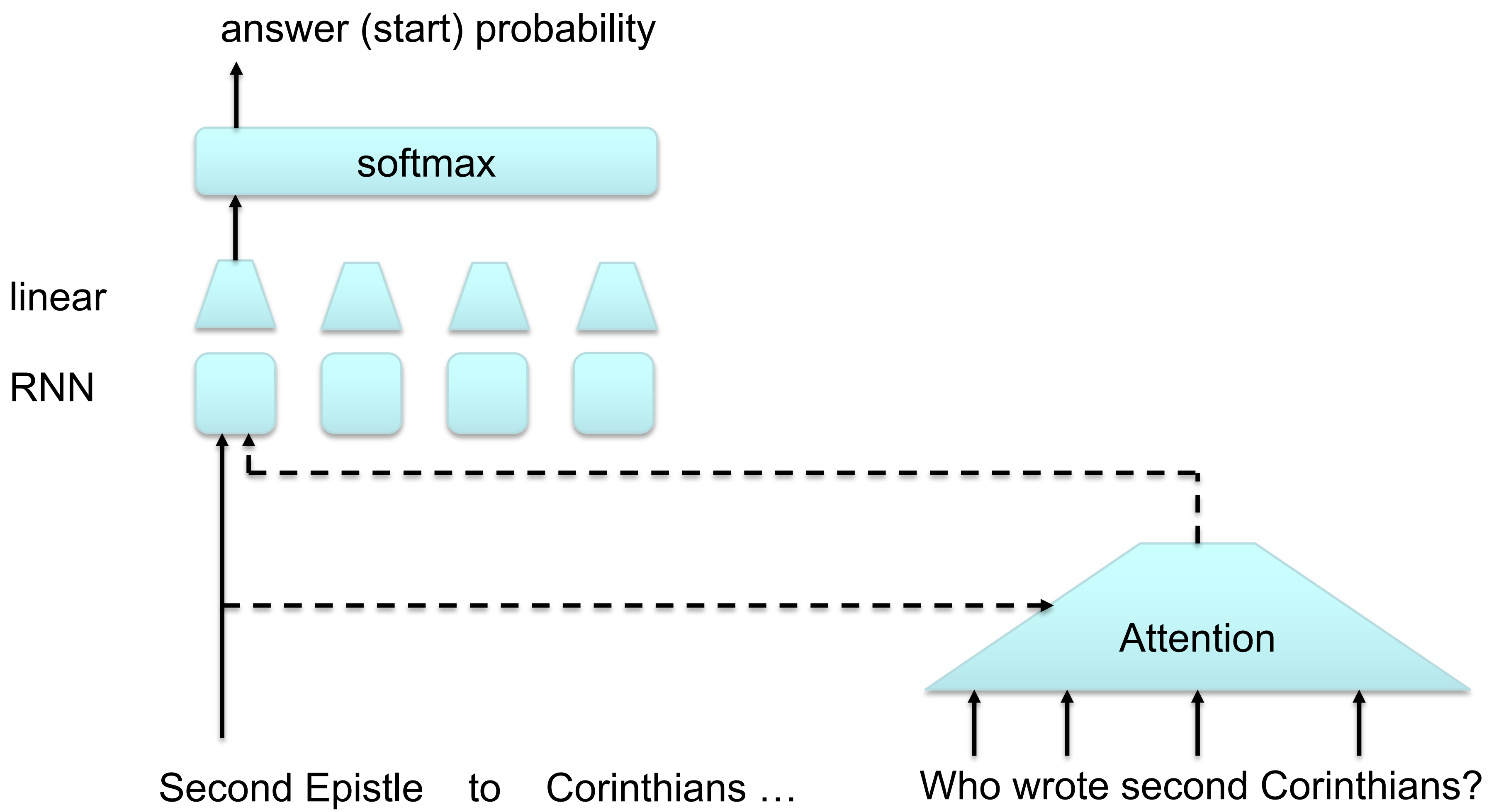
Attention Mechanism

- A mechanism to dynamically summarize a sequence of vectors
- Many variants exist:
  - Concat and linear (Bhadanau et al., 2015)
    - Memory intensive
  - Bilinear (Luong et al., 2015)
  - Transformer-style (Vaswani et al., 2017)
    - Memory efficient

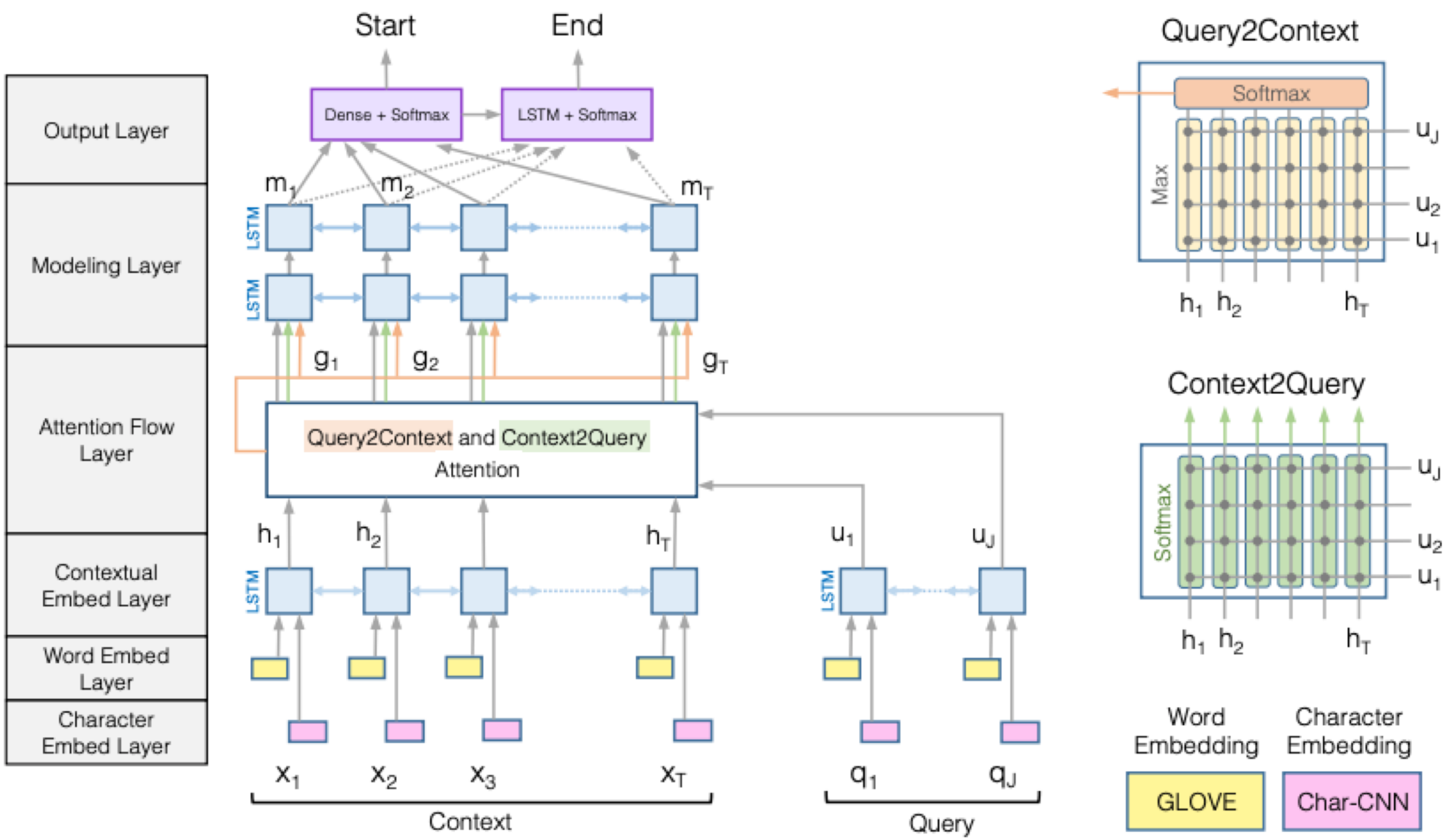
Cross-Attention

- Common in early SQuAD models
- Choose what part of the question to look at for each context
  - **match-LSTM** (Wang et al., 2017)
- Often times the other direction is also considered
  - **bi-attention** (Seo et al., 2017)
  - **co-attention** (Xiong et al., 2017)

Cross-Attention Model



Cross-Attention Example



Seo et al. Bidirectional attention flow for machine comprehension. ICLR 2017.

SQuAD Leaderboard

Date	Model	F1	EM
Aug 2016	+Cross-Attention	75~78%	66~68%

Cross-Attention Demo (BiDAF)

Bi-directional Attention Flow Demo

for Stanford Question Answering Dataset (SQuAD)

Direction : Select a paragraph and write your own question. The answer is always a subphrase of the paragraph - remember it when you ask a question!

Select Paragraph

[05] Teacher

Paragraph

The role of teacher is often formal and ongoing, carried out at a school or other place of formal education. In many countries, a person who wishes to become a teacher must first obtain specified professional qualifications or credentials from a university or college. These professional qualifications may include the study of pedagogy, the science of teaching. Teachers, like other professionals, may have to continue their education after they qualify, a process known as continuing professional development. Teachers may use a lesson plan to facilitate student learning, providing a course of study which is called the curriculum.

Question

Where do most teachers get their credentials from?

new question!

Answer

a university or college

Reference : Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi. "Bidirectional Attention Flow for Machine Comprehension" [link]

Demo by : Sewon Min

http://allgood.cs.washington.edu:1995/

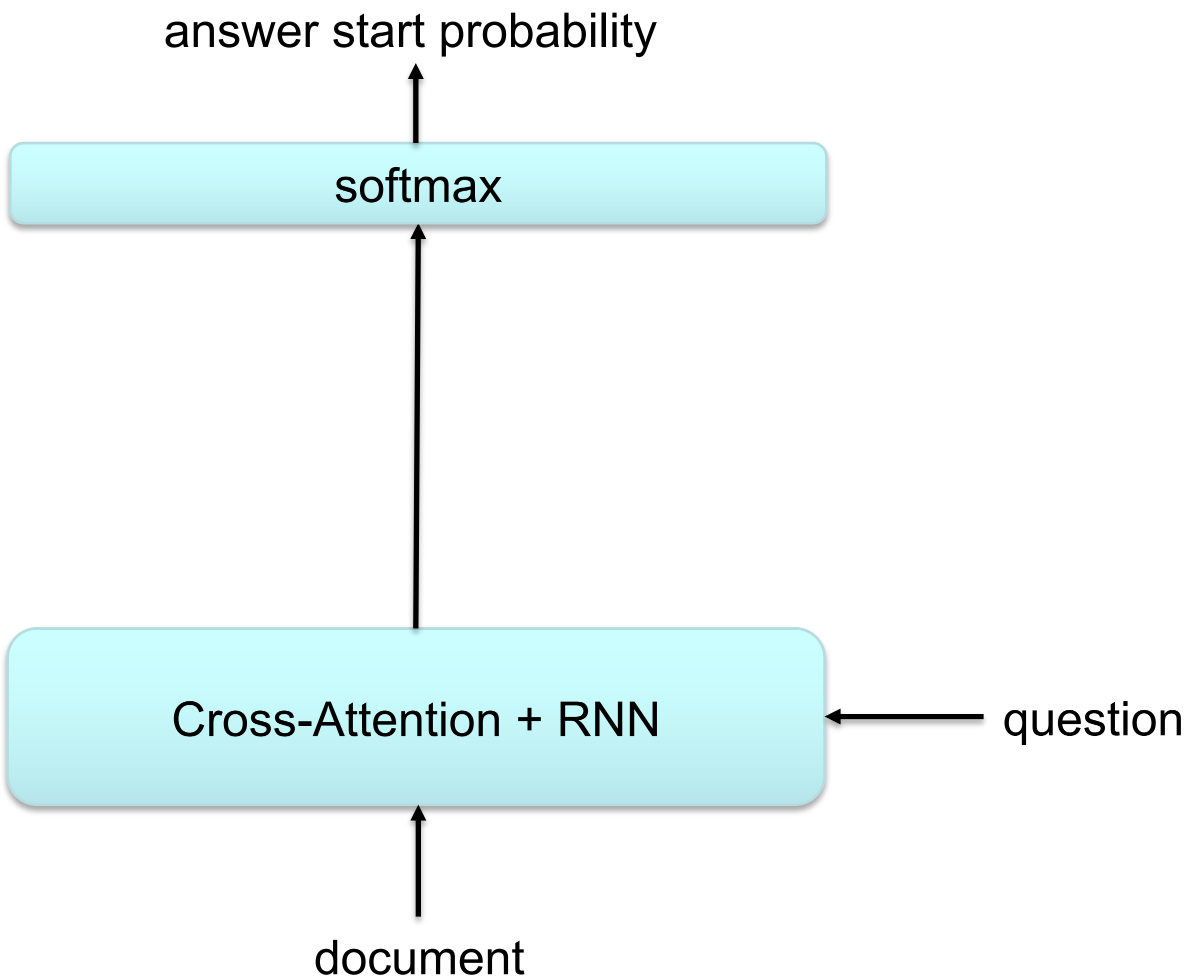
Issues with Cross-Attention

- Sequential models are not great for long-term dependency
  - even with gating mechanism
- Context (document) is long (200+ words)
  - Coreference?
  - Long sentence?

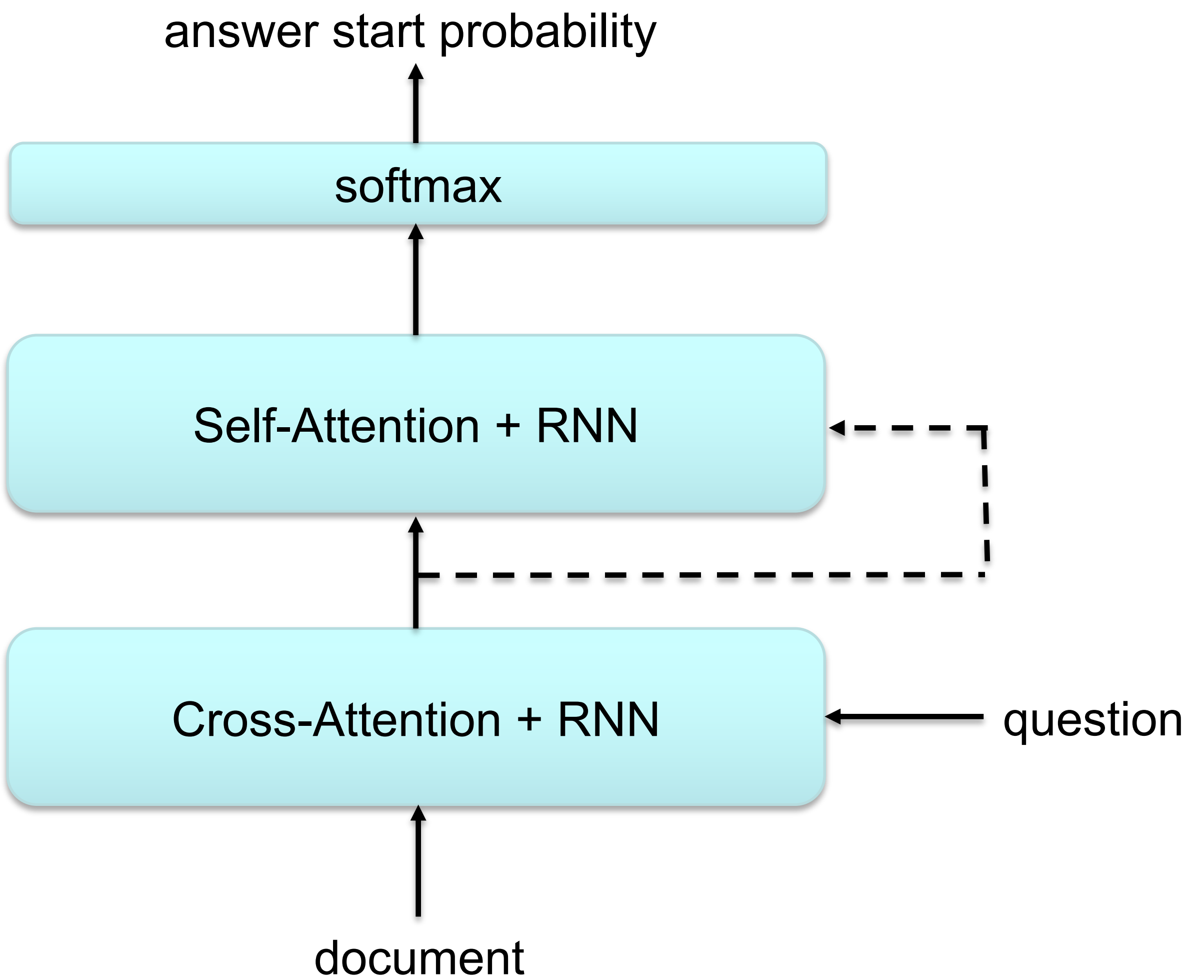
Self-Attention

- ”Attend” on itself!
- Self-attention allows direct access to distant words
- Usually on top of cross-attention

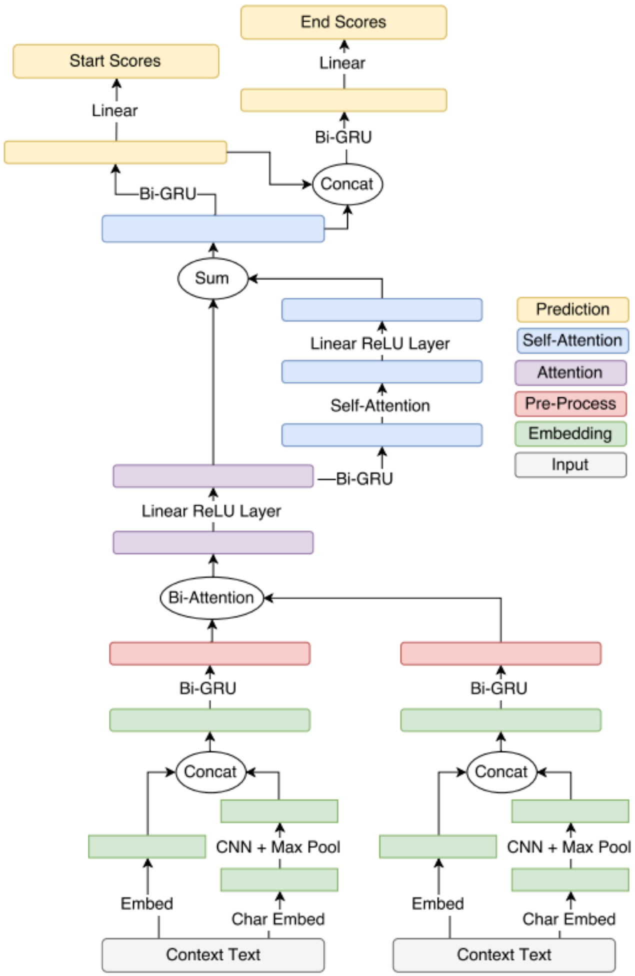
Self-Attention



Self-Attention



Self-Attention Example



Clark and Gardner. Simple and effective multi-paragraph reading comprehension. 2017.

SQuAD Leaderboard

Date	Model	F1	EM
Aug 2016	+Cross-Attention	75~78%	66~68%
Mar 2017	+Self-Attention	80~82%	71-73%

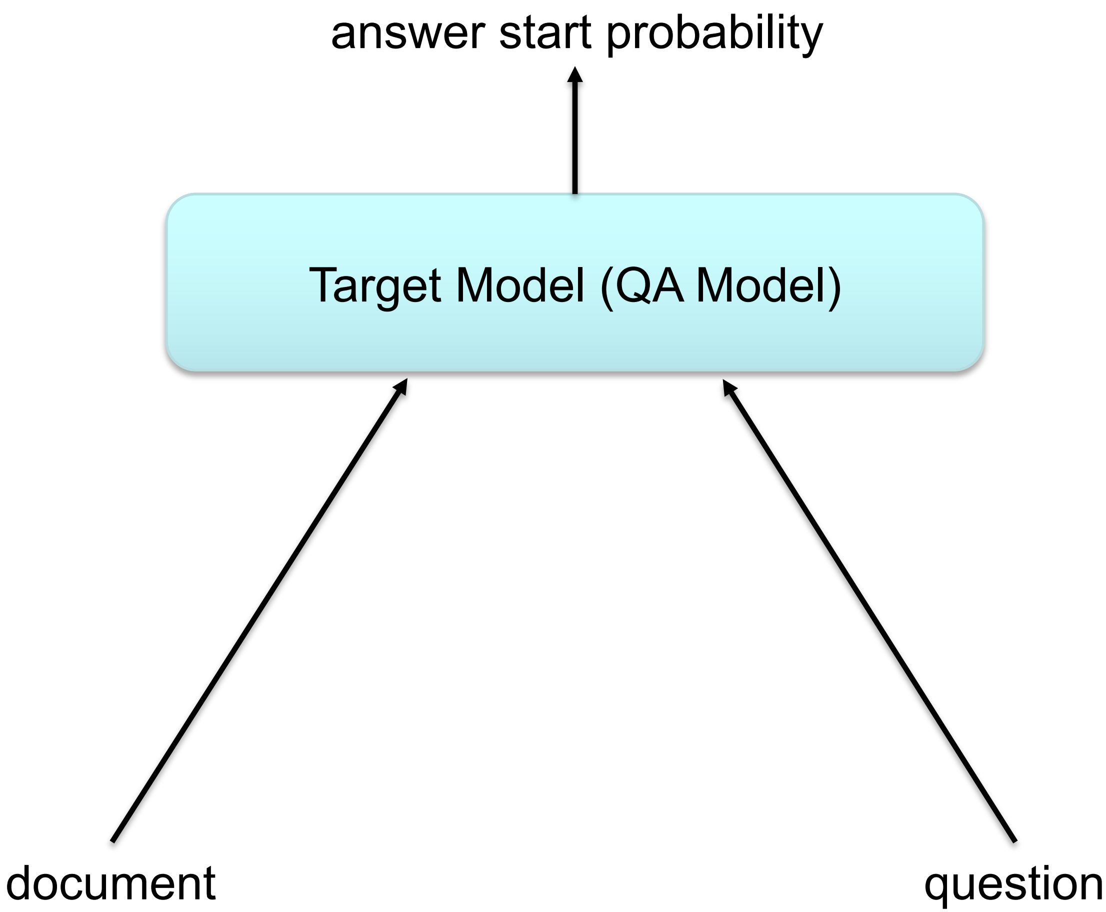
Is SQuAD big enough?

- Larger training data never hurts!
- Can we benefit from other larger corpus?

Transfer Learning

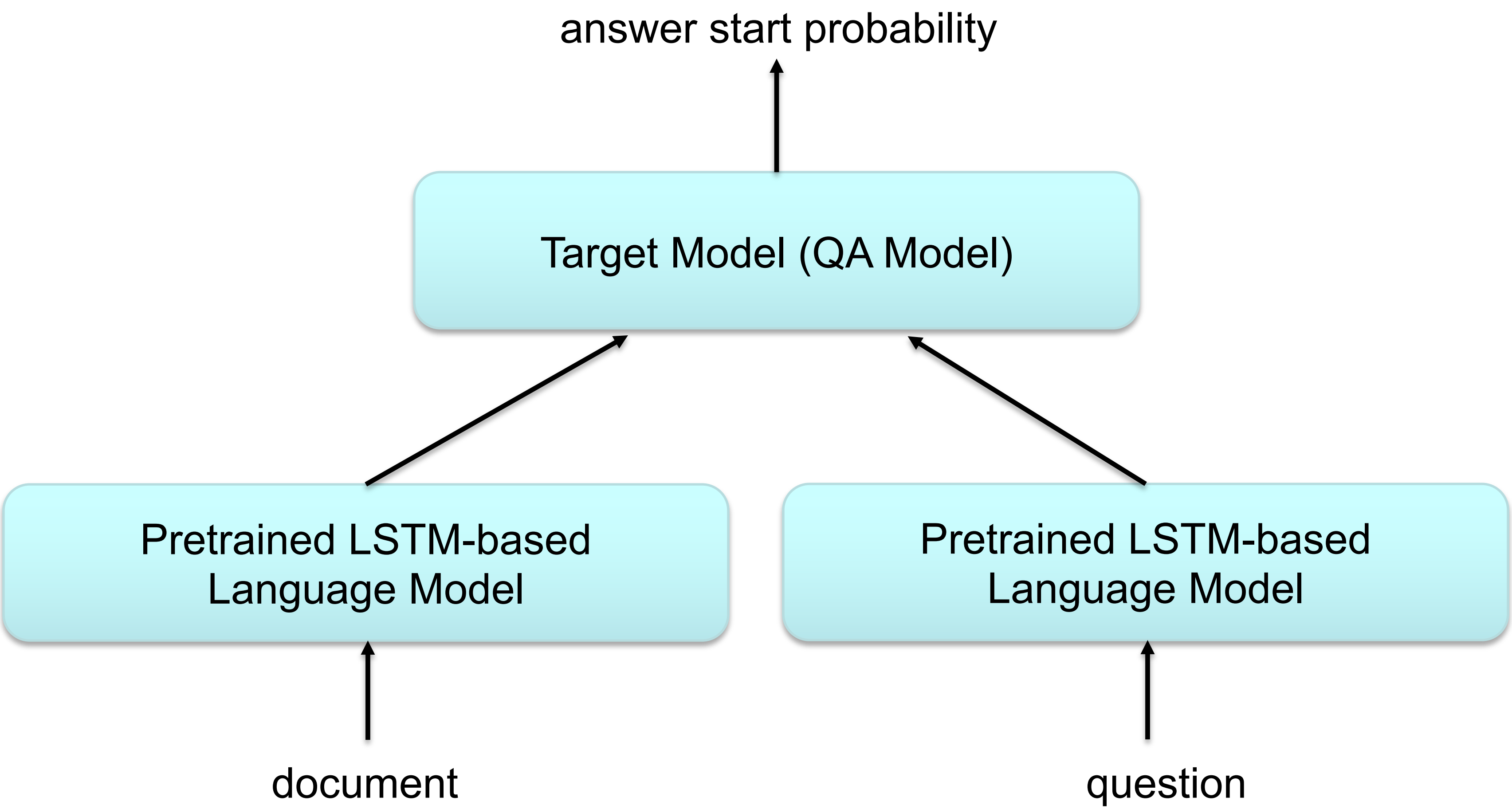
- Data Augmentation via back translation
  - QANet (Yu et al., ICLR 2018)
- Transfer learning from MT Model
  - CoVe (McCann et al., ICML 2018)
- Transfer learning from Language Model, *trained on a large unlabeled corpus*
  - ELMo (Peters et al., NAACL 2018)

Pretrained Language Model



Peters et al. Deep contextualized word representations. NAACL 2018.

Pretrained Language Model



Peters et al. Deep contextualized word representations. NAACL 2018.

SQuAD Leaderboard

Date	Model	F1	EM
Aug 2016	+Cross-Attention	75~78%	66~68%
Mar 2017	+Self-Attention	80~82%	71-73%
Nov 2017	+Transfer Learning	84~86%	77~79%

Other tricks

- Sparse feature
  - Putting 0/1 flag of whether each context word appears in the question
  - POS, NER
- Finetuning with RL
  - Optimize for F1 (soft) rather than EM (hard)

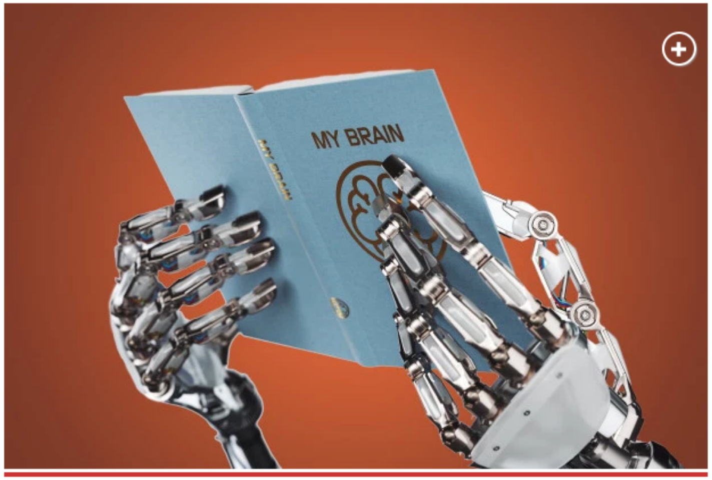
Super-human (Jan 2018)

TECH

AI systems are beating humans in reading comprehension

By Associated Press

January 24, 2018 | 2:25pm



iStockphoto

MORE ON:
ARTIFICIAL INTELLIGENCE

Swiss bank digitally 'clones' chief economist

PROVIDENCE, RI — Seven years ago, a computer beat two human quizmasters on a "Jeopardy" challenge. Ever since, the tech industry has been training its machines to make them even better at amassing knowledge and answering questions.

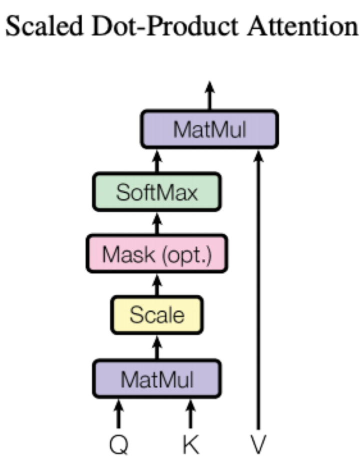
SQuAD Leaderboard

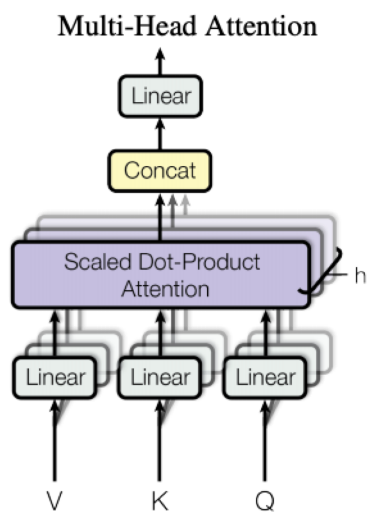
Date	Model	F1	EM
Aug 2016	+Cross-Attention	75~78%	66~68%
Mar 2017	+Self-Attention	80~82%	71-73%
Nov 2017	+Transfer Learning	84~86%	77~79%
Jan 2018	+Tricks	88~90%	82~85%

We thought it was the end…

Just Attention?

- Transformer
  - Multi-head self-attention
  - No LSTM (unlike ELMo)





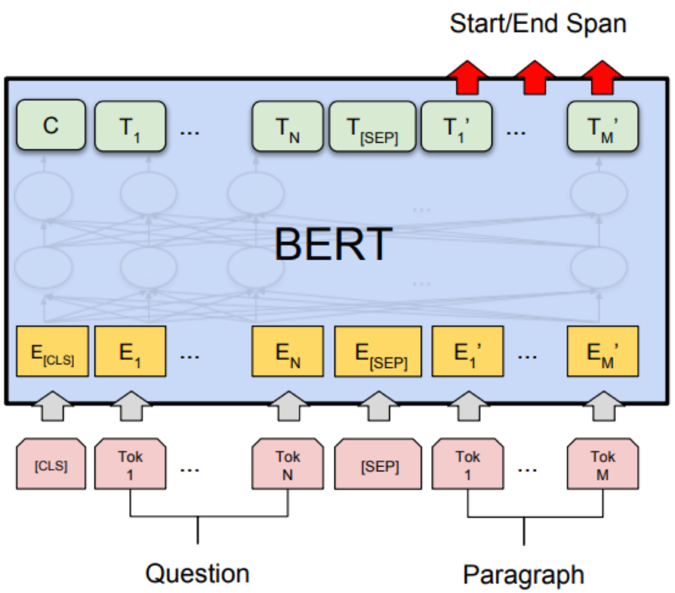
Vaswani et al. Attention is all you need. NIPS 2017.

Just Attention?

- Transformer
  - Multi-head self-attention
  - No LSTM (unlike ELMo)
- Concat context and question
- Trained masked language model on a large unlabeled corpus
  - cloze test instead of next word prediction
- Super-large model (64 TPUs)

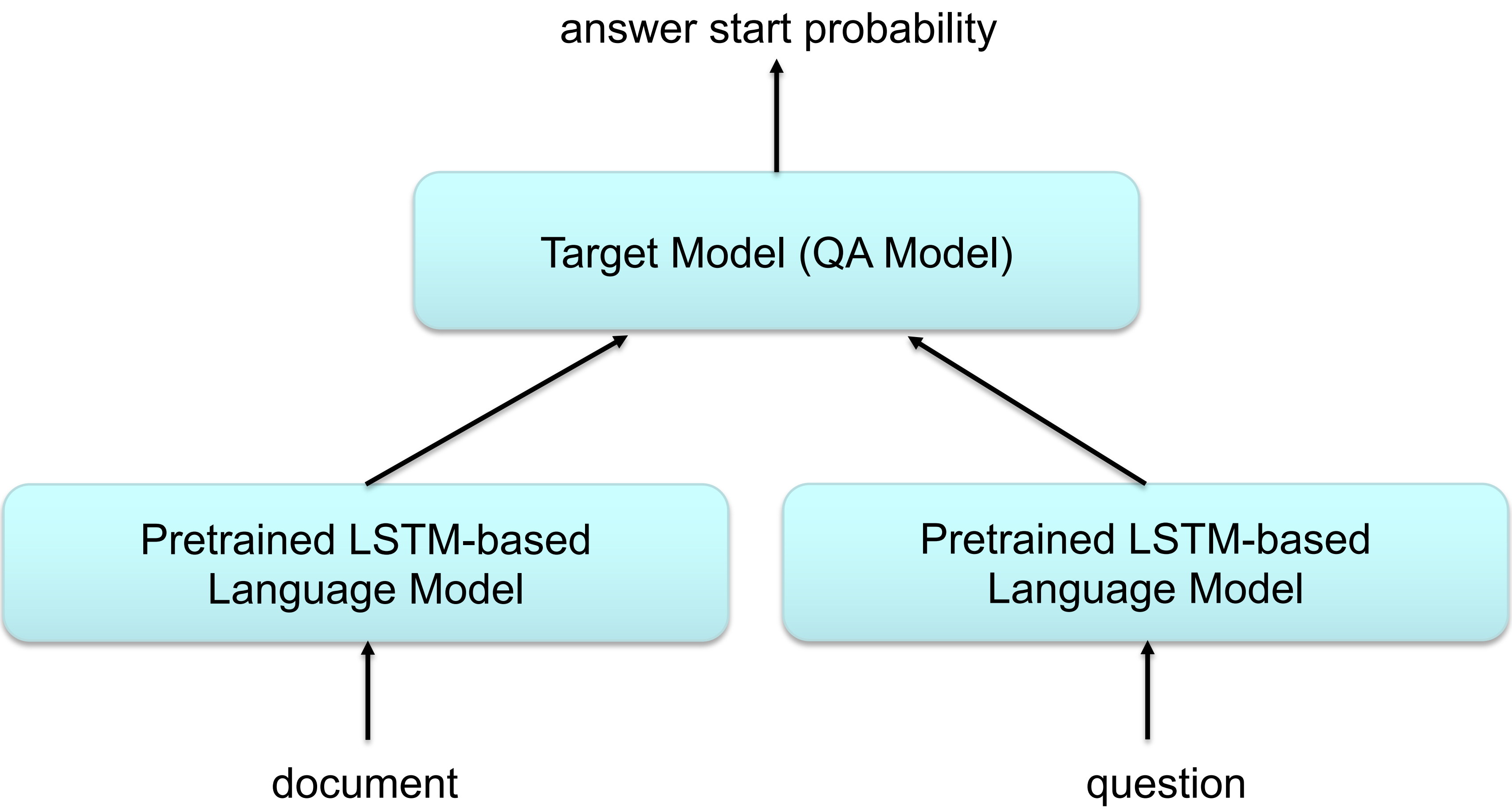
Just Attention: BERT

- Basically, doing all three at once:
  - **Cross-attention**: via concat
  - **Self-attention**: using Transformer
  - **Transfer learning**: via masked LM



Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. 2018.

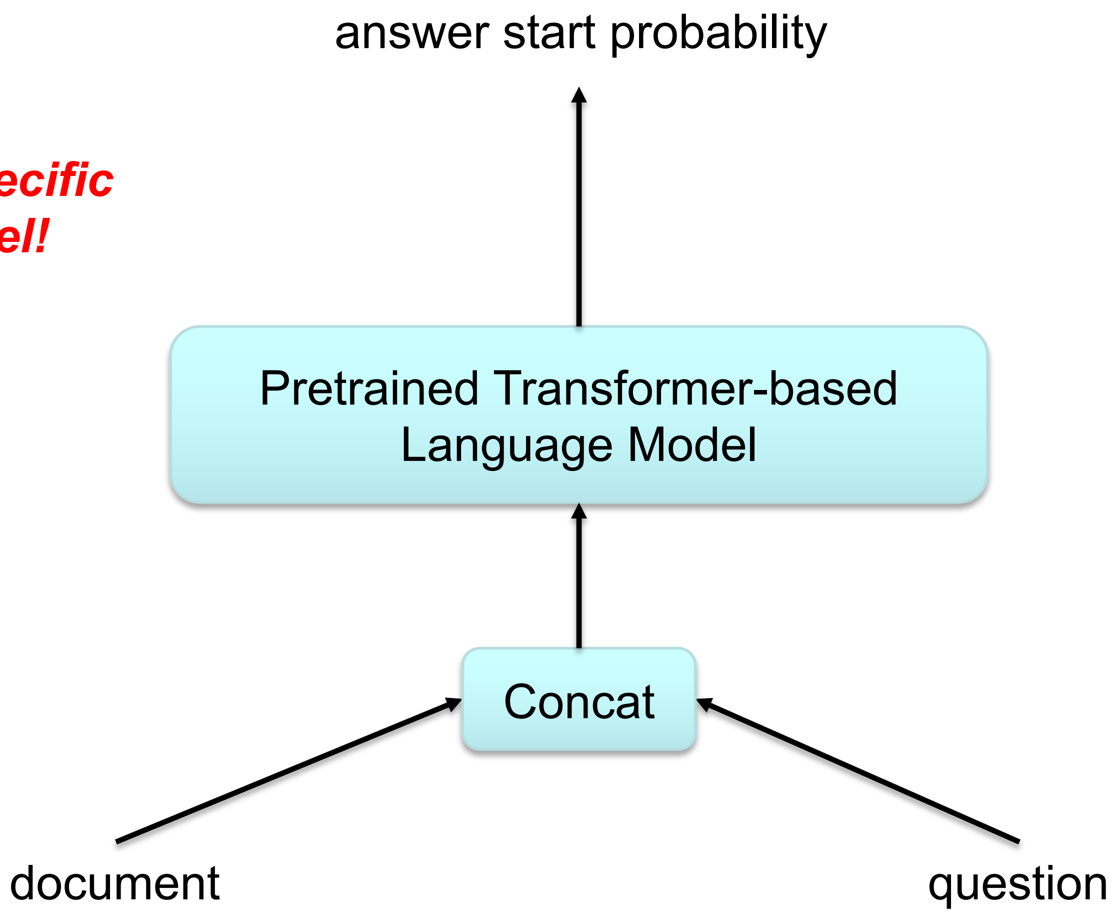
Pretrained Language Model



Peters et al. Deep contextualized word representations. NAACL 2018.

Pretrained LM with Transformer

No task-specific target model!



Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. 2018.

SQuAD Leaderboard

Date	Model	F1	EM
Aug 2016	+Cross-Attention	75~78%	66~68%
Mar 2017	+Self-Attention	80~82%	71-73%
Nov 2017	+Transfer Learning	84~86%	77~79%
Jan 2018	+Tricks	88~90%	82~84%
Sep 2018	Just Attention	93%	87%

SQuAD Leaderboard

Date	Model	F1	EM
Aug 2016	+Cross-Attention	75~78%	66~68%
Mar 2017	+Self-Attention	80~82%	71-73%
Nov 2017	+Transfer Learning	84~86%	77~79%
Jan 2018	+Tricks	88~90%	82~84%
Sep 2018	Just Attention	93%	87%
	Human	91%	82%

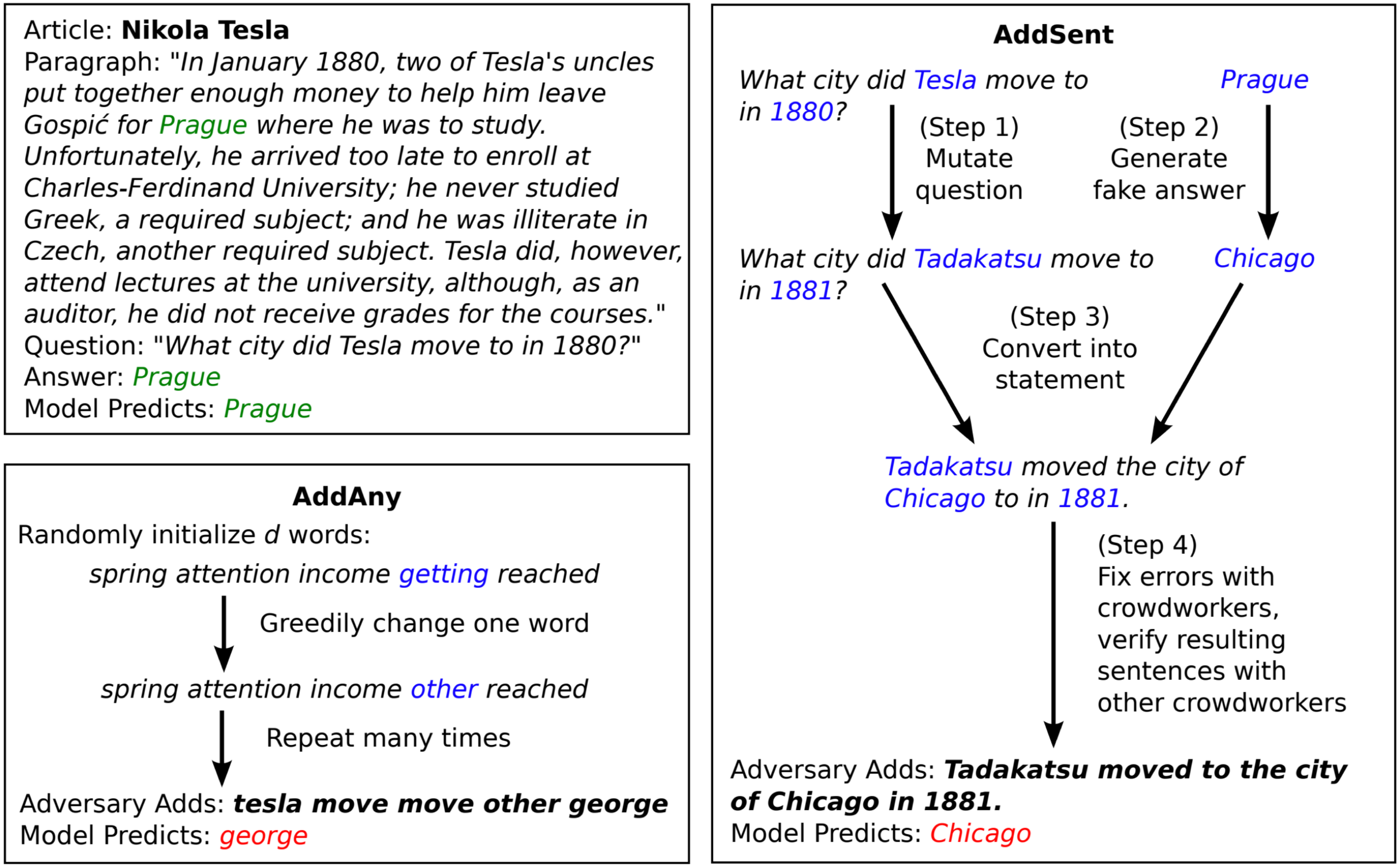
5% above human!

AI! Really? Adversarial Examples

(Jia and Liang, 2017)

- It is more about pattern matching!
- Can systems answer questions about paragraphs that contain adversarially inserted sentences automatically generated to distract without changing the correct answer or misleading humans?
- In this adversarial setting, the accuracy of sixteen published models drops from an average of 75% F1 score to 36%!
- When the adversary is allowed to add ungrammatical sequences of words, average accuracy on four models decreases further to 7%!

AI! Really? Adversarial Examples



(Jia and Liang 2017)

TriviaQA (Joshi et al., 2017)

- 95k Trivia questions
- Wikipedia articles are retrieved via IR.
- Distant supervision (string match)
- Simulates open-domain QA better
  - SQuAD is relatively artificial; the questions are created by looking at the documents

More recent datasets

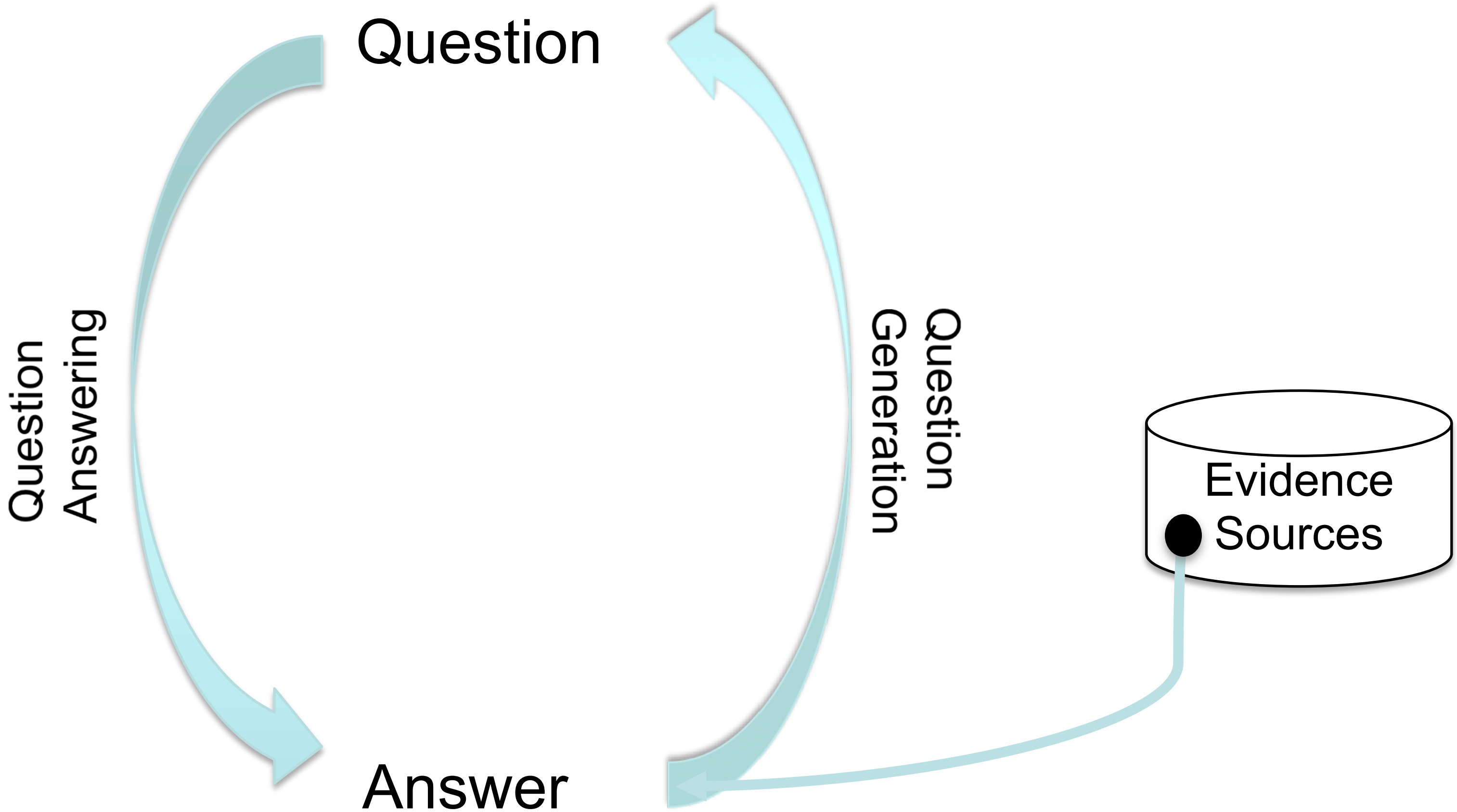
- RACE (Lai et al. 2017)
  - Collected from real English exams for middle and high school Chinese students
- NarrativeQA (Kočiský et al. 2017)
  - System must answer questions by reading the entire narrative (books or movie scripts).
- MultiRC (Khashabi et al. 2018)
  - Questions can only be answered by taking into account information from multiple sentences
- SQuAD 2.0 (Rajarpurkar et al. 2018)
  - Unanswerable questions written adversarially by crowdworkers that are similar to answerable ones. Systems must determine when no answer is supported by the passage and abstain from answering

More new datasets at EMNLP

- QuAC (Choi et al., 2018)
  - Multi-turn question answering in a conversation
- CoQA (Reddy et al. 2018)
  - Similarly to QuAC, multi-turn QA in a conversation
- HotpotQA (Yang et al. 2018)
  - Similarly to MultiRC, need to look at multiple sentences, but the answer is a span of the context

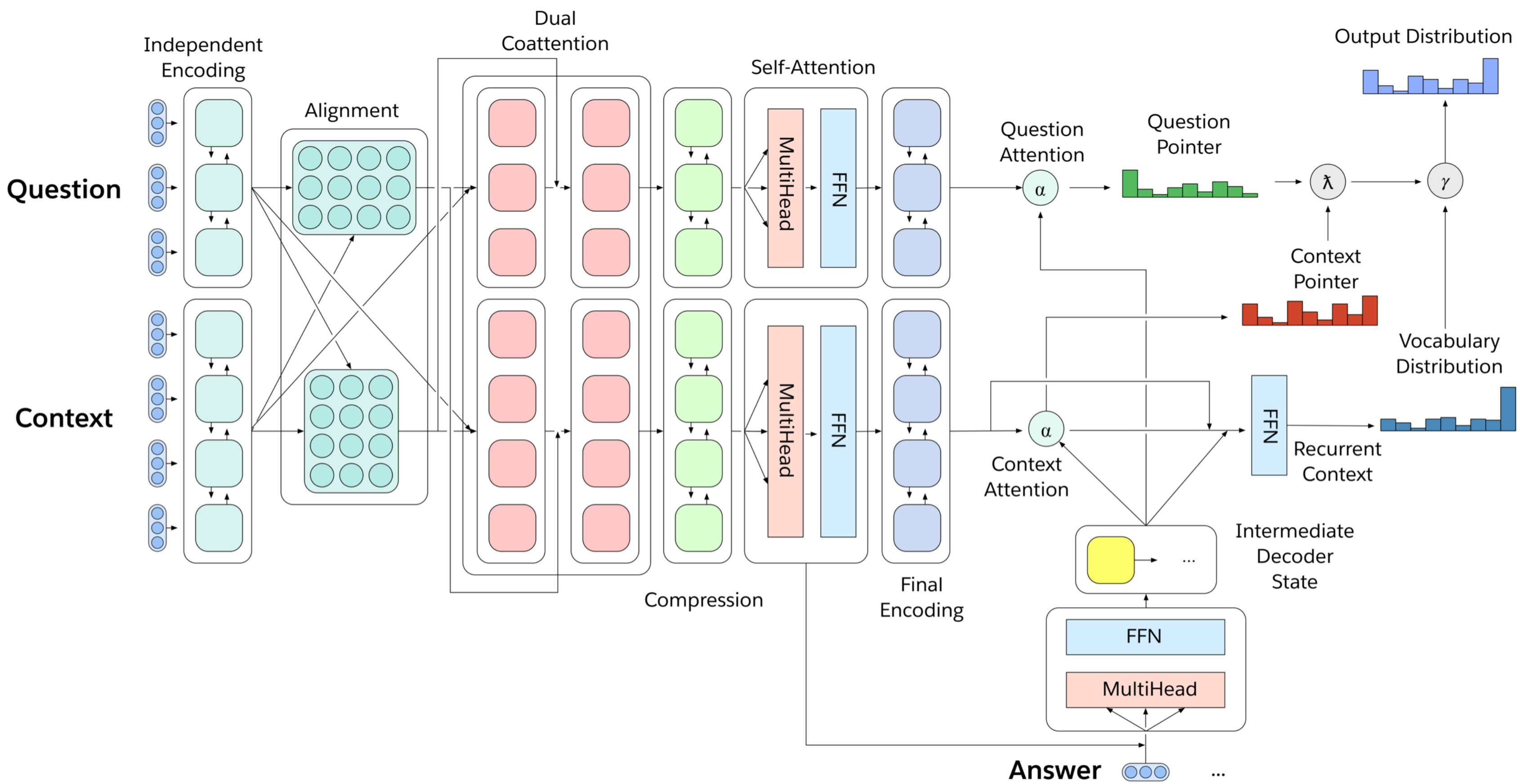
What's next? Semi-supervised Learning

Joint Question Answering and Question Generation



Tang et al. 2017 – Constraints based on the duality of QA and QG
Sachan et al. 2018 – Self Training

What's next? Multi-task learning



McCann et al. The Natural language decathlon. Multitask learning as question answering. 2018.

What's next? Transfer learning

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

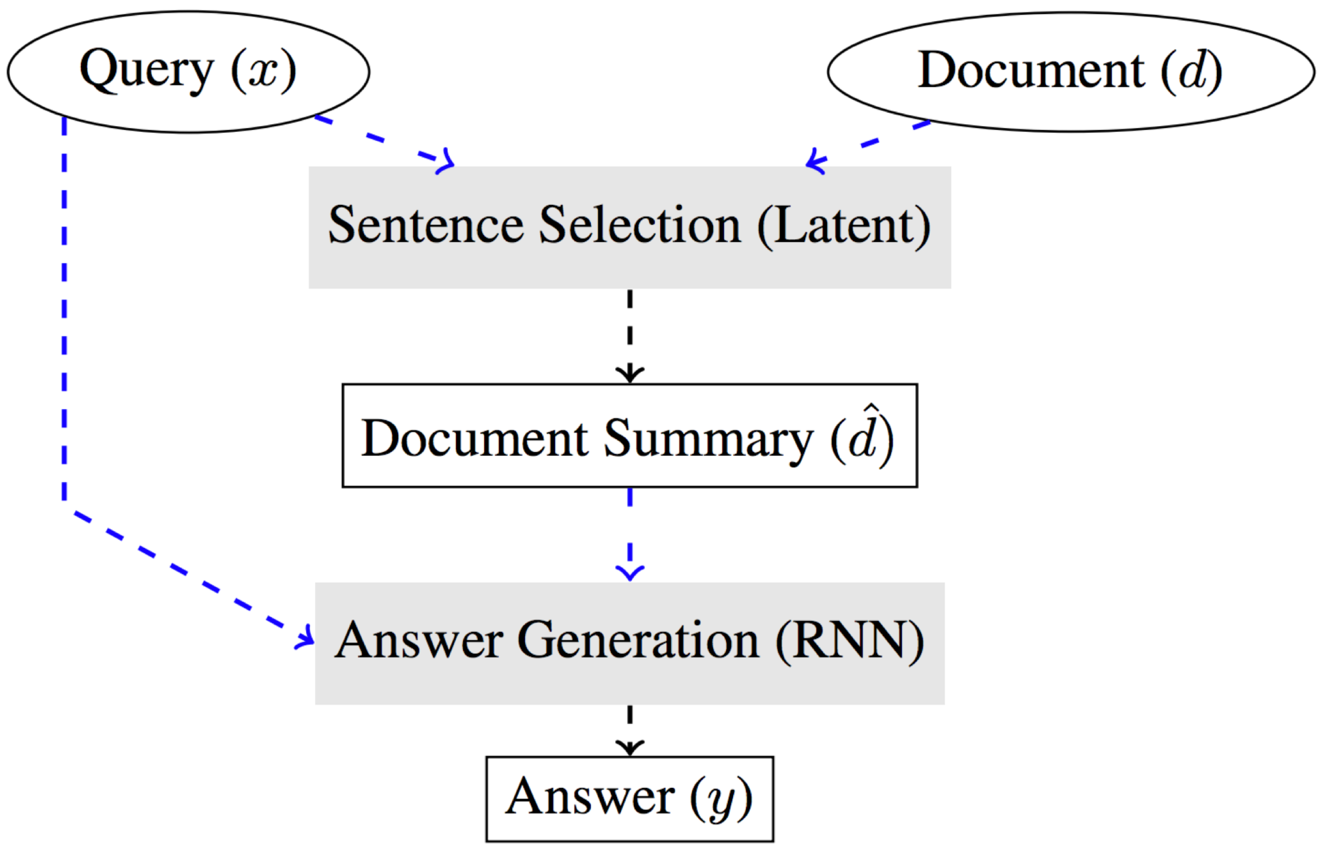
System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. 2018.

What's next? Large-scale QA

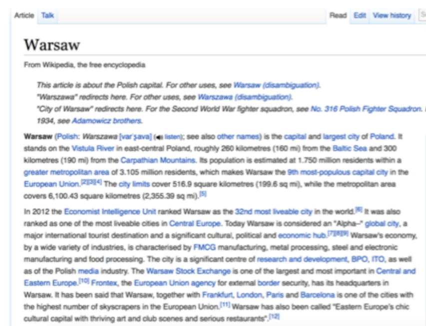


Choi et al. 2017

What's next? Open-domain QA

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever



Article Talk Read Edit View history

Warsaw

From Wikipedia, the free encyclopedia

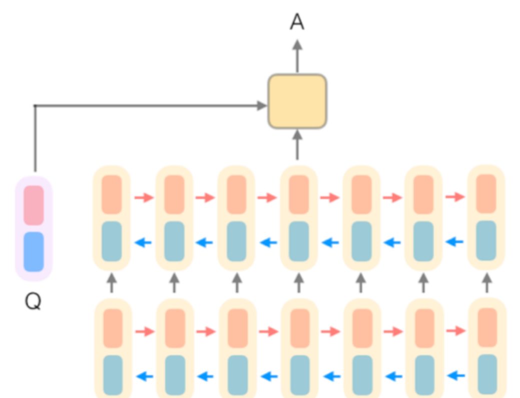
This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F 1934, see Adamowicz brothers.

Warsaw (Polish: *Warszawa* [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500



DrQA

SQuAD, TREC, WebQuestions, WikiMovies

Chen et al. 2017

RC vs QA?

- *Reading comprehension* is for **evaluating** machine's text understanding ability
- *Question answering* is a useful **application** for users

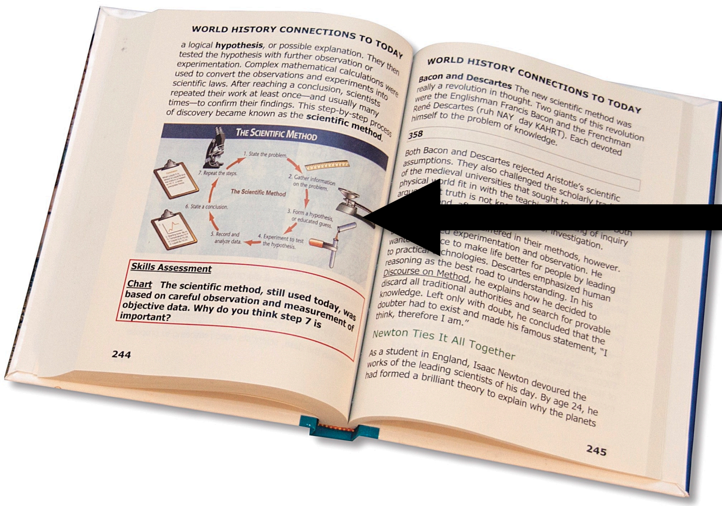
- They are correlated, but have different goals.
- Recent trends focus more on QA.

Outline

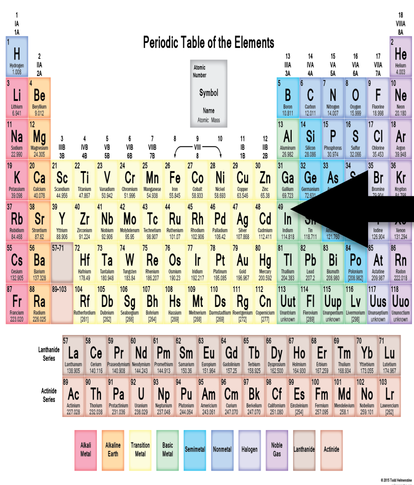
- *Machine Reading for Question Answering:*
  Reading Comprehension
  - *Feature Driven Models*
    - *MCTest*
  - *Deep Learning Models*
    - *WikiQA*
    - *CNN & DailyMail*
    - *SQUAD*
    - *Etc.*

  Beyond Reading Comprehensions
    - Elementary-level Science Exams
    - *Diagram QA*
    - *Textbook QA*
- *Mathematical Question Answering:*
  Advanced Math and Science Problems
  - *Algebra Word Problems*
  - *Geometry Problems*
  - *Newtonian Physics Problems*

Student Learning



Diagrams/Images



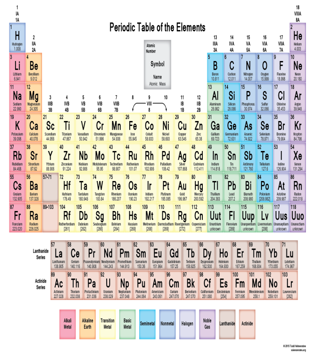
Knowledge

- Elementary Science Exams
  - State exams
- Diagram/Visual QA
  - Given a diagram or image, answer questions about it
- Textbook QA
  - Read textbook, answer questions

- **Elementary-level Science Exam**
  - **State exams**
- Diagram/Visual QA
  - Given a diagram or image, answer questions about it
- Textbook QA
  - Read textbook, answer questions

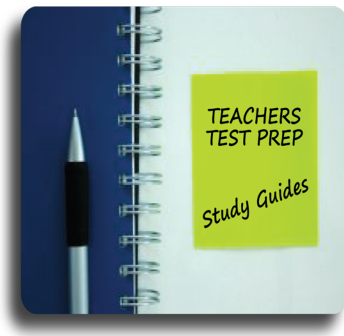
Project Aristo@AI2

Textbooks
(Unstructured)



Periodic Table

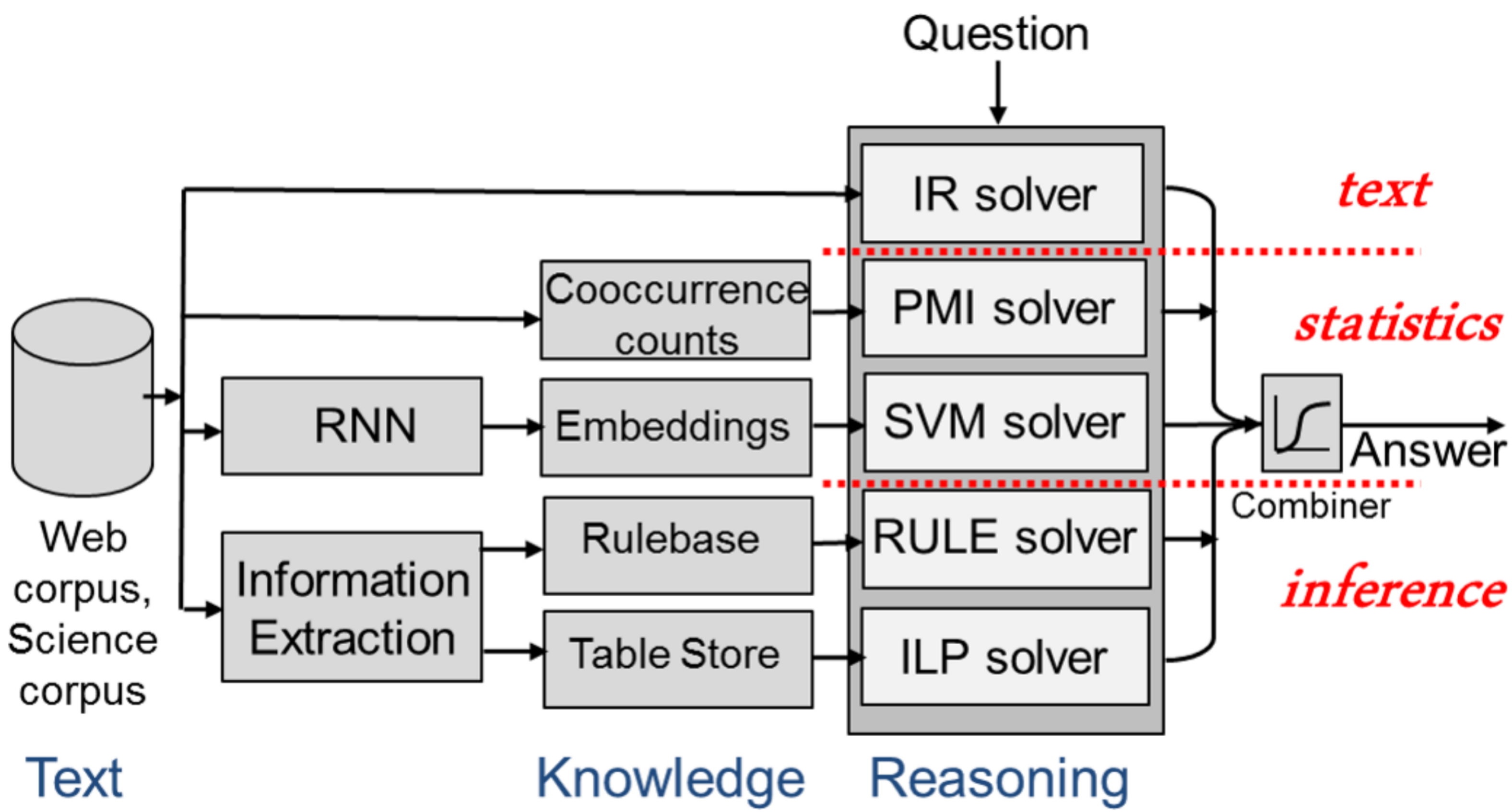
Dictionaries
(Semi-Structured)



Study Guides …

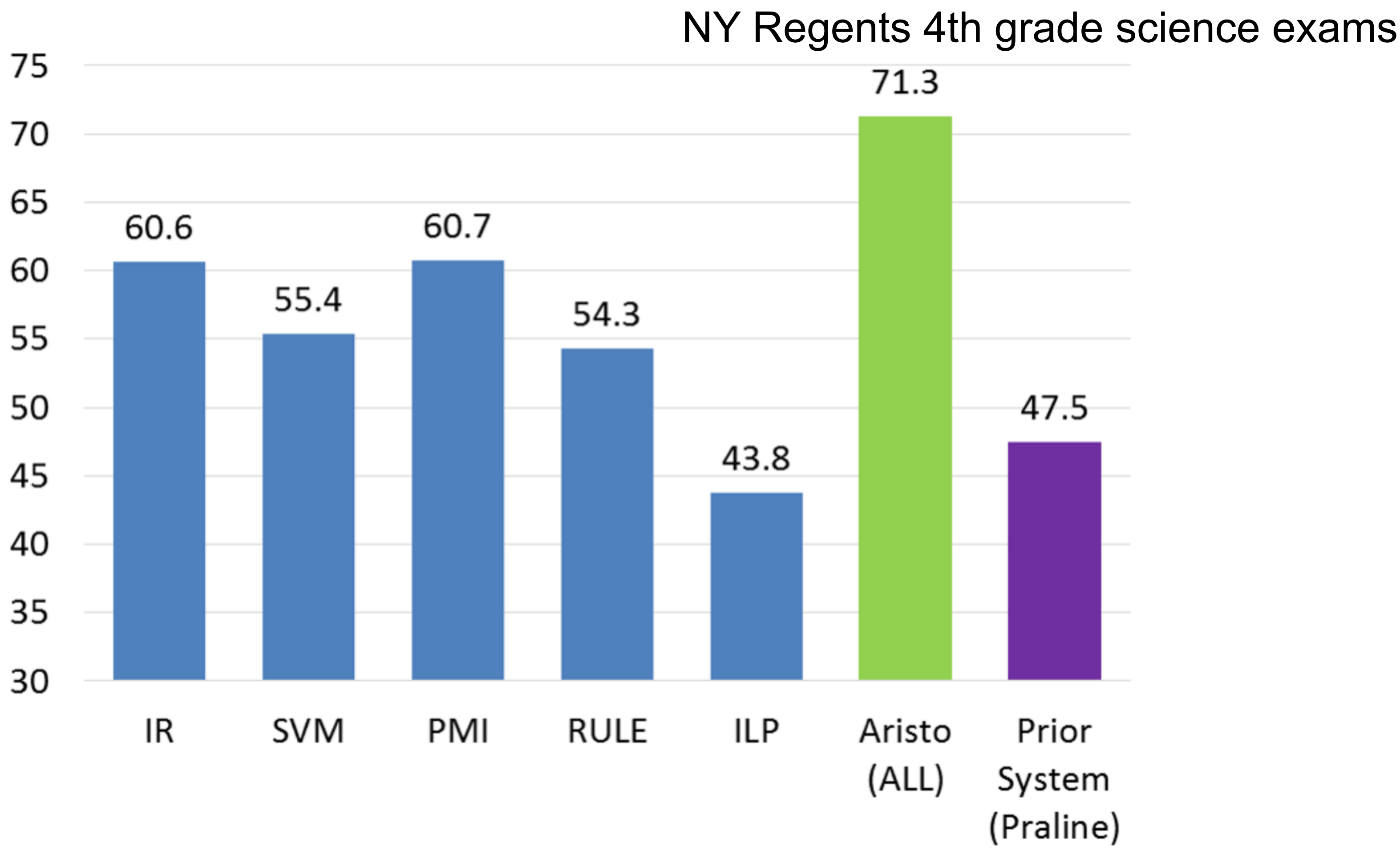
Q: Which of the following gases cause the greenhouse effect?
A) O_2 and CO_2
B) O_2, O_3 and CFC
C) O_2, CO, CO_2 and CFC
D) CO_2, CH_4, O_3 and CFC ✓

Project Aristo@AI2



Clark et al. AAAI 2016, Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions.

Project Aristo@AI2



Clark et al. AAAI 2016, Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions.

Structured Knowledge

Country	*Location*
France	north hemisphere
USA	north hemisphere
. . .	
Brazil	south hemisphere
Zambia	south hemisphere
. . .	

Hemisphere	*Orbital Event*	*Month*
northern	summer solstice	Jun
northern	winter solstice	Dec
northern	autumn equinox	Sep
. . .		
southern	summer solstice	Dec
southern	autumn equinox	Mar
. . .		

In USA, when is the summer solstice? (A) June

A path through joined rows in the knowledge tables match the question + answer

Domain-Targeted, High Precision Knowledge Extraction (Dalvi et al., TACL'17)
IKE - An Interactive Tool for Knowledge Extraction (Dalvi et al., AKBC'16)
Automatic Construction of Inference-Supporting Knowledge Bases (Clark et al., AKBC'14)

Allen AI Science Challenge

8th Grade Science questions

Team Name	Kernel	Team Members	Score
Cardal			0.59307
poweredByTalkwalker		+4	0.58344
Alejandro Mosquera			0.58256
Capuccino Monkeys			0.56241
A Pure Logical Approach			0.56154

Unfortunately, all the top models are fancy IR methods!

- **IR features** applied by searching over corpora compiled from **various sources** (study-guides, quiz-building websites, open source textbooks, Wikipedia).
- **Features based on properties of questions -** length of question and answer, form of answer like numeric answer, answers containing none of the above, and relationships among answer options.
- **Various weightings and stemming** strategies. All the top models used **gradient boosted trees**!

Aristo Demo

Aristo Quiz: https://aristo-quiz.allenai.org/

Aristo Demo: http://aristo-demo.allenai.org/

- Elementary-level Science Exam
  - State exams
- **Diagram/Visual QA**
  - **Given a diagram or image, answer questions about it**
- Textbook QA
  - Read textbook, answer questions

Visual Question Answering

Who is wearing glasses?

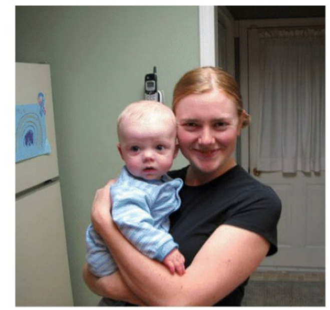
man

woman

Where is the child sitting?

fridge

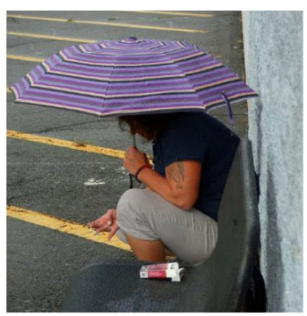
arms



Is the umbrella upside down?

yes

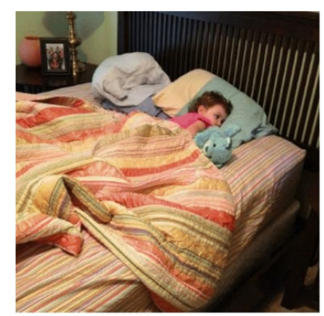
no



How many children are in the bed?

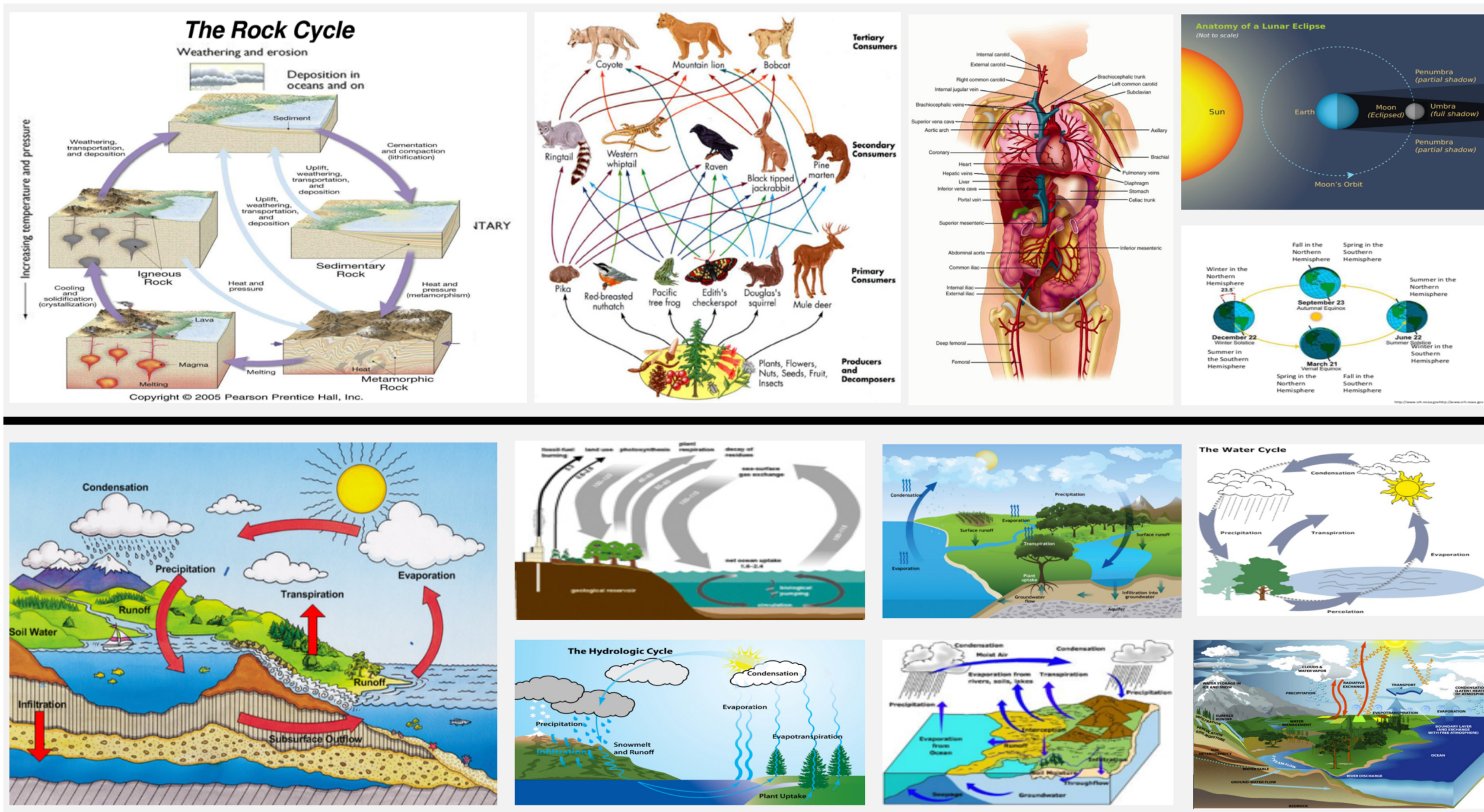
2

1



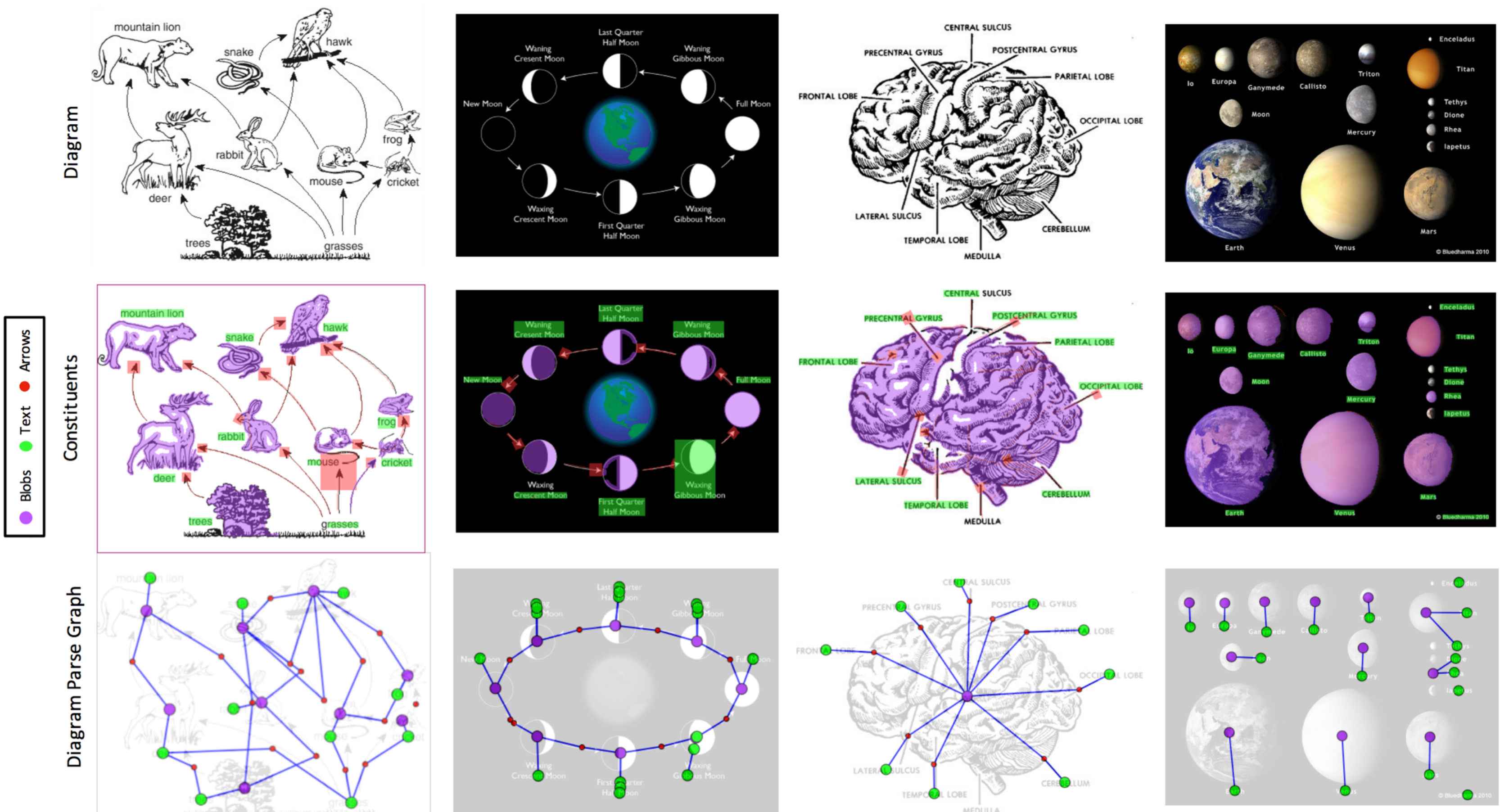
VisualQA Challenge http://visualqa.org/

Diagram Question Answering



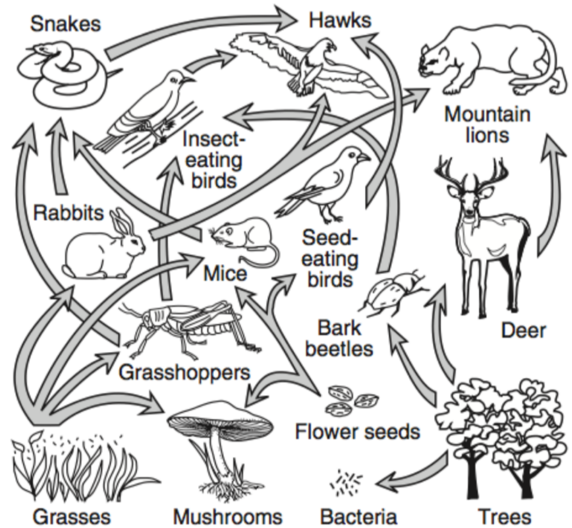
Khembhavi et al. ECCV 2016, A Diagram Is Worth A Dozen Images

Diagrams to Graph Representations



A Diagram Is Worth A Dozen Images. Khembhavi et al. 2016

Semantic Parsing to Probabilistic Programs



1. According to the given food chain, what is the number of organisms that eat deer? (A) 3 (B) 2 (C) 4 **(D) 1**

2. Which organism is both predator and prey? (A) Bark Beetles **(B) Insect-eating birds** (C) Deer (D) Hawks

3. Based on the given food web, what would happen if there were no insect-eating birds? **(A) The grasshopper population would increase.** (B) The grasshopper population would decrease. (C) There would be no change in grasshopper number.

Krishnamurthy et al. EMNLP 2016

Semantic Parsing to Probabilistic Programs

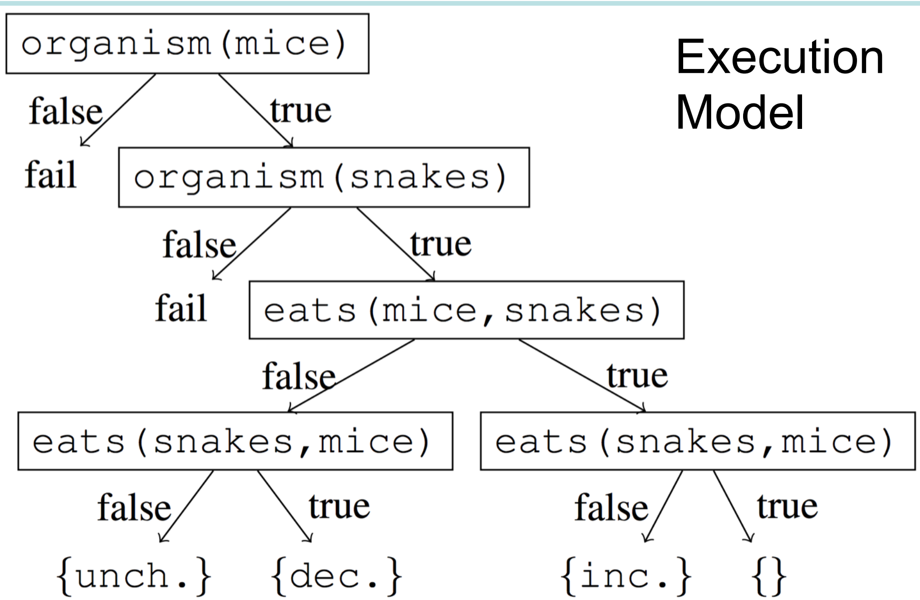
if	mice	die	snakes	will __ ?
$S/N/S : \lambda x.\lambda y.\lambda f.\,\textsc{cause}(x, f(y))$	$N : \textsc{mice}$	$S\backslash N : \lambda x.\textsc{decrease}(x)$	$N : \textsc{snakes}$	$skip$

$$S : \textsc{decrease}(\textsc{mice})$$

$$S/N : \lambda y.\lambda f.\textsc{cause}(\textsc{decrease}(\textsc{mice}), f(y))$$

$$S : \lambda f.\textsc{cause}(\textsc{decrease}(\textsc{mice}), f(\textsc{snakes}))$$

Semantic Parser

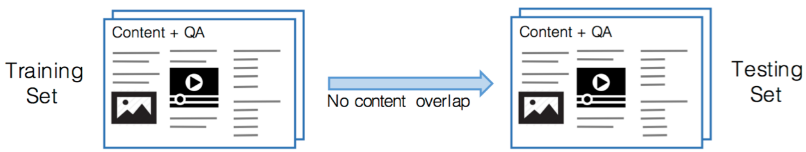


Krishnamurthy et al. EMNLP 2016

- Elementary-level Science Exam
  - State exams
- Diagram/Visual QA
  - Given a diagram or image, answer questions about it
- **Textbook QA**
  - **Read textbook, answer questions**

Textbook QA

Multi-modal Machine Comprehension (M^{3}C)



Textbook Question Answering (TQA)

1076 lessons from middle school curricula

Life Science	Earth Science	Physical Science

78,338 sentences
3,455 images
26,260 questions

Lessons in TQA

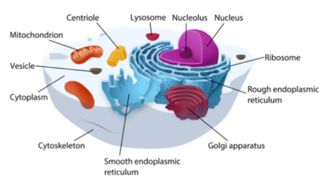
Cell Structures

Introduction

In some ways, a cell resembles a plastic bag full of Jell-O. Its basic structure is a cell membrane filled with cytoplasm. The cytoplasm of a eukaryotic cell is like Jell-O containing mixed fruit. It also contains a nucleus and other organelles.

Cell Membrane

The cell membrane is like the bag holding the Jell-O. It encloses the cytoplasm of the cell. It forms a barrier between the cytoplasm and the environment outside the cell. The function of the cell membrane is to protect and support the cell. It also controls what enters or leaves the cell. It allows only certain substances to pass through. It keeps other substances inside or outside the cell.



Cell Membrane Structure

Cytoplasm

Organelles

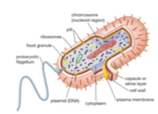
Lesson Summary

- The cell membrane consists of two layers of phospholipids.
- The cytoplasm consists of watery cytosol and cell structures.
- Eukaryotic cells contain a nucleus and other organelles

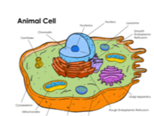
Vocabulary

Cell Wall	rigid layer that surrounds the cell membrane of a plant cell or fungal cell and that supports and protects the cell
Cyto-skeleton	structure in a cell consisting of filaments and tubules that crisscross the cytoplasm and help maintain the cells shape
Central Vacuole	large storage sac found in the cells of plants

Instructional Diagrams



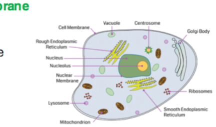
The image below shows the Prokaryotic cell. A prokaryote is a single-celled organism that lacks a membrane-bound nucleus (karyon), mitochondria, or any other membrane-bound organelle. In the prokaryotes, all the intracellular water-soluble components (proteins, DNA and metabolites) are located together in the cytoplasm enclosed by the cell membrane, rather than in separate cellular compartments.



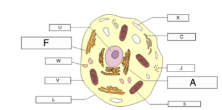
This diagram shows the anatomy of an Animal cell. Animal Cells have an outer boundary known as the plasma membrane. The nucleus and the organelles of the cell are bound by this membrane. The cell organelles have a vast range of functions to perform like hormone and enzyme production to providing energy for the cells. They are of various sizes and have irregular shapes. Most of the cells size range between 1 and 100 micrometers and are visible only with help of microscope.

Questions

What is the outer surrounding part of the Nucleus?
a. **Nuclear Membrane**
b. Golgi Body
c. Cell Membrane
d. Nucleolus



Which component forms a barrier between the cytoplasm and the environment outside the cell?
a. J
b. **L**
c. X
d. U



Which statement about the cell membrane is false?
a. It encloses the cytoplasm
b. It protects and supports the cell
c. **It keeps all external substances out of the cell**
d. none of the above

Kembhavi et al. CVPR 2017, Are You Smarter Than A Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension.

Textbook QA

Text Question Leaderboard - Final

Rank	Entrant	Accuracy
1	mlh	0.4208
2	beethoven	0.4200
3	tuanluu	0.4100
4	Daesik	0.4021
5	akshay107	0.3822
6	freerailway	0.3436

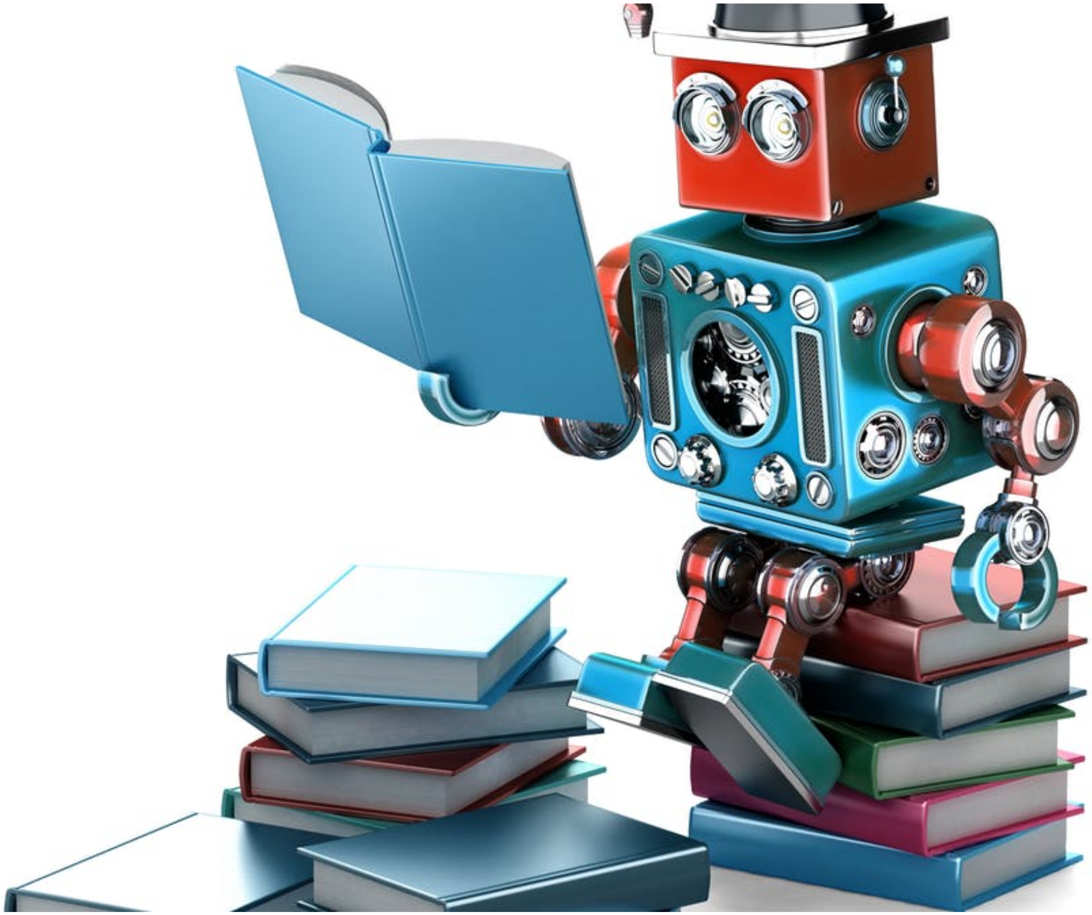
Diagram Question Leaderboard - Final

Rank	Entrant	Accuracy
1	beethoven	0.3175
2	mlh	0.3139
3	tuanluu	0.3050
4	Daesik	0.2588
5	akshay107	0.2581

CVPR 2017 Workshop on Visual Understanding Across Modalities

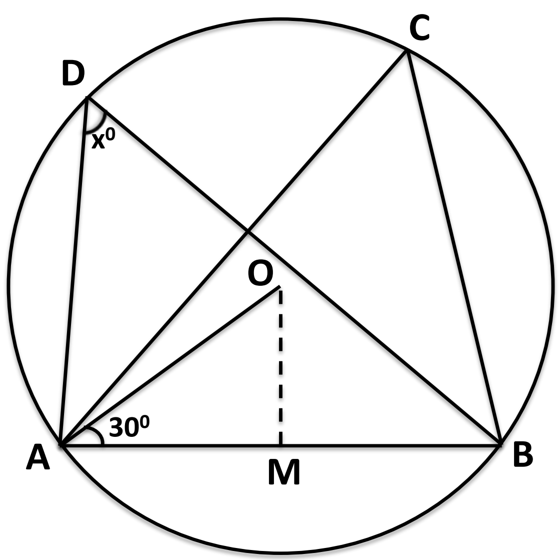
Challenging Task!

Story so far



What’s to come

Liz had 9 black kittens. She gave some of her kittens to John. John now has 11 kittens. Liz has 5 kittens left and 3 have spots. How many kittens did John get?



As shown in the Figure, ∠MAO = 30^0 and the radius of the circle with center O is 4cm. Find the value of x.

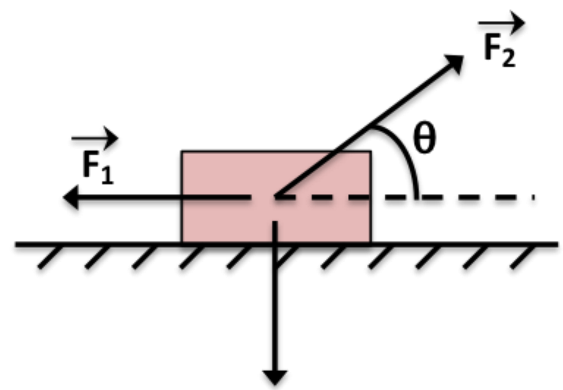


Figure above shows three forces applied to a trunk that moves leftward by 3.00 m over a frictionless floor. The force magnitudes are F_1 = 5.00N, F_2 = 9.00N, and F_3 = 3.00N, and the indicated angle is θ = 60.0° . During the displacement, what is the net work done on the trunk by the three forces?

Coffee Break!

Outline

- *Machine Reading:*
  Reading Comprehension
  - *Feature Driven Models*
    - *MCTest*
  - *Deep Models*
    - *WikiQA*
    - *CNN & DailyMail*
    - *SQUAD*
    - *Etc.*

  Multi-View/Multi-modal QA
    - Elementary-level Science Exams
    - *Diagram QA*
    - *Textbook QA*
- *Mathematical Question Answering:*
  Advanced Math and Science Problems
  - *Algebra Word Problems*
  - *Geometry Problems*
  - *Newtonian Physics Problems*

Outline

- *Machine Reading for Question Answering:*
  Reading Comprehension
  - *Feature Driven Models*
    - *MCTest*
  - *Deep Learning Models*
    - *WikiQA*
    - *CNN & DailyMail*
    - *SQUAD*
    - *Etc.*

  Beyond Reading Comprehensions
    - Elementary-level Science Exams
    - *Diagram QA*
    - *Textbook QA*

- *Mathematical Question Answering:*
  Advanced Math and Science Problems
  - *Algebra Word Problems*
  - *Geometry Problems*
  - *Newtonian Physics Problems*

Outline

- *Machine Reading for Question Answering:*

  Reading Comprehension
  - *Feature Driven Models*
    - *MCTest*
  - *Deep Learning Models*
    - *WikiQA*
    - *CNN & DailyMail*
    - *SQUAD*
    - *Etc.*

  Beyond Reading Comprehensions
    - Elementary-level Science Exams
    - *Diagram QA*
    - *Textbook QA*

- *Mathematical Question Answering:*

  Advanced Math and Science Problems
  - *Algebra Word Problems*
  - *Geometry Problems*
  - *Newtonian Physics Problems*

Arithmetic Word Problems

Liz had 9 black kittens. She gave some of her kittens to John. John now has 11 kittens. Liz has 5 kittens left and 3 have spots. How many kittens did John get?

- Solving math and science problems is a long-standing AI challenge, since 1963!
- Interesting problem for NLP

Quantitative Reasoning

Emanuel spent ***$13.6 million*** *from July until* ***the Feb. 24 election*** *and spent* ***an additional $6.3 million*** *in* ***the following five weeks****.*

How can you answer how much money Emanuel spent ?

For each number one needs to extract

- Unit of the number (which numbers indicate currency)
- Associated verb ("spent" implies expenditure)
- Associated arguments (Knowing that "Emanuel" is the subject for both numbers)

Reasoning About Quantities in Natural Language (Roy et al., TACL 2015).

Quantity Entailment

T: A bomb in a Hebrew University cafeteria killed **five Americans** and **four Israelis**.

H: A bombing at Hebrew University in Jerusalem killed **nine people**, including **five Americans**.

Given a statement **T** and a quantity **q** in **H**, do the **quantities in T** entail **q** ?
(Assuming upward monotonicity to be true)

Does the **quantities in T** entail **"nine people"** ?

Reasoning About Quantities in Natural Language (Roy et al., TACL 2015).

Challenges

- Variety of problems across different domains

Daily life	Shopping	Science Experiment
There are 7 crayons in the drawer. Mary took 3 crayons out of the drawer. How many crayons are there now?	Sally paid $12.32 total for peaches, after a 3 dollar coupon, and $11.54 for cherries. In total, how much money did Sally spend?	A petri dish originally contained 600 bacteria. A scientist let the bacteria grow and now there are 8917 of them. How many more bacteria are there now?

- No prior constraint on syntax or vocabulary
  - Requires world knowledge

Challenges

- Variety of problems across different domains

Daily life	Shopping	Science Experiment
There are 7 crayons in the drawer. Mary took 3 crayons out of the drawer. How many crayons are there now?	Sally paid $12.32 total for peaches, after a 3 dollar coupon, and $11.54 for cherries. In total, how much money did Sally spend?	A petri dish originally contained 600 bacteria. A scientist let the bacteria grow and now there are 8917 of them. How many more bacteria are there now?

- No prior constraint on syntax or vocabulary
  - Requires world knowledge

There were 7,000,000 people living in a country. Last year, 90,000 children were born, and 16,000 people immigrated to it. How many new people began living in the country last year?

Challenges

- Verbs vs. other words

Books, toy aircrafts, cookies

Liz had 9 black kittens.
She gave some of her kittens to John.
John now has 11 kittens.
Liz has 5 kittens left and 3 have spots.
How many kittens did John get?

received

Challenges

- Irrelevant Information

> Liz had 9 black kittens. She gave some of her kittens to John. John now has 11 kittens. Liz has 5 kittens left and 3 have spots. How many kittens did John get?

- Missing information

> There were 6 roses in the vase. Mary cut some more roses from her flower garden. There are now 16 roses in the vase. How many roses did she cut?

Challenges

- Ambiguity (requires context):

Sara's high school won 5 basketball games this year. They lost 3 games. How many games did they play in all? $5+3=x$

John has 8 orange balloons, but lost 2 of them. How many orange balloons does John have now? $8-2=x$

Is now an active area of research in AI:

[Lei et al, 2018, Roy et al 2018, Wang et al. 2017, Shyam et al, 2016, Hosseini et al 2014, Kushman et al 2014, Roy and Roth 2015, Zhou et al., 2015, etc]

A Historical Perspective

- STUDENT program (Bobrow, 1964)
  - Restricted set of English language
  - A set of rules form a set of equations representing the problem
- WORDPRO (Fletcher, 1985)
  - Introduced the concept of “schemas”, Rule based
- Domain specific solvers
  - CHIPS (Briars, 1984), ARITHPRO (Dellarosa, 1986) and ROBUST (Bakman, 2007) – word problems
  - CARPS i.e. Calculus Rate Problem Solver (Charniak, 1968)
  - HAPPINESS (Gelb, 1971) - simple probability questions

Datasets

- **AddSub** Data from Learning to Solve Arithmetic Word Problems with Verb Categorization (Hosseini et al., 2014).
- **SingleOp** Data from Reasoning About Quantities in Natural Language (Roy et al., 2015).
- **MultiArith** Data from Solving General Arithmetic Word Problems (Roy and Roth, 2015)
- **SingleEQ** Data from Parsing Algebraic Word Problems into Equations (Koncel-Kedziorski et al., 2015)
- **Algebra**.com Data from Learning to Automatically Solve Algebra Word Problems (Kushman et al., 2014).

3221 Questions are paired with equations

Koncel-Kedziorski et al. NAACL 2016, MAWPS: A Math Word Problem Repository http://lang.ee.washington.edu/MAWPS/

More Datasets

- Math23K

23K math word problems from a couple of online education web sites for elementary school students

Linear algebra questions with only one variable

More Datasets

- **AQUA-RAT (Algebra Question Answering with Rationales)**

100,000 crowdsourced algebraic word problems with natural language rationales (Ling et al. 2017)

Problem 1:
Question: Two trains running in opposite directions cross a man standing on the platform in 27 seconds and 17 seconds respectively and they cross each other in 23 seconds. The ratio of their speeds is:
Options: A) 3/7 B) 3/2 C) 3/88 D) 3/8 E) 2/2
Rationale: Let the speeds of the two trains be x m/sec and y m/sec respectively. Then, length of the first train = 27x meters, and length of the second train = 17 y meters. (27x + 17y) / (x + y) = 23 $\rightarrow$ 27x + 17y = 23x + 23y $\rightarrow$ 4x = 6y $\rightarrow$ x/y = 3/2.
Correct Option: B

Problem 2:
Question: From a pack of 52 cards, two cards are drawn together at random. What is the probability of both the cards being kings?
Options: A) 2/1223 B) 1/122 C) 1/221 D) 3/1253 E) 2/153
Rationale: Let s be the sample space.
Then n(s) = 52C2 = 1326
E = event of getting 2 kings out of 4
n(E) = 4C2 = 6
P(E) = 6/1326 = 1/221
Answer is C
Correct Option: C

Techniques

- Template-based
  - Kushman et al., 2014
- Verb-Categorization
  - Hosseini et al., 2014
- Parsing word problems to equation trees
  - Koncel-Kedziorski et al., 2015
  - Roy et al., 2015
- Deep Learning

Learning to Automatically Solve Algebra Word Problems

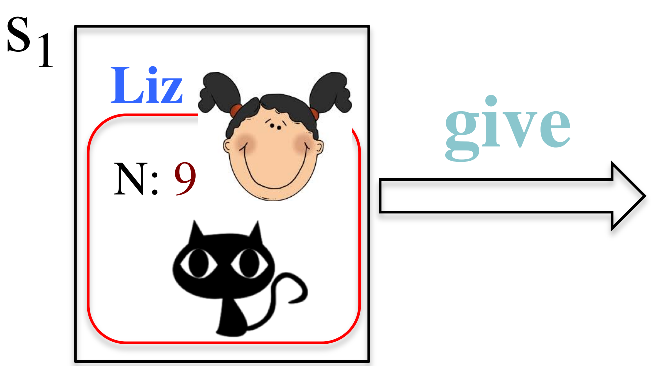
Derivation 1	
Word problem	An amusement park sells 2 kinds of tickets. Tickets for children cost $ 1.50. Adult tickets cost $ 4. On a certain day, 278 people entered the park. On that same day the admission fees collected totaled $ 792. How many children were admitted on that day? How many adults were admitted?
Aligned template	$u_1^1 + u_2^1 - n_1 = 0$ $n_2 \times u_1^2 + n_3 \times u_2^2 - n_4 = 0$
Instantiated equations	$x + y - 278 = 0$ $1.5x + 4y - 792 = 0$
Answer	$x = 128$ $y = 150$

Derivation 2	
Word problem	A motorist drove 2 hours at one speed and then for 3 hours at another speed. He covered a distance of 252 kilometers. If he had traveled 4 hours at the first speed and 1 hour at the second speed, he would have covered 244 kilometers. Find two speeds?
Aligned template	$n_1 \times u_1^1 + n_2 \times u_2^1 - n_3 = 0$ $n_4 \times u_1^2 + n_5 \times u_2^2 - n_6 = 0$
Instantiated equations	$2x + 3y - 252 = 0$ $4x + 1y - 244 = 0$
Answer	$x = 48$ $y = 52$

Kushman et al., 2014

Learn to Solve Word Problems with Verb Categorization

- Representation:
  - State transitions

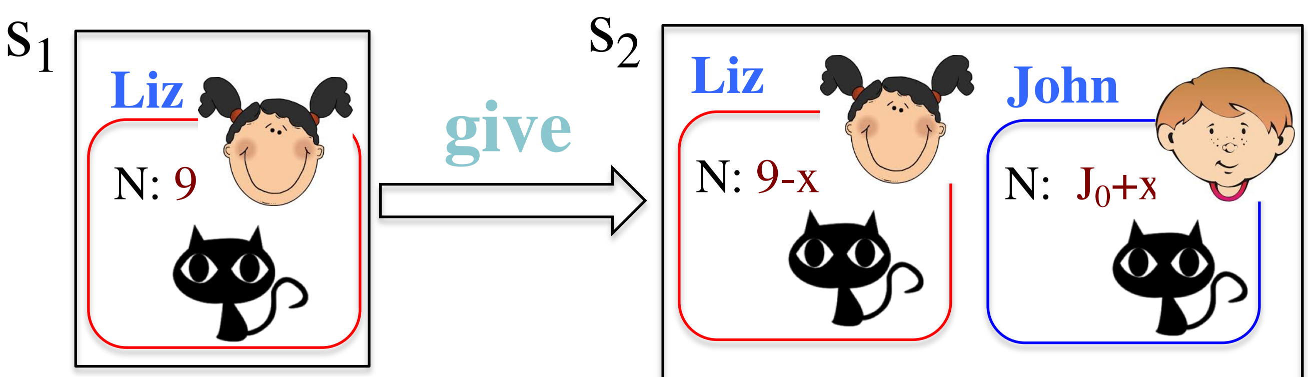


Liz gave some of her kittens to John.

(Hosseini et al. EMNLP 2014)

Learn to Solve Word Problems with Verb Categorization

- Representation:
  - State transitions
- Learning:
  - Learn state transitions based on verb categories
- Inference:
  - Form equations based on state transitions

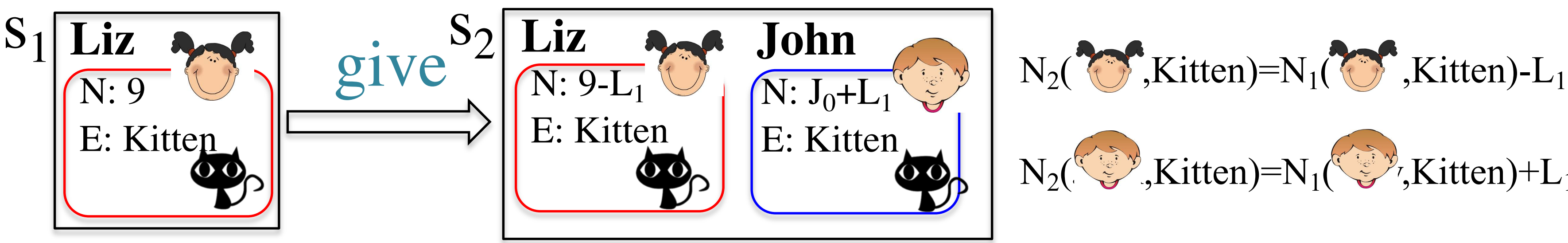


Liz gave some of her kittens to John.

(Hosseini et al. EMNLP 2014)

Representation: State Transitions

Transitions based on verb categories and containers



Give: transfer entities from one container$_1$ to container$_2$

- Verb Categories: Transfer or initialization of quantities in containers {Construction, destruction, positive, negative, positive transfer, negative transfer, initialization}

(Hosseini et al. EMNLP 2014)

Algorithm: ARIS

Liz had 9 black kittens.
She gave some of her kittens to John.

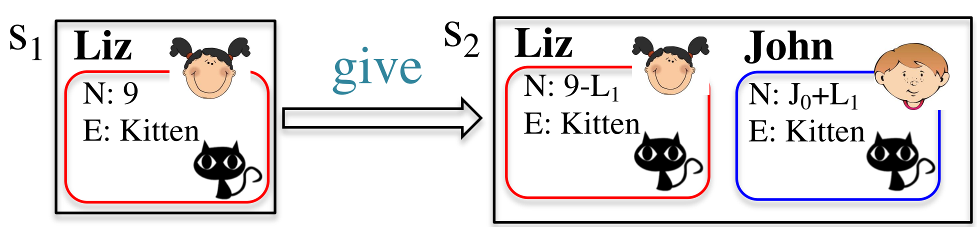
Grounding into Entities and Containers

(give, 9)

Learning: Training for verb categories

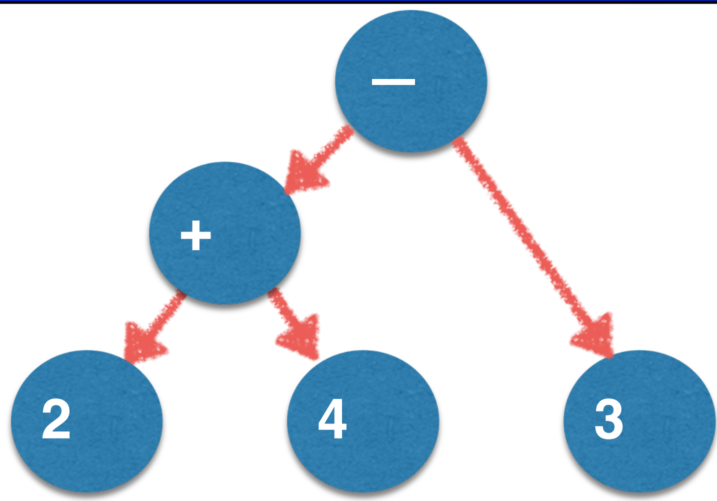
(Liz,Give):negative
(John,Give):positive

Inference: Forming state transitions and equations



(Hosseini et al. EMNLP 2014)

Expression Trees



Expression Tree for **2 + 4 - 3**

LCA(2, 3) = **subtraction**

LCA(4, 3) = **subtraction**

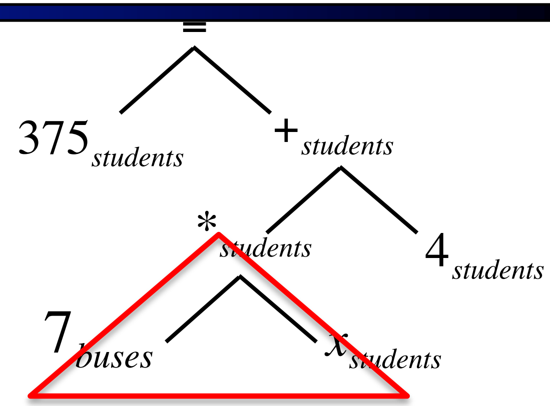
LCA(2, 4) = **addition**

Decompose word problems into simpler decision problems, where each decision problem is to predict **lowest common ancestor operation** for a pair of numbers in the problem.

Solving General Arithmetic Word Problems (Roy et al., EMNLP 2015).

Typed Equation Trees

On Monday, 375 students went on a trip to the zoo. All 7 buses were filled and 4 students had to travel in cars. How many students were in each bus?



- Semantically augmented equation trees:
  - Leaves: typed entities
  - Intermediate nodes: math operations

Koncel-Kedziorski et al., 2015

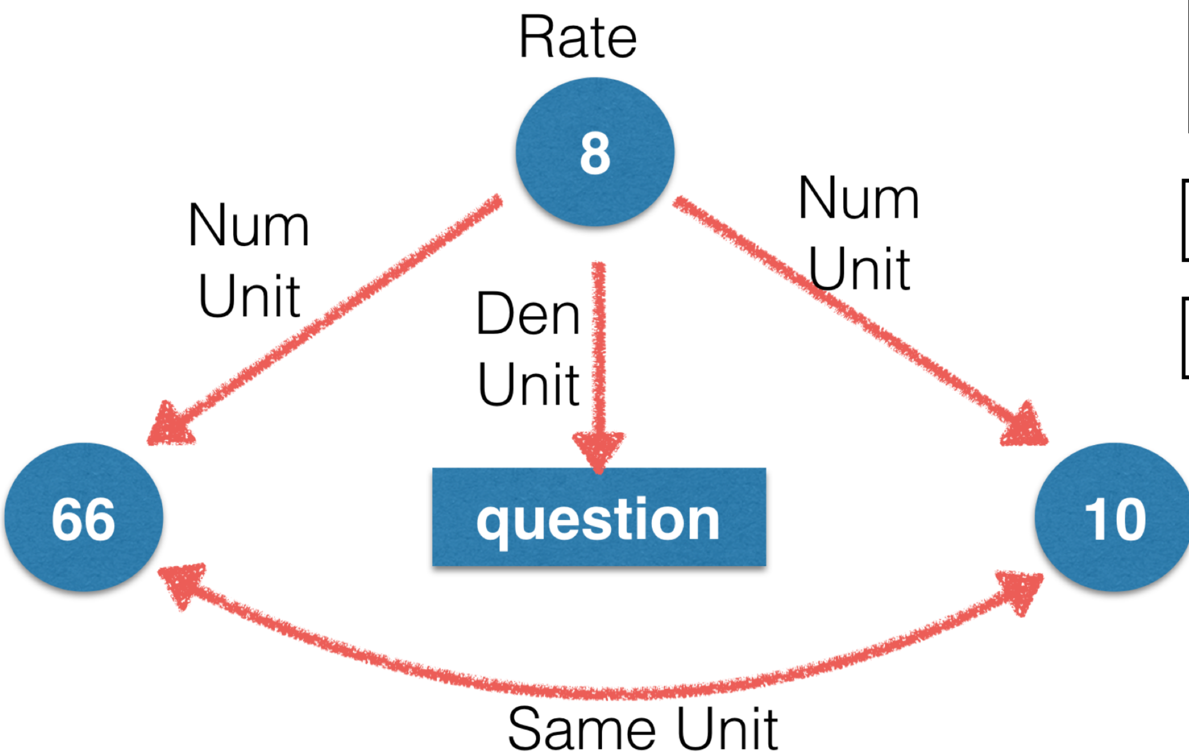
Units

- **Units** associated with quantities provide information essential to support quantitative reasoning.
- **Unit Dependency Graph** as a way to capture and reason about units mentioned in a problem.
- Reduces the error of math solvers by over 10%

Roy and Roth 2017

Unit Dependency Graph

Isabel picked 66 flowers for her friends wedding. She was making bouquets with 8 flowers in each one. If 10 of the flowers wilted before the wedding, how many bouquets could she still make?



Num Unit of 8: **flower**
Den Unit of 8: **bouquet**

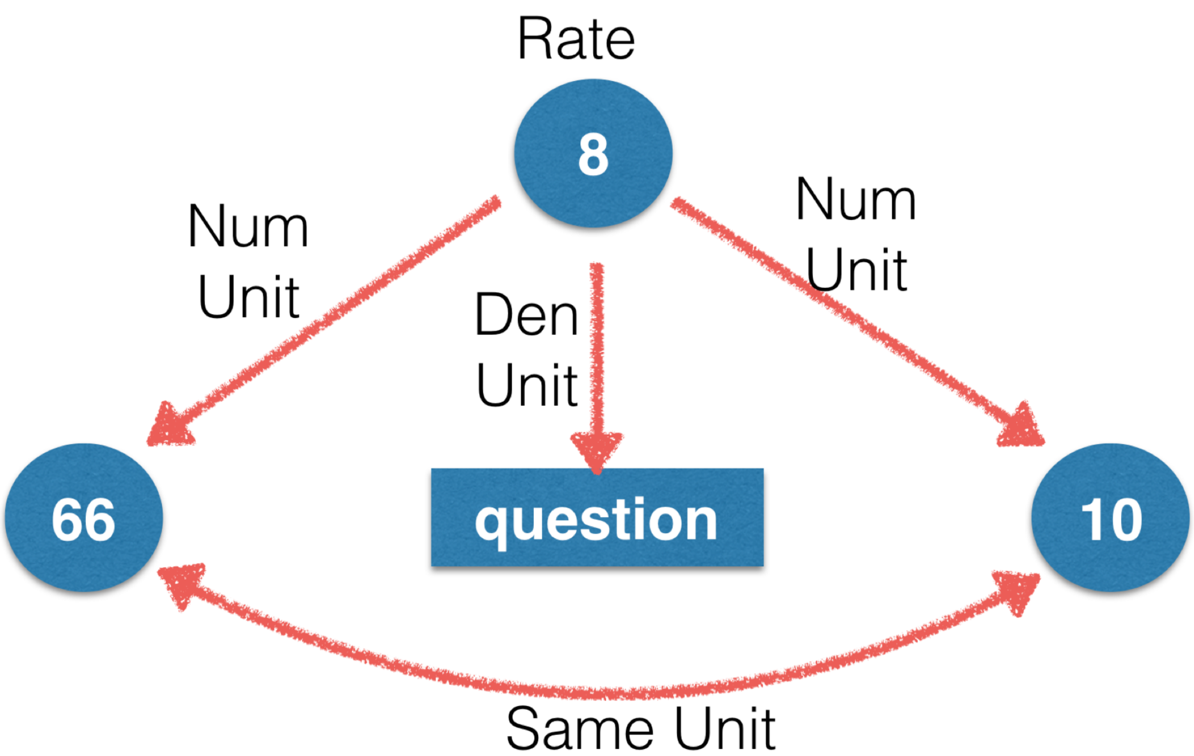
Unit of 66 and 10: **flower**

Unit of question: **bouquet**

Roy and Roth 2017

Unit Dependency Graph

Isabel picked 66 flowers for her friends wedding. She was making bouquets with 8 flowers in each one. If 10 of the flowers wilted before the wedding, how many bouquets could she still make?

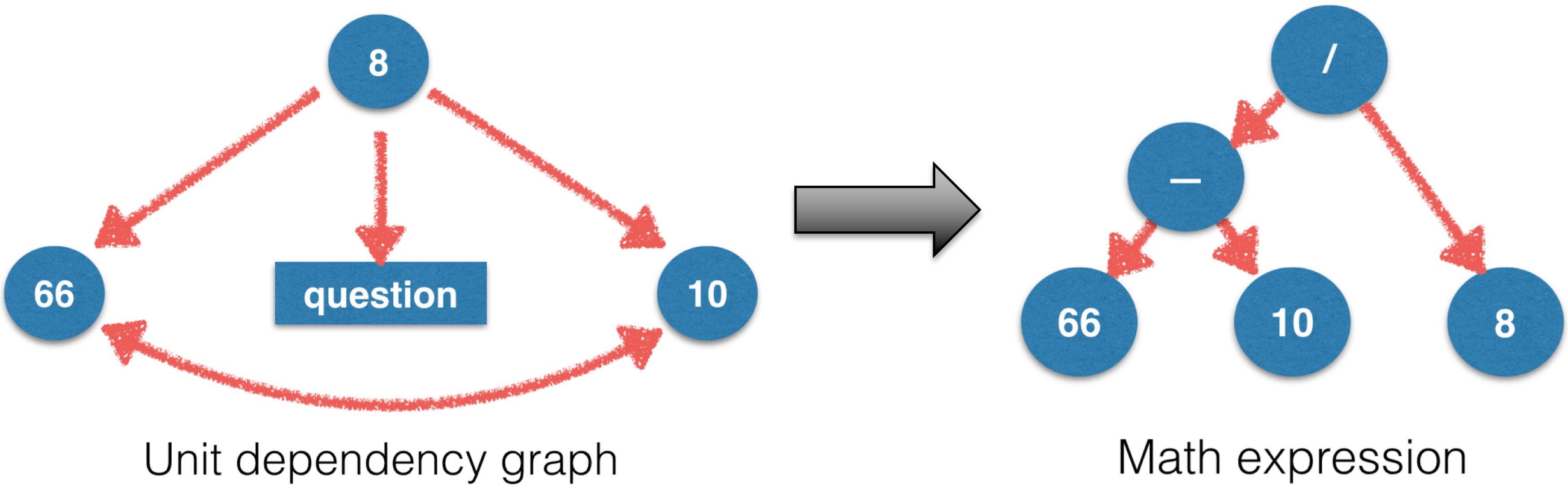


Same Unit => addition or subtraction

Den Unit to question => Something will be divided by it to get answer's unit

Roy and Roth 2017

Unit Dependency Graph



Roy and Roth 2017

Mapping to Declarative Rules

4 Declarative Knowledge Classes

Transfer	Stephen has 5 apples. Daniel gave him 4 apples. How many apples does Stephen have ?
Unit Dependencies	Stephen has 5 bags. Each bag has 4 apples. How many apples does Stephen have ?
Explicit Math	Stephen has 5 apples. Daniel has 4 more apples than Stephen. How many apples does Daniel have ?
Subset Relations	There are 5 red apples. There are 6 green apples. How many apples are there in all ?

For each class, we have a few declarative rules

Roy and Roth 2018

Deep Neural Solver

- seq2seq model for transforming problem text to a math equation

Method	Math23K	Alg514
ZDC (Zhou et al., 2015 - Improved version of Kushman et al., 2014)	42.1%	79.7%
Seq2seq Model	58.1%	16.1%

Wang et al. 2017

Program Induction by Rationale Generation

Problem 1:
Question: Two trains running in opposite directions cross a man standing on the platform in 27 seconds and 17 seconds respectively and they cross each other in 23 seconds. The ratio of their speeds is:
Options: A) 3/7 B) 3/2 C) 3/88 D) 3/8 E) 2/2
Rationale: Let the speeds of the two trains be x m/sec and y m/sec respectively. Then, length of the first train = 27x meters, and length of the second train = 17 y meters. (27x + 17y) / (x + y) = 23 → 27x + 17y = 23x + 23y → 4x = 6y → x/y = 3/2.
Correct Option: B

Problem 2:
Question: From a pack of 52 cards, two cards are drawn together at random. What is the probability of both the cards being kings?
Options: A) 2/1223 B) 1/122 C) 1/221 D) 3/1253 E) 2/153
Rationale: Let s be the sample space.
Then n(s) = 52C2 = 1326
E = event of getting 2 kings out of 4
n(E) = 4C2 = 6
P(E) = 6/1326 = 1/221
Answer is C
Correct Option: C

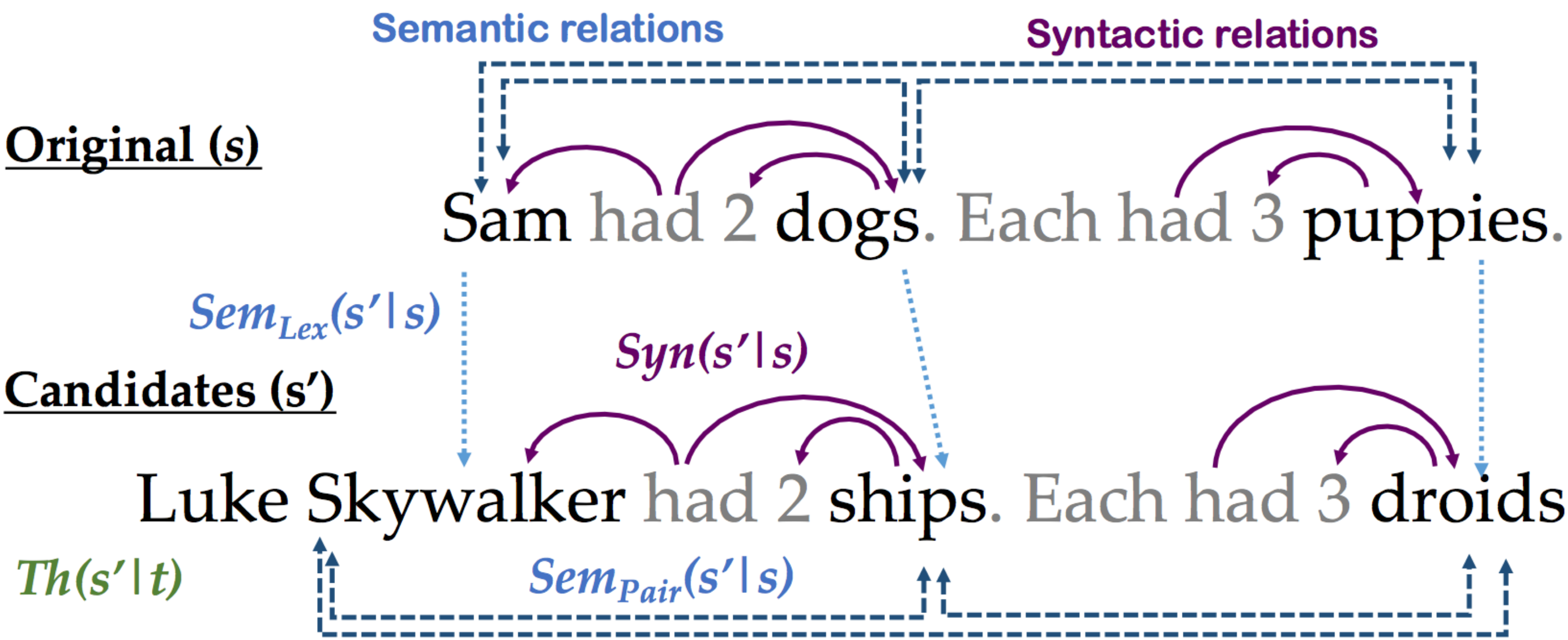
Ling et al. 2017

Generation

Jim walked 0.2 of a mile from school to David's house and 0.7 of a mile from David's house to his own house. How many miles did Jim walk in all?
Star Wars Uncle Owen walked 0.2 of a mile from hangar to Luke Skywalker's room and 0.7 of a mile from Luke Skywalker's room to his own room. How many miles did Uncle Owen walk in all?
Cartoon Finn squished 0.2 of a mile from cupboard to Melissa's dock and 0.7 of a mile from Melissa's dock to his own dock. How many miles did Finn squish in all?
Western Duane strolled 0.2 of a mile from barn to Madeline's camp and 0.7 of a mile from Madeline's camp to his own camp. How many miles did Duane stroll in all?

Koncel-Kedziorski et al. 2016, A Theme-Rewriting Approach for Generating Algebra Word Problems

Generation



Koncel-Kedziorski et al. 2016, A Theme-Rewriting Approach for Generating Algebra Word Problems

Illinois, AI2 Demo

AI2 Demo: http://euclid.allenai.org/

Illinois Demo: https://cogcomp.org/page/demo_view/Math

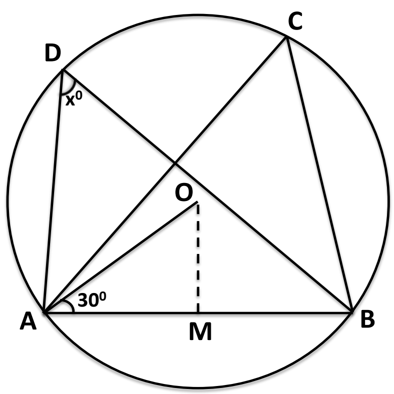
Outline

- *Machine Reading for Question Answering:*
  Reading Comprehension
  - *Feature Driven Models*
    - *MCTest*
  - *Deep Learning Models*
    - *WikiQA*
    - *CNN & DailyMail*
    - *SQUAD*
    - *Etc.*

  Beyond Reading Comprehensions
    - Elementary-level Science Exams
    - *Diagram QA*
    - *Textbook QA*
- *Mathematical Question Answering:*
  Advanced Math and Science Problems
  - *Algebra Word Problems*
  - *Geometry Problems*
  - *Newtonian Physics Problems*

Situated Question Answering

- Situated QA requires the system to answer questions about a very large, yet, constrained environment.



As shown in the Figure, $\angle MAO = 30^0$ and the radius of the circle with center O is 4cm. Find the value of x.

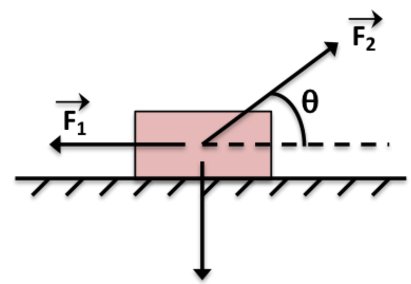


Figure above shows three forces applied to a trunk that moves leftward by 3.00 m over a frictionless floor. The force magnitudes are F_1 = 5.00N, F_2 = 9.00N, and F_3 = 3.00N, and the indicated angle is $\theta = 60.0^\circ$. During the displacement, what is the net work done on the trunk by the three forces?

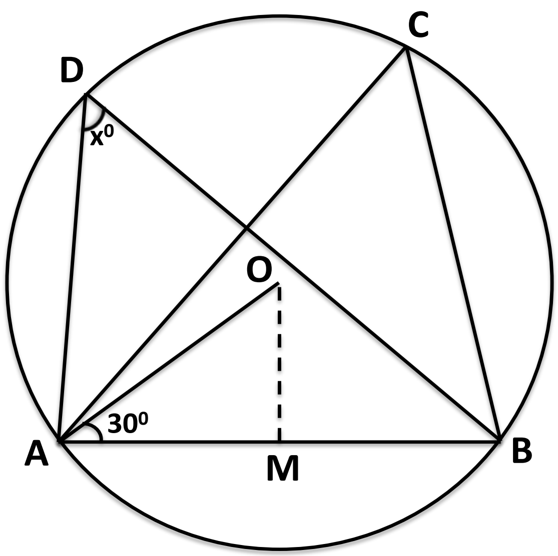
- Situated QA poses two key challenges:
  - How to **interpret the question**
  - How to build **background knowledge** about the environment (i.e. subject knowledge) and then how to the use background knowledge to determine the answer.

Geometry QA: A Historical Perspective

- Theorem Proving for geometry
  - (Feigenbaum and Feldman 1963)
  - Wus method (Wen-Tsun 1986)
  - Grobner basis method (Kapur1986)
  - Angle method (Chou et al. 1994)
- Geometric analogies (Evans, 1964)
- Tutoring systems:
  - Geometry Expert (Gao and Lin, 2002)
  - Geometry Explorer (Wilson and Fleuriot 2005)
- Synthesizing geometry problems:
  - Synthesize constructions given logical constraints (Gulwani et al. 2011, Itzhaky et al. 2013)
  - Generate geometric proof problems (Alvin et al. 2014a)

Physics QA: A Historical Perspective

- MECHO (Bundy et al. 1979)
  - Mechanics problems (pulley problems, statics problems, motion on smooth complex paths and motion under constant acceleration) stated in English
- ISAAC (Novak, 1976)
  - Read, understand, solve and draw pictures of physics problems stated in English
- ALBERT (Oberem, 1987)
  - A tutoring system that understands and solves physics (kinematics) problems but can teach a student how to solve them
- Chang et al. (2014)
  - Simple vector addition, tension, and gravitation ranking problems
- Klenk et al. (2005)
  - Physical reasoning problems by analyzing sketches



As shown in the Figure, ∠MAO = 30⁰ and the radius of the circle with center O is 4cm. Find the value of x.

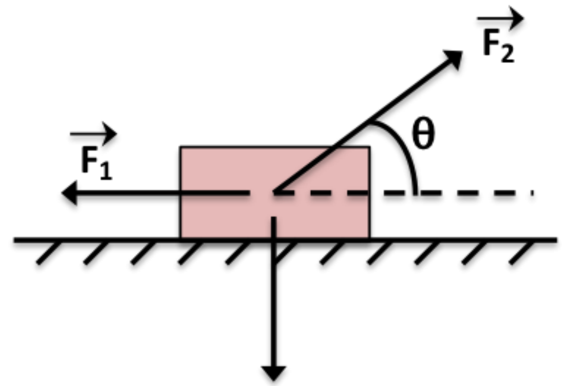
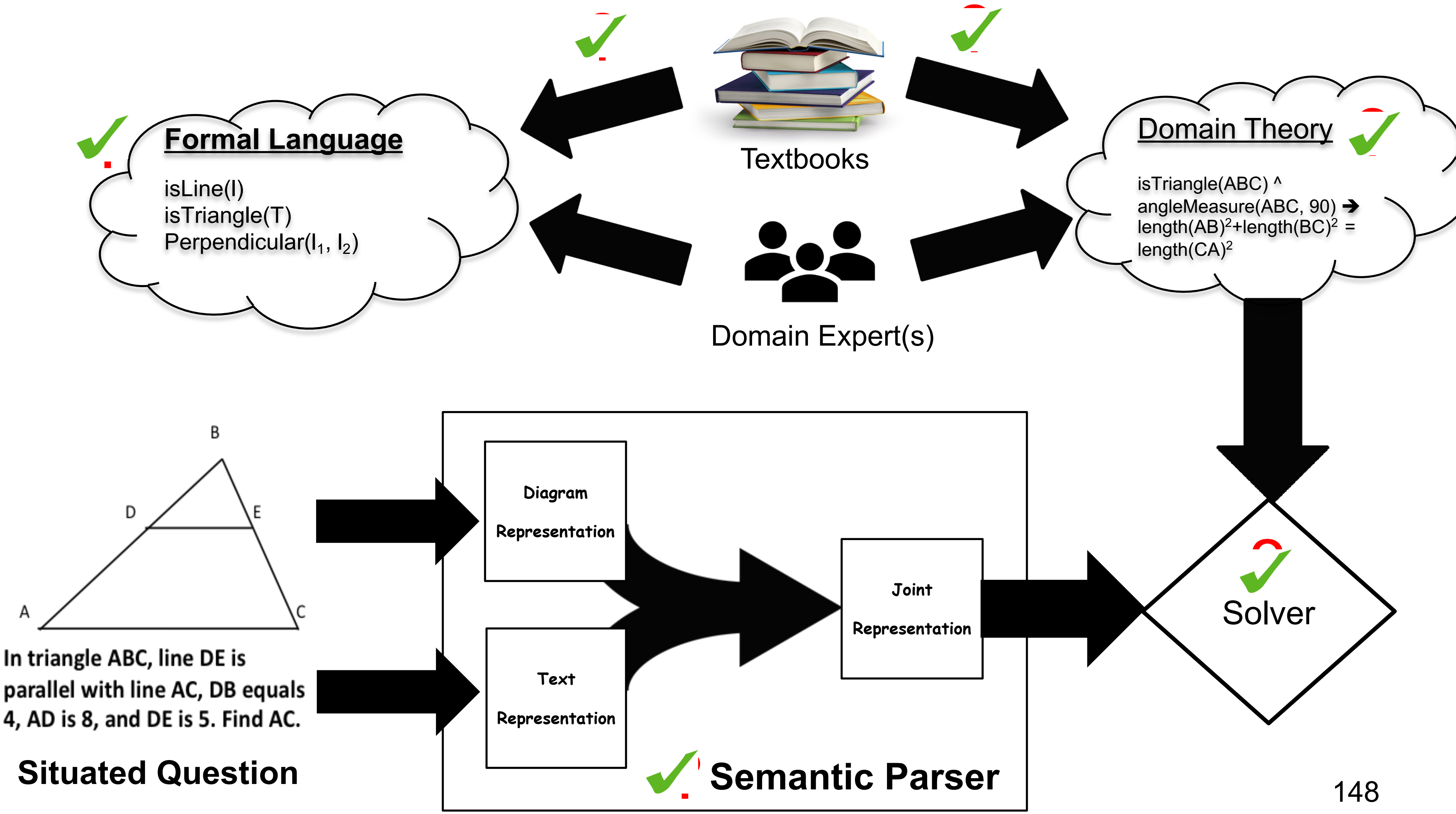


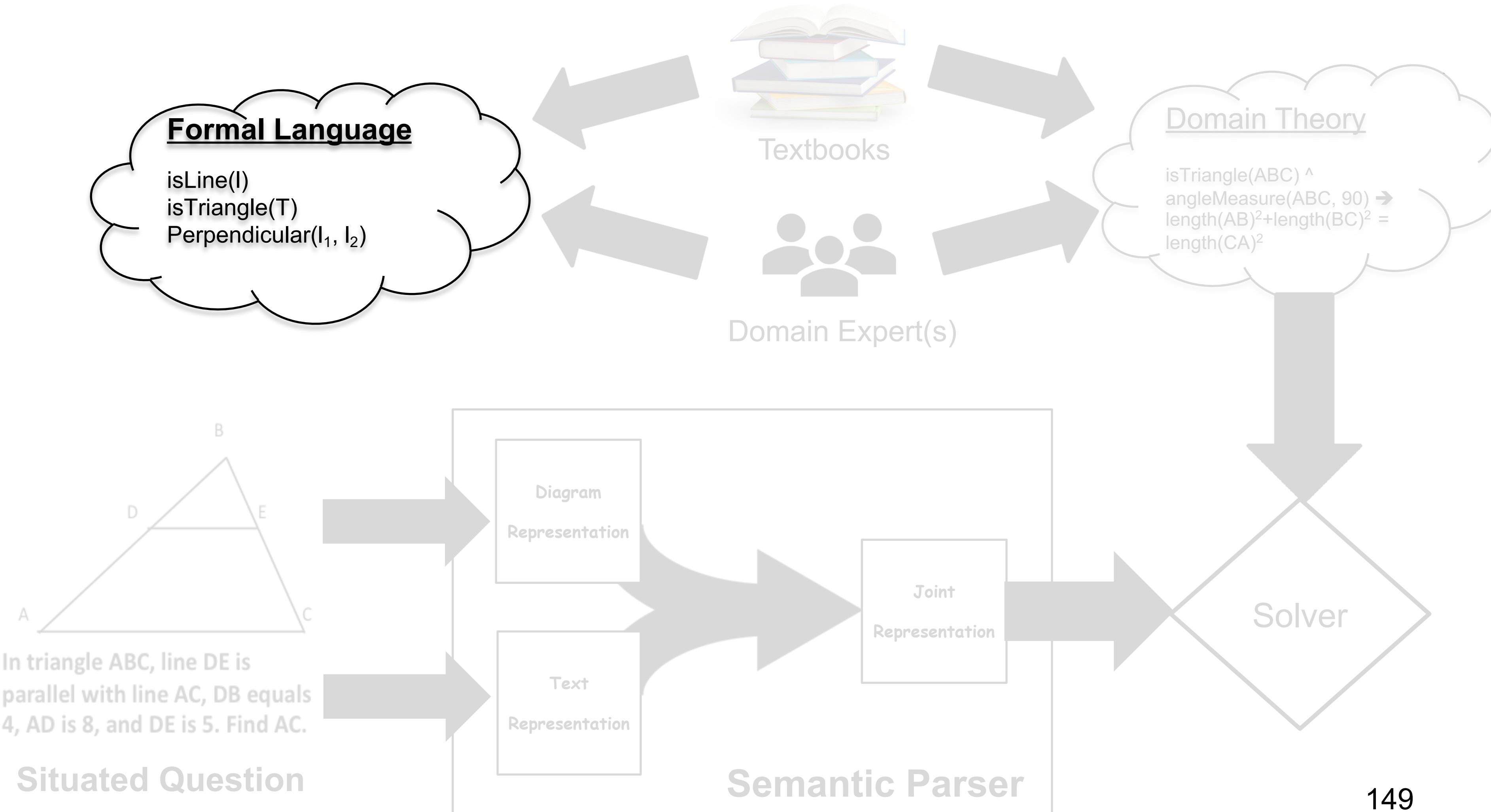
Figure above shows three forces applied to a trunk that moves leftward by 3.00 m over a frictionless floor. The force magnitudes are F_1 = 5.00N, F_2 = 9.00N, and F_3 = 3.00N, and the indicated angle is θ = 60.0° . During the displacement, what is the net work done on the trunk by the three forces?

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi and Oren Etzioni. Diagram understanding in geometry questions. In AAAI 2014
Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni and Clint Malcolm. Solving geometry problems: combining text and diagram interpretation. In EMNLP 2015
Mrinmaya Sachan, Avinava Dubey and Eric P. Xing. From Textbooks to Knowledge: A Case Study in Harvesting Axiomatic Knowledge from Textbooks to Solve Geometry Problems*. In EMNLP 2017*
Mrinmaya Sachan, Eduard Hovy and Eric P. Xing. Discourse in Multimedia: A Case Study in Information Extraction.
Mrinmaya Sachan, Eric P. Xing. Parsing to Programs: A Framework for Situated QA*. In KDD 2018*
Mrinmaya Sachan, Avinava Dubey, Tom Mitchell, Dan Roth and Eric P. Xing. Learning Pipelines with Limited Data and Domain Knowledge: A Study in Parsing Physics Problems. In NIPS 2018.
Mrinmaya Sachan, Minjoon Seo, Hannaneh Hajishirzi and Eric P. Xing. Parsing to Programs: A Framework for Situated Question Answering

Parsing to Programs



Formal Language



The Formal Language

- A subset of typed first-order logic
  - **Constants**
    - Known numbers or geometry/physics entities
      - e.g. 5 cm, 60^0, 3.00m, 5.00N
  - **Variables**
    - Unknown numbers or geometry/physics entities
      - e.g. O, AB, F_1, θ
  - **Predicates**
    - Geometric/Physical or arithmetic relations
      - e.g. *isLine*, *isTriangle*, *isAtRest*
  - **Functions**
    - Properties of geometrical/physical entities
      - e.g. *lengthOf*, *areaOf, mass, distance, force, momentum, work…*

The Formal Language

- Every element in the language has either **boolean** (e.g. true), **numeric** (e.g. 4), or **entity** (e.g., *line*, *circle*, *object*, *force*, *mass*, *velocity*) type.
- We refer to all symbols in the language as **concepts**.
- We use the term **literal** to refer to the application of a predicate to a sequence of arguments (e.g., IsTriangle(ABC)).
- Questions are represented as (Weighted) Logical formulas containing constants, variables, functions, existential quantifiers and conjunctions over literals (e.g., $\exists x$, IsTriangle(x) $\wedge$ isIsosceles(x)).
  - Weight corresponds to our model's confidence in it.

Lexicon

- We built lexicon from training data and textbooks
- Lexicon maps geometry-related words (or phrases) to *concepts*
- Some concepts are obtained via simple regular expressions
- Single word can map to two or more concepts

Word or phrase	Concept
“Perpendicular”	`Perpendicular`
“Lies on”	`PointLiesOnLine`, `PointLiesOnCircle`
“CD”	`line`, `arc`
“ABC”	`triangle`, `angle`

Question Parsing

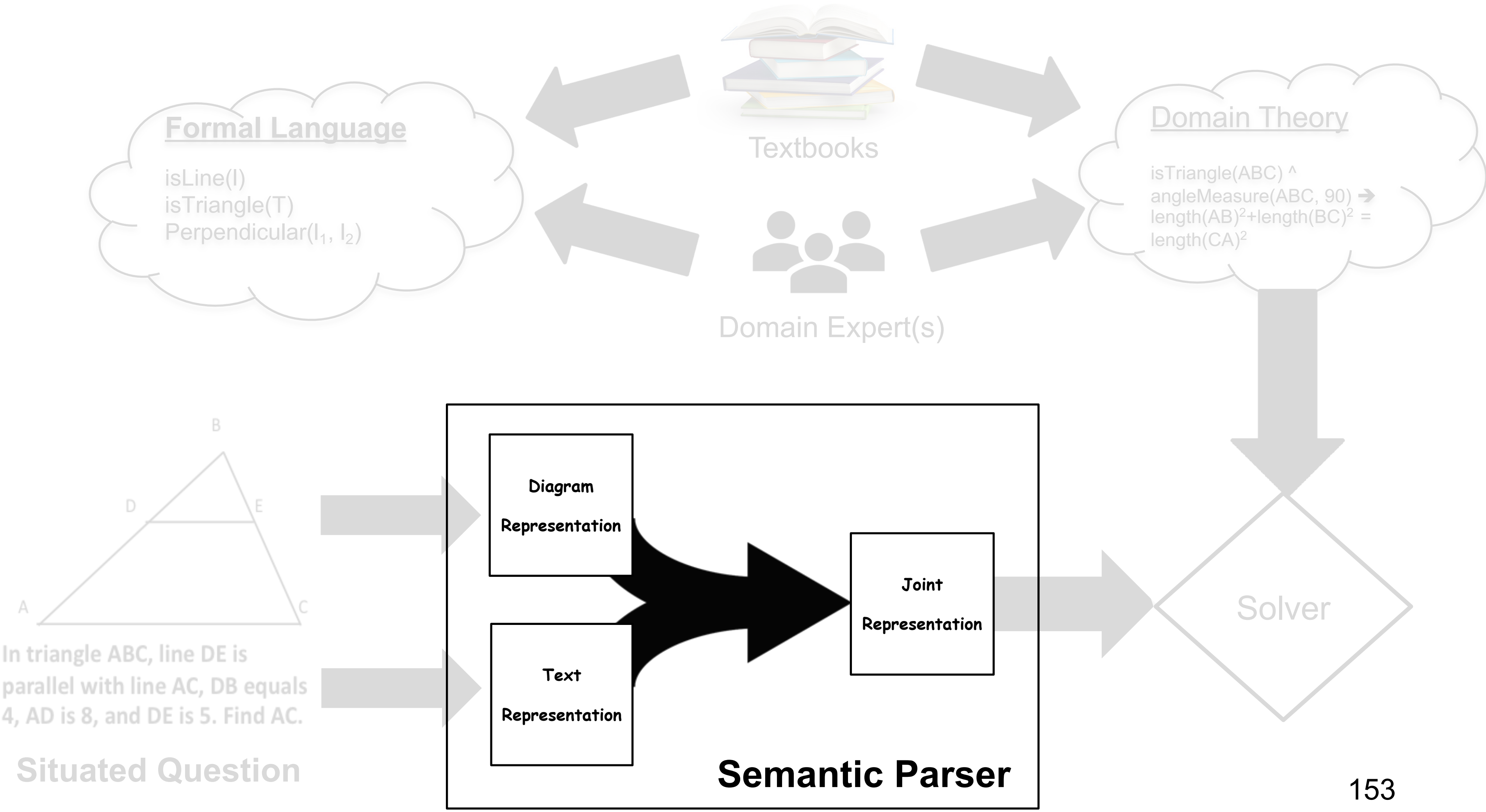
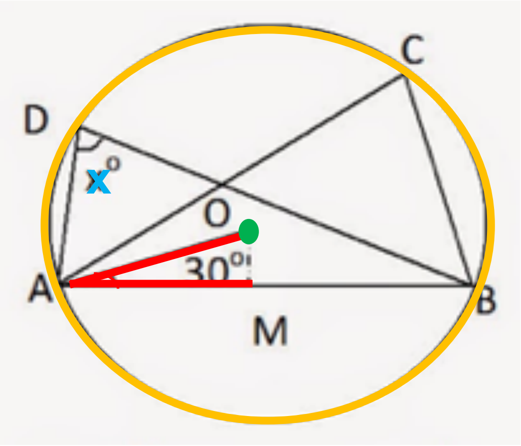


Diagram Parsing

As shown in the figure, ∠MAO = 30^0 and the radius of the circle with center O is 4cm. Find the value of x.



G-Aligner - **Seo et. al. 2014**

Use both diagram and text

Text Parsing

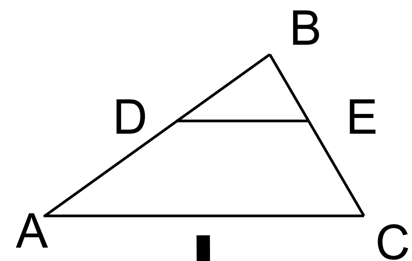
GEOS - **Seo et. al. 2015**

- Concept Identification
  - Identify numbers and explicit variables (e.g. “5”, “AB”, “O”) using regular expressions
- Relation Identification
  - Predict if a particular relation holds between concepts

Diagram-aided text parsing

Text Input

In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
(a) 9 (b) 10 (c) 12.5 (d) 15 (e) 17



Our method

Over-generated literals	**Text scores**	**Diagram scores**
IsTriangle(ABC)	*0.96*	*1.00*
Parallel(AC, DE)	*0.91*	*0.99*
Parallel(AC, DB)	*0.74*	*0.02*
Equals(LengthOf(DB), 4)	*0.97*	*n/a*
Equals(LengthOf(AD), 8)	*0.94*	*n/a*
Equals(LengthOf(DE), 5)	*0.94*	*n/a*
Equals(4, LengthOf(AD))	*0.31*	*n/a*
…	…	…

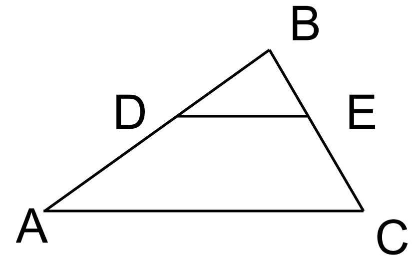
Selected subset

Logical form

IsTriangle(ABC) ∧ *Parallel(AC, DE)* ∧
Equals(LengthOf(DB), 4) ∧ *Equals(LengthOf(AD), 8)* ∧
Equals(LengthOf(DE), 5) ∧ *Find(LengthOf(AC))*

Step 1. Literal over-generation

In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
(a) 9 (b) 10 (c) 12.5 (d) 15 (e) 17



Over-generated literals

IsTriangle(ABC)
Parallel(AC, DE)
Parallel(AC, DB)
Equals(LengthOf(DB), 4)
Equals(LengthOf(AD), 8)
Equals(LengthOf(DE), 5)
Equals(4, LengthOf(AD))
…

Step 1. Generating literals

“Lines AB and CD are perpendicular to EF”

```
IsLine(AB)
IsLine(CD)
IsLine(EF)
Perpendicular(AB, CD)
Perpendicular(CD, EF)
Perpendicular(AB, EF)
```

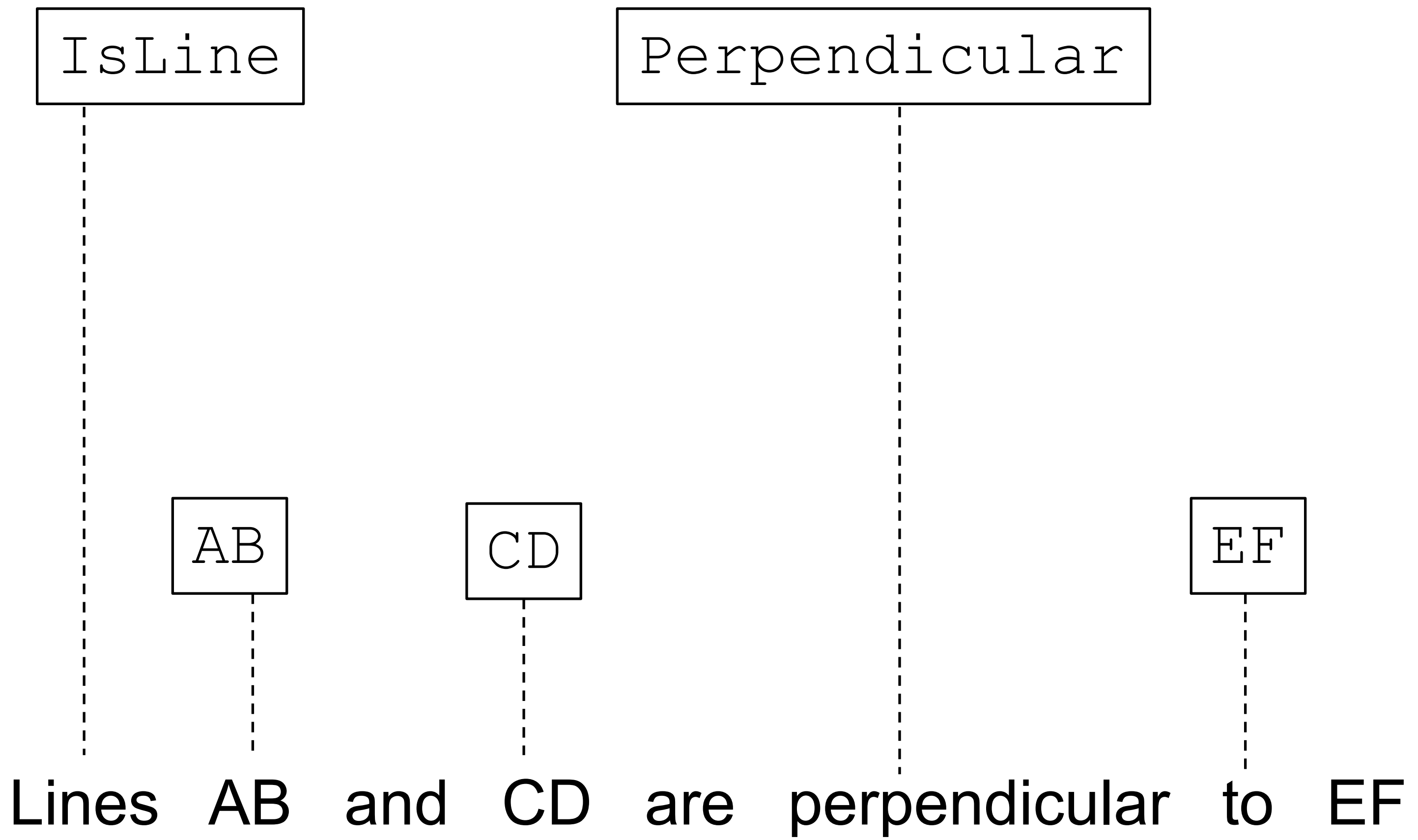
Step 1. Generating literals

"Lines AB and CD are perpendicular to EF"

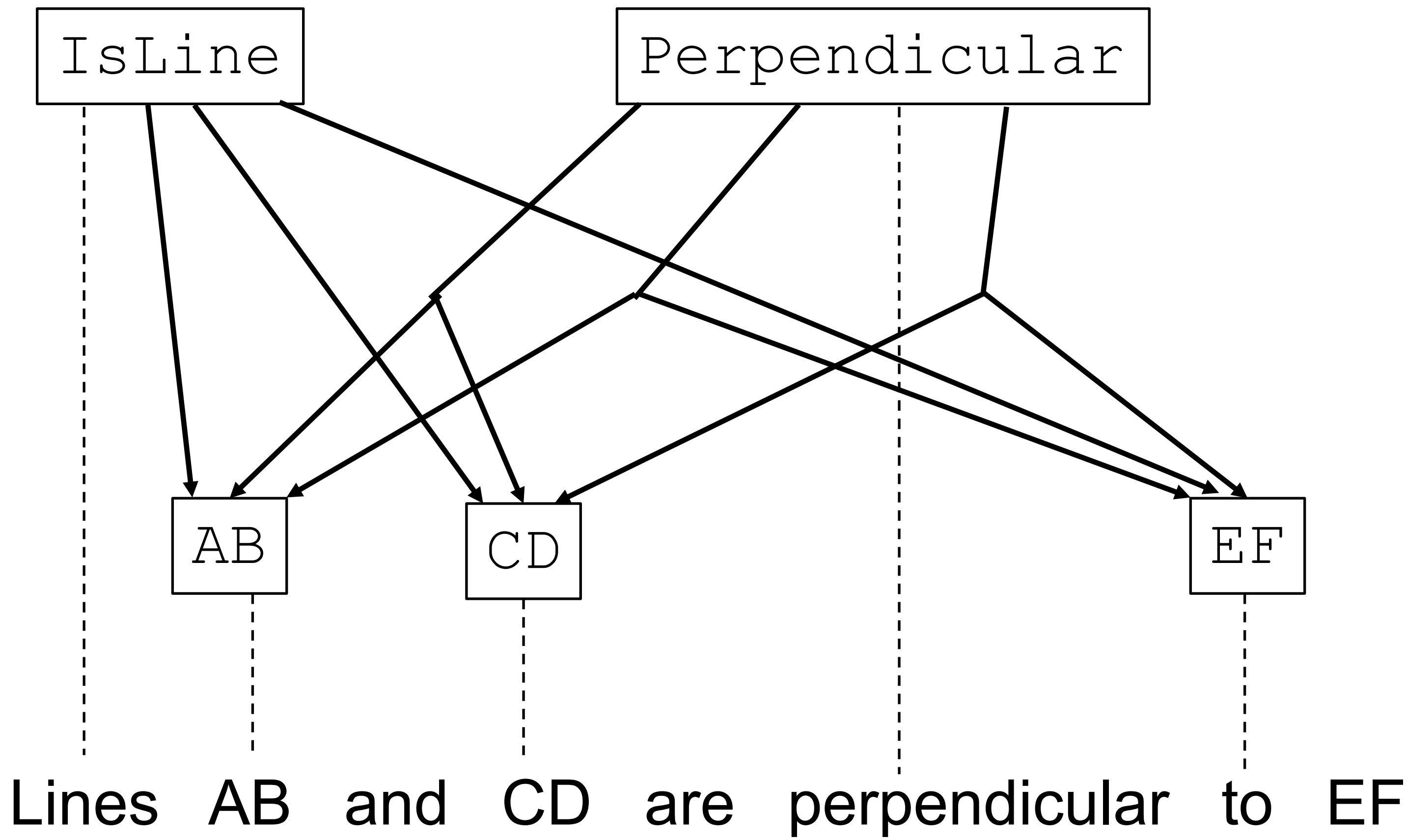
```
IsLine(AB)
IsLine(CD)
IsLine(EF)
Perpendicular(AB, CD)
Perpendicular(CD, EF)
Perpendicular(AB, EF)
```

Red literals *are false.*

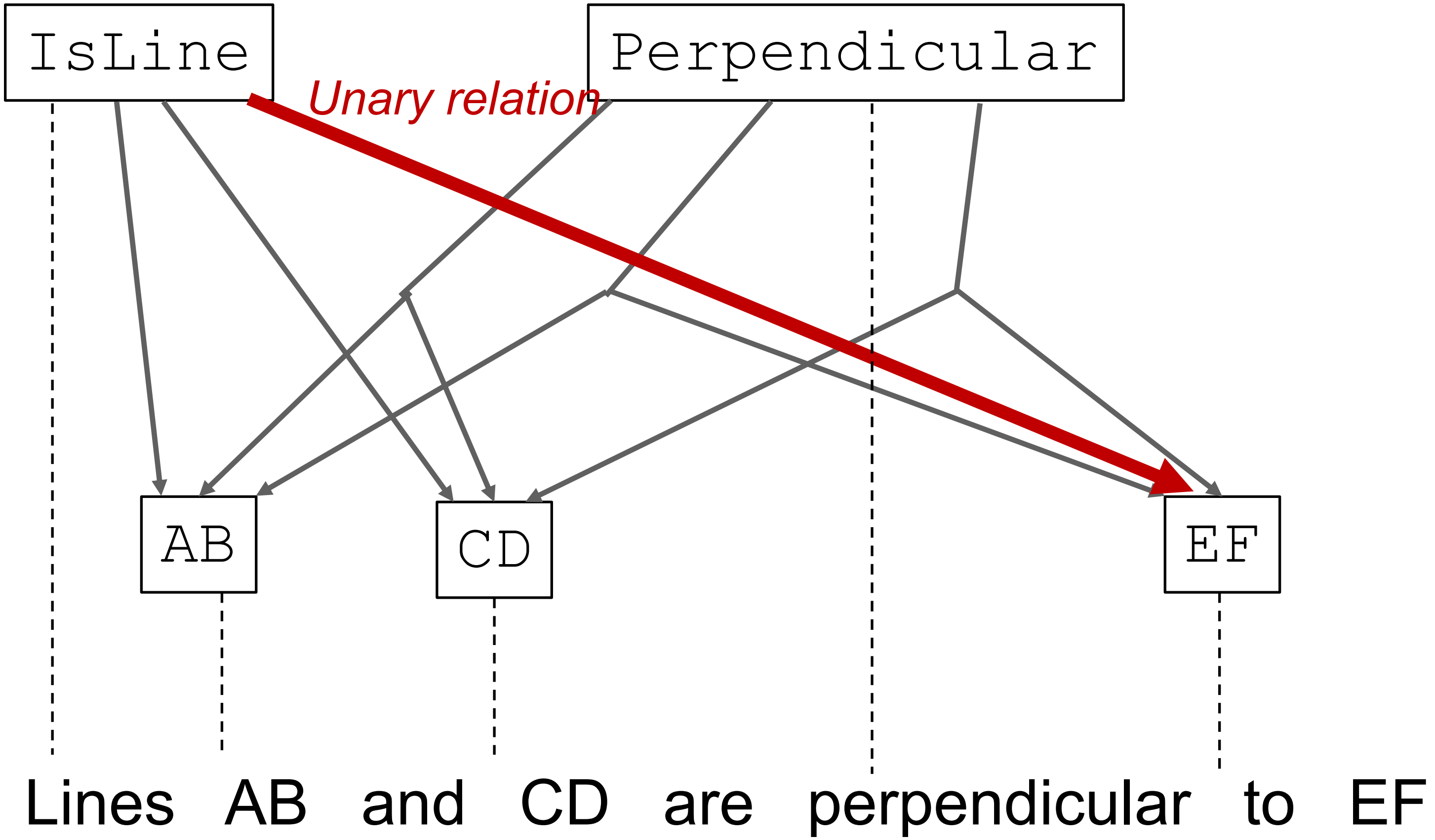
Concepts


IsLine
Perpendicular
AB
CD
EF
Lines AB and CD are perpendicular to EF

Relations

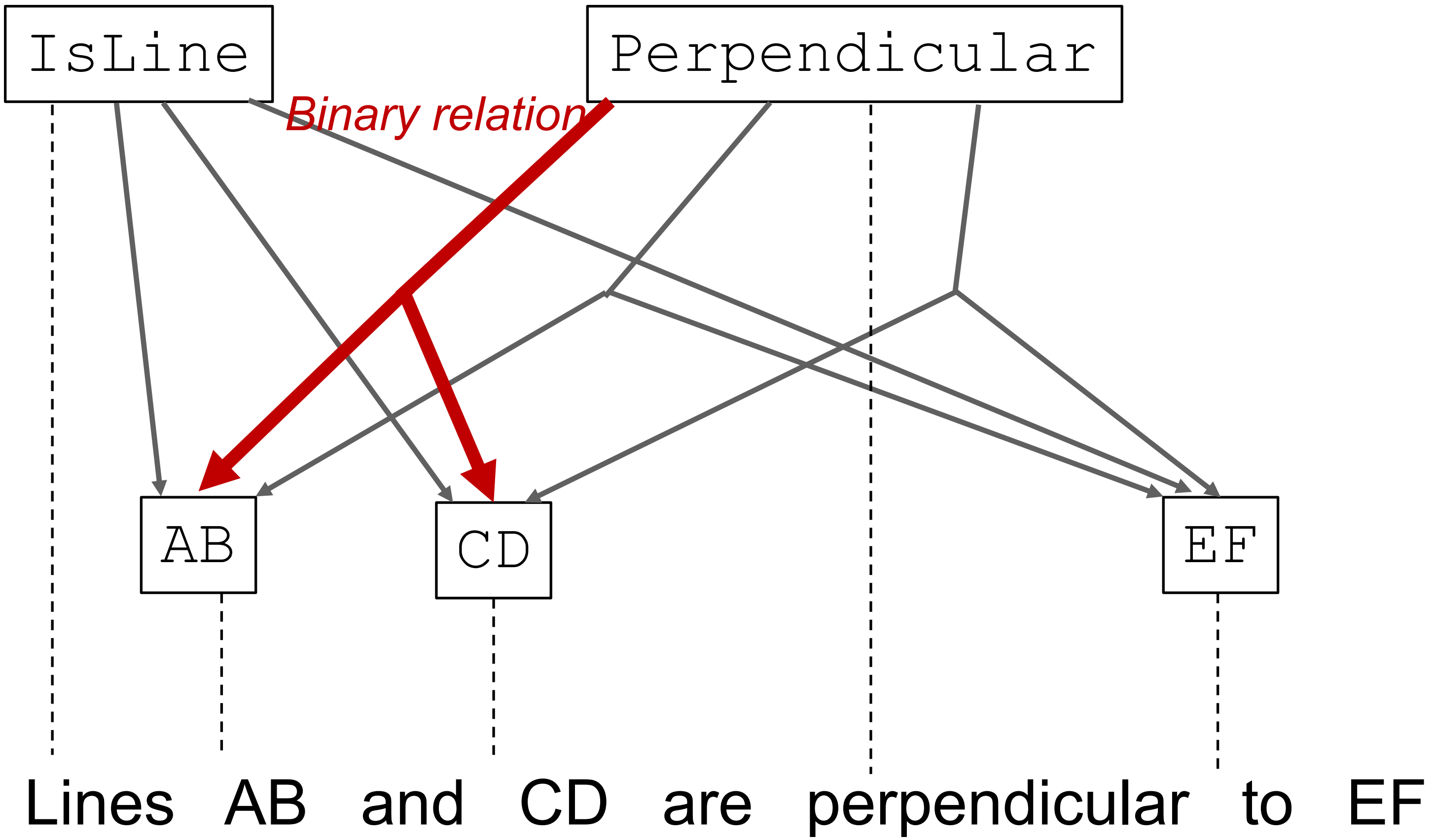

IsLine
Perpendicular
AB
CD
EF
Lines AB and CD are perpendicular to EF

Relations



IsLine(EF)

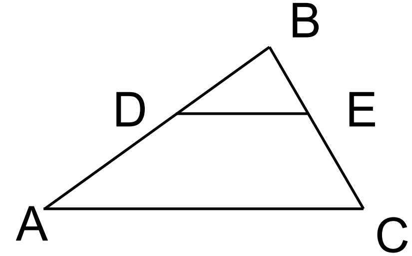
Relations



Perpendicular(AB, CD)

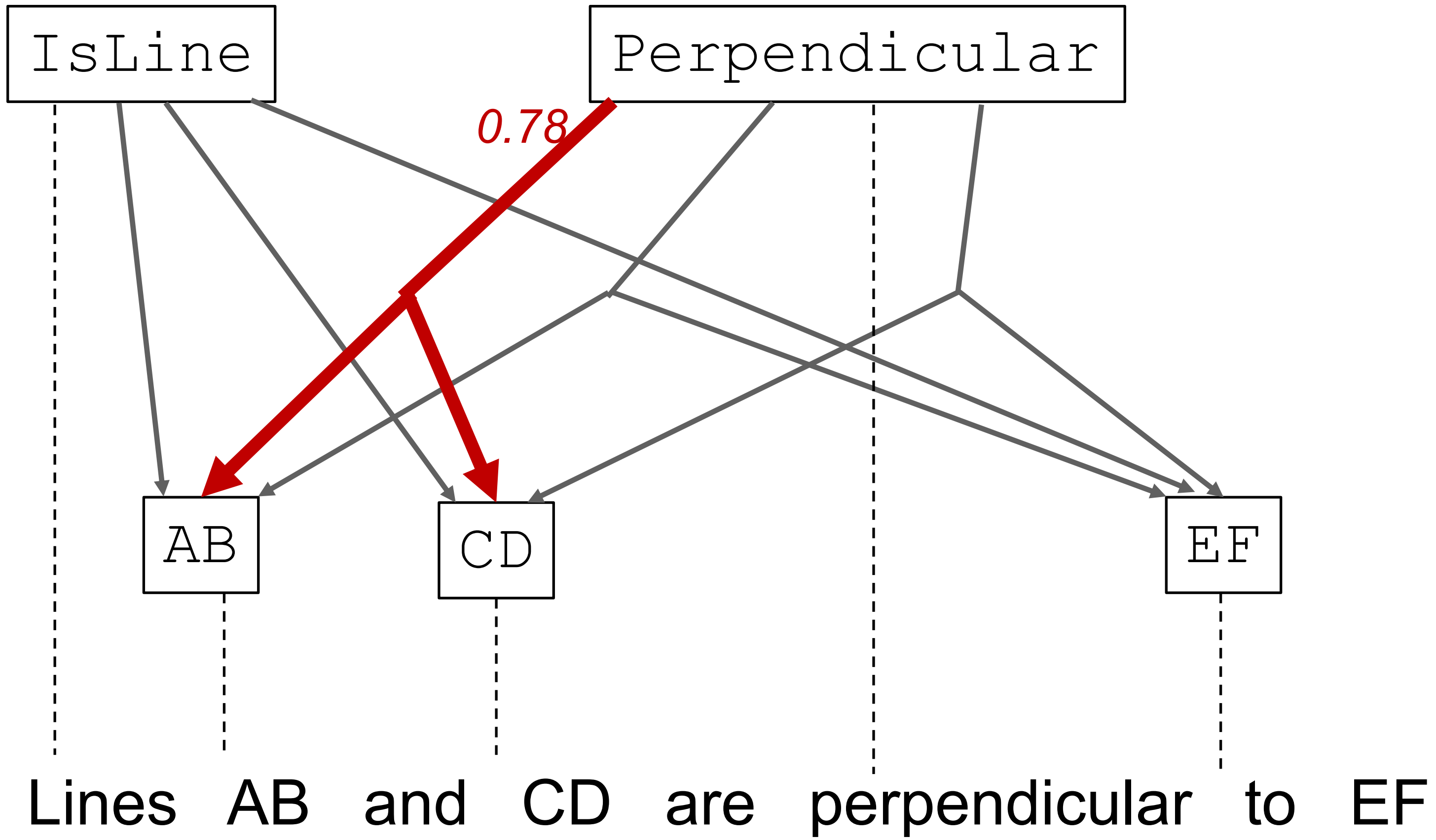
Step 2. Text scores of literals

In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
(a) 9 (b) 10 (c) 12.5 (d) 15 (e) 17



Over-generated literals	**Text scores**
IsTriangle(ABC)	*0.96*
Parallel(AC, DE)	*0.91*
Parallel(AC, DB)	*0.74*
Equals(LengthOf(DB), 4)	*0.97*
Equals(LengthOf(AD), 8)	*0.94*
Equals(LengthOf(DE), 5)	*0.94*
Equals(4, LengthOf(AD))	*0.31*
…	…

Relation score



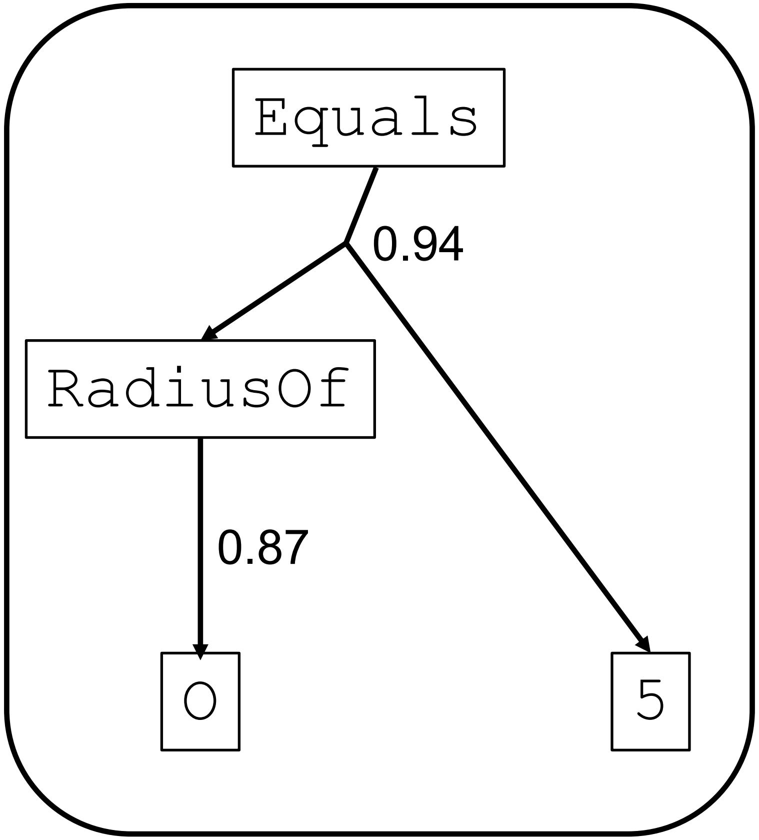
Relation classification

- **Supervision**: annotated logical forms
- **Training data**: all possible relations from training questions
  - Relations found in annotations: **positive**
  - All others: **negative**

```
IsLine->AB
IsLine->CD
IsLine->EF
Perpendicular->AB, CD
Perpendicular->CD, EF
Perpendicular->AB, EF
```

- Logistic regression with L2 regularization
- Features:
  - Stanford dependency parse
  - Part of speech tags
  - Type of concept (line, circle, triangle, predicate, etc.)

Text scores of literals



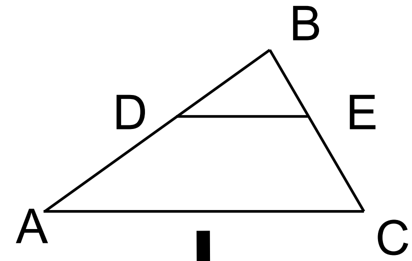
"Circle O has radius of 5"

$$\mathcal{A}_{text}(l) = \sum \log P_\theta(y_i = 1 | r_i, t)$$

l	Literal
y_i	Label for edge
r_i	Edge (relation)
t	Question text
θ	Logistic regression parameters to be learned

Step 3. Diagram scores of literals

In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
(a) 9   (b) 10   (c) 12.5   (d) 15   (e) 17

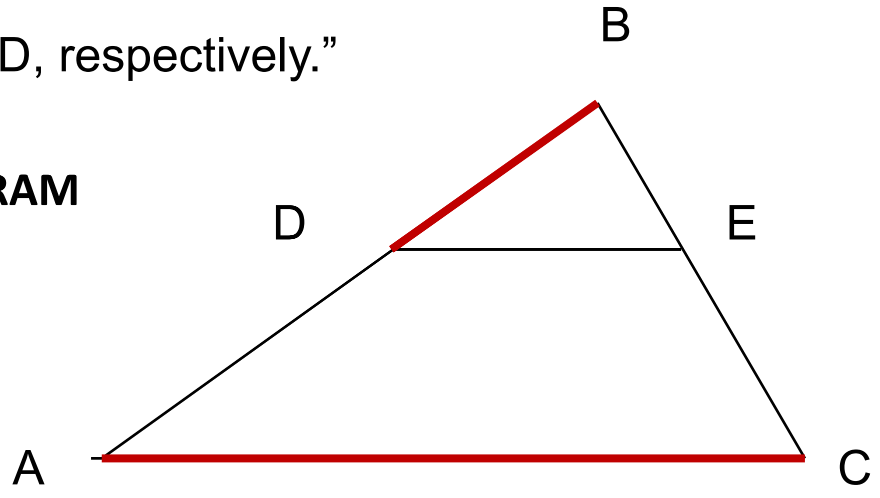


Over-generated literals	Text scores	Diagram scores
IsTriangle(ABC)	*0.96*	*1.00*
Parallel(AC, DE)	*0.91*	*0.99*
Parallel(AC, DB)	*0.74*	*0.02*
Equals(LengthOf(DB), 4)	*0.97*	*n/a*
Equals(LengthOf(AD), 8)	*0.94*	*n/a*
Equals(LengthOf(DE), 5)	*0.94*	*n/a*
Equals(4, LengthOf(AD))	*0.31*	*n/a*
…	…	…

Step 3. Diagram scores of literals

"AC and DB are parallel with DE and AD, respectively."

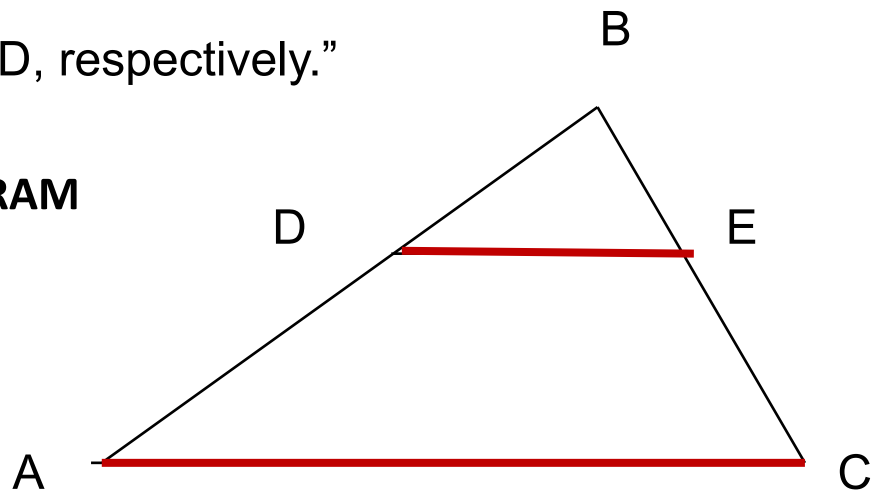
	TEXT	DIAGRAM
`Parallel(AC, DB)`	0.74	0.02



Step 3. Diagram scores of literals

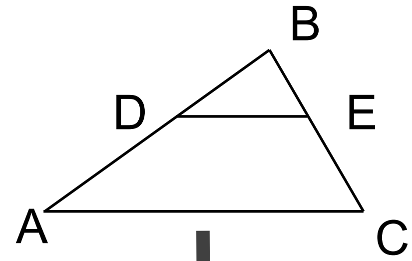
"AC and DB are parallel with DE and AD, respectively."

	TEXT	DIAGRAM
`Parallel(AC, DB)`	0.74	0.02
`Parallel(AC, DE)`	**0.78**	**0.99**



Step 4. Subset selection

In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
(a) 9 (b) 10 (c) 12.5 (d) 15 (e) 17



Over-generated literals	Text scores	Diagram scores
IsTriangle(ABC)	*0.96*	*1.00*
Parallel(AC, DE)	*0.91*	*0.99*
Parallel(AC, DB)	*0.74*	*0.02*
Equals(LengthOf(DB), 4)	*0.97*	*n/a*
Equals(LengthOf(AD), 8)	*0.94*	*n/a*
Equals(LengthOf(DE), 5)	*0.94*	*n/a*
Equals(4, LengthOf(AD))	*0.31*	*n/a*
…	…	…

Selected subset

IsTriangle(ABC) ∧ *Parallel(AC, DE)* ∧
Equals(LengthOf(DB), 4) ∧ *Equals(LengthOf(AD), 8)* ∧
Equals(LengthOf(DE), 5) ∧ *Find(LengthOf(AC))*

Step 4. Subset selection

L

IsTriangle(ABC)
Parallel(AC, DE)
Parallel(AC, DB)
Equals(LengthOf(DB), 4)
Equals(LengthOf(AD), 8)
Equals(LengthOf(DE), 5)
Equals(4, LengthOf(AD))
Equals(LengthOf(AC), 4)
Parallel(DE, DB)
Equals(LengthOf(DB, 8)
Find(LengthOf(AC))
Find(LengthOf(DE))
Find(LengthOf(DB))
…

$\longrightarrow$

L^*

IsTriangle(ABC)
Parallel(AC, DE)
Equals(LengthOf(DB), 4)
Equals(LengthOf(AD), 8)
Equals(LengthOf(DE), 5)
Find(LengthOf(AC))

- High text and diagram scores
- Cover all facts in text
- Literals don't conflict

$$L^* = \underset{L' \subset L}{\operatorname{argmax}} \mathcal{F}(L')$$

0ptimization algorithm

$$L^* = \underset{L' \subset L}{\operatorname{argmax}} \mathcal{F}(L')$$

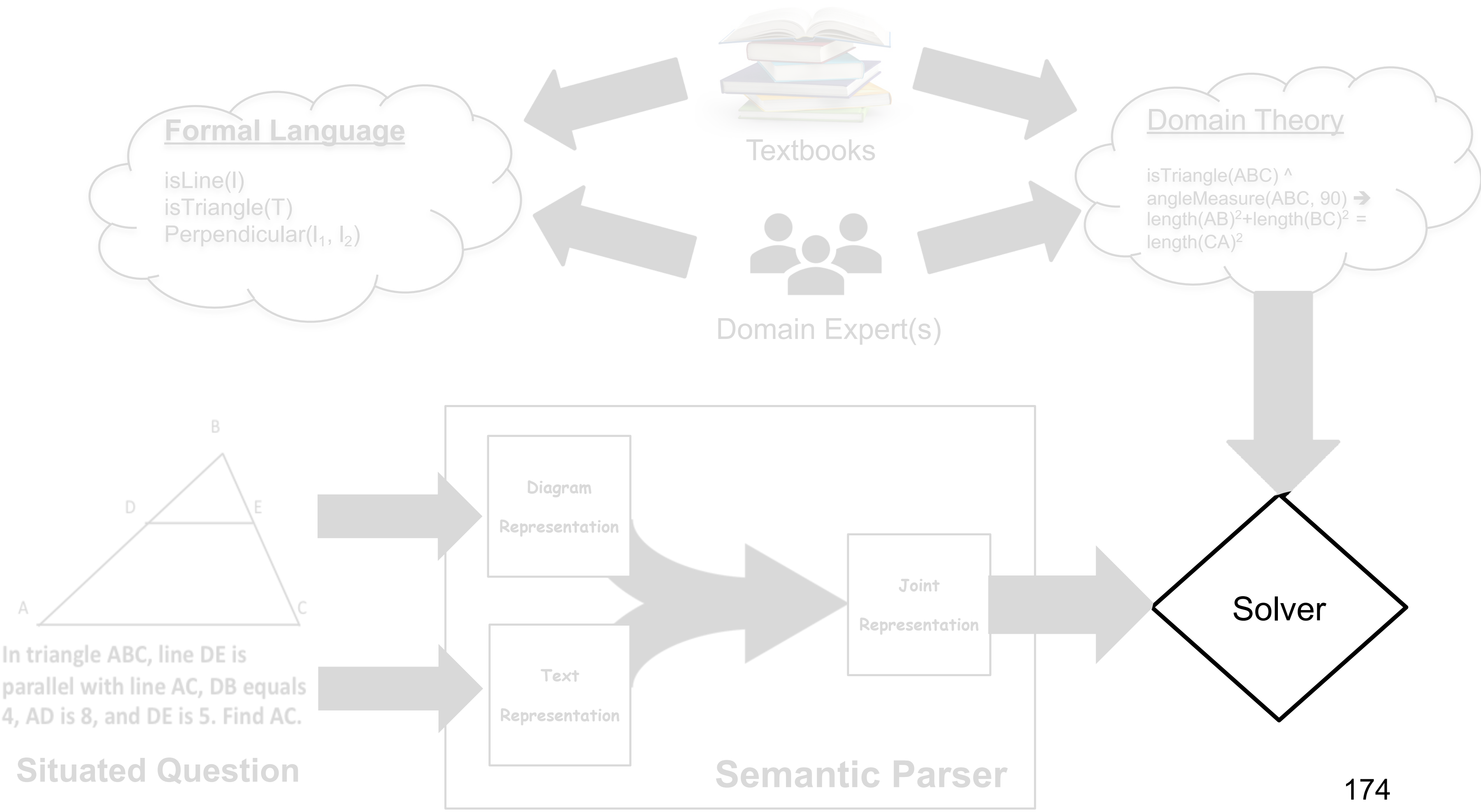
Bad news: combinatorial optimization is NP-hard
Good news: objective function is *submodular*

Greedy algorithm efficiently finds a solution with bounded distance to the optimum.

Starting from empty set, greedily add the next best literal to the set.

$$l_j = \underset{l_j \in L \setminus L'}{\operatorname{argmax}} \mathcal{F}\left(L' \cup \{l_j\}\right) - \mathcal{F}(L')$$

Solver


Formal Language
isLine(l)
isTriangle(T)
Perpendicular(l_1, l_2)
Textbooks
Domain Expert(s)
Domain Theory
isTriangle(ABC) ^
angleMeasure(ABC, 90) →
length(AB)^2+length(BC)^2 =
length(CA)^2
B
D
E
A
C
In triangle ABC, line DE is parallel with line AC, DB equals 4, AD is 8, and DE is 5. Find AC.
Situated Question
Diagram Representation
Text Representation
Joint Representation
Semantic Parser
Solver

Programmatic Solving: **Numerical solver**

- Translate literals to numeric equations

Literal	Equation
Equals(LengthOf(AB),d)	$(A_x-B_x)^2+(A_y-B_y)^2-d^2 = 0$
Parallel(AB, CD)	$(A_x-B_x)(C_y-D_y)-(A_y-B_y)(C_x-D_x) = 0$
PointLiesOnLine(B, AC)	$(A_x-B_x)(B_y-C_y)-(A_y-B_y)(B_x-C_x) = 0$
Perpendicular(AB,CD)	$(A_x-B_x)(C_x-D_x)+(A_y-B_y)(C_y-D_y) = 0$

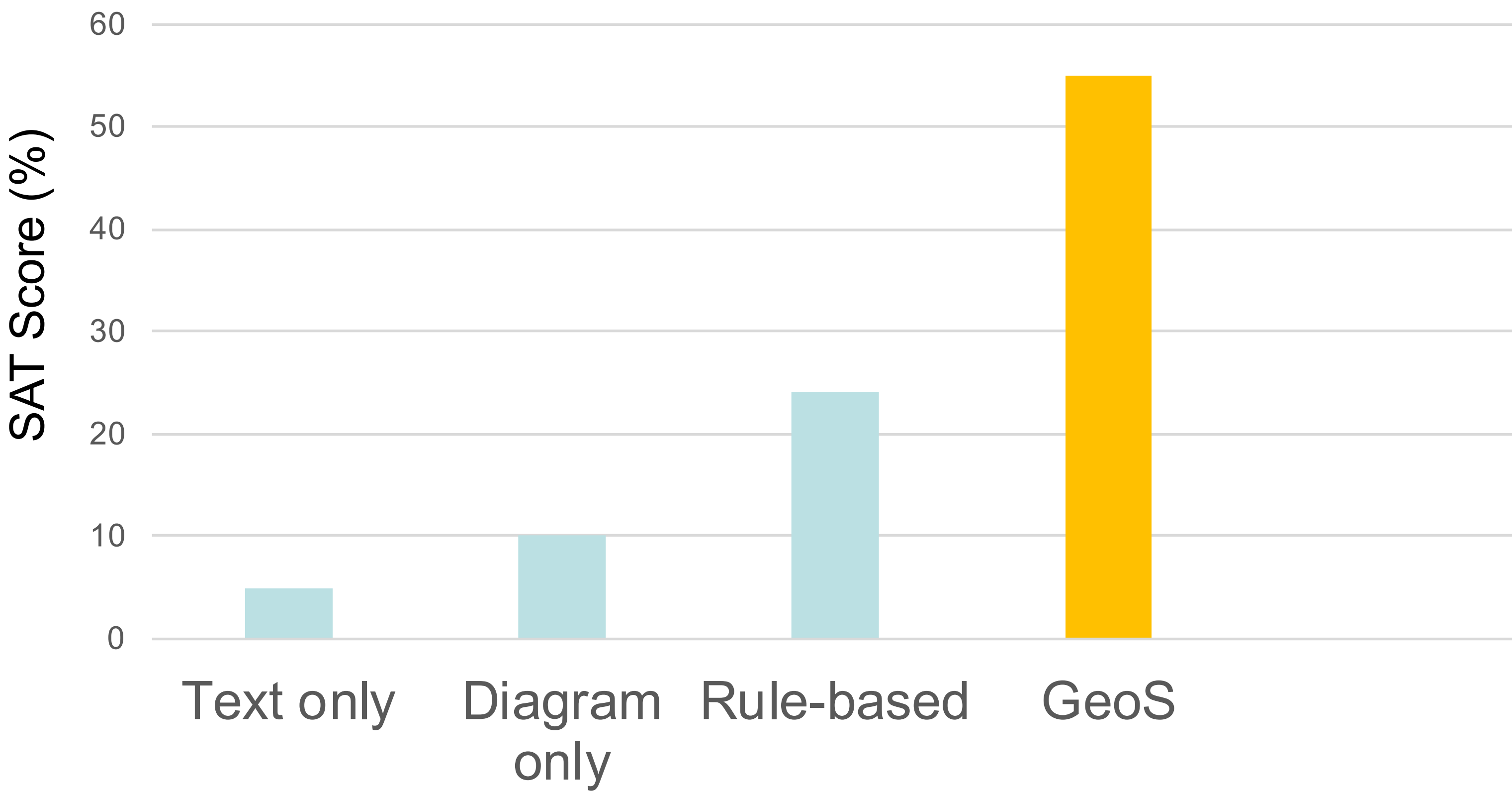
- Find the solution to the equation system
- Use off-the-shelf numerical minimizers (Wales and Doye, 1997; Kraft, 1988)
- Numerical solver can choose <u>not</u> to answer question

Dataset

- **Training questions** (67 questions, 121 sentences)
  - Seo et al., 2014
  - High school geometry questions
- **Test questions** (119 questions, 215 sentences)
  - We collected them
  - SAT (US college entrance exam) geometry questions
- Manually annotated the text parse of all questions
- Dataset is publicly available at:

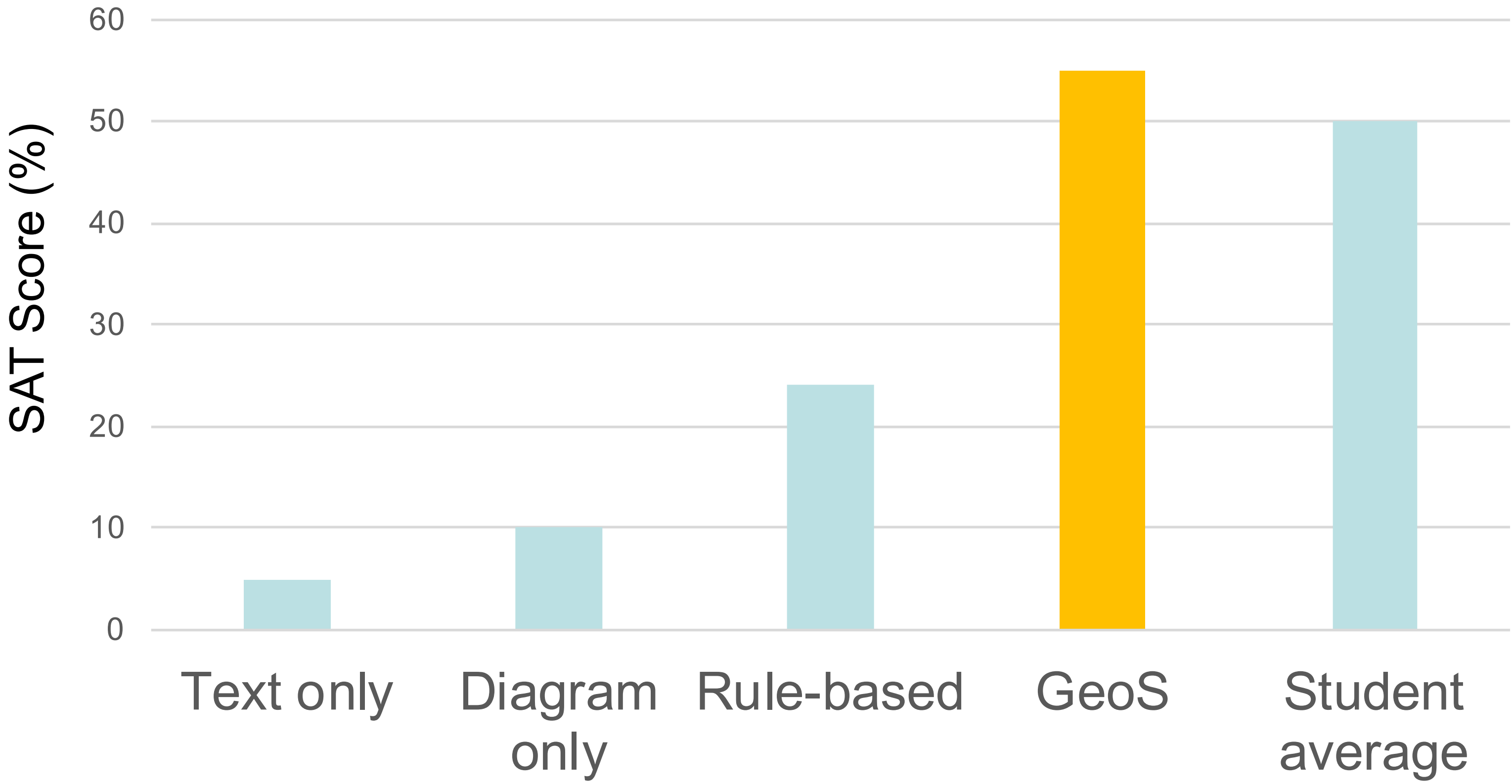
  `geometry.allenai.org`

Results



*** 0.25 penalty for incorrect answer

Results



*** 0.25 penalty for incorrect answer

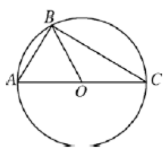
Demo (geometry.allenai.org/demo)

geometry.allenai.org/demo/

Search

GeoS Demo – An End to End Geometry Problem Solver

In the figure to the left, triangle ABC is inscribed in the circle with center O and diameter AC. If AB=AO, what is the degree measure of angle ABO?

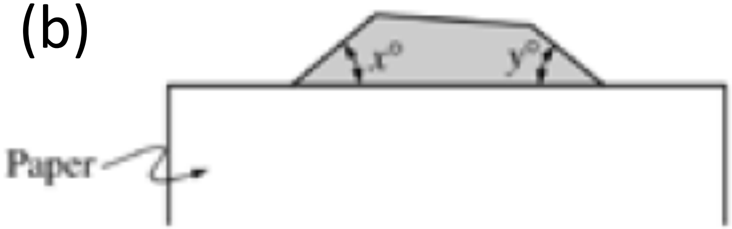


(A) 15°
(B) 30°
(C) 45°
(D) 60°
(E) 90°

Solve Problem

But some are really hard

(b)

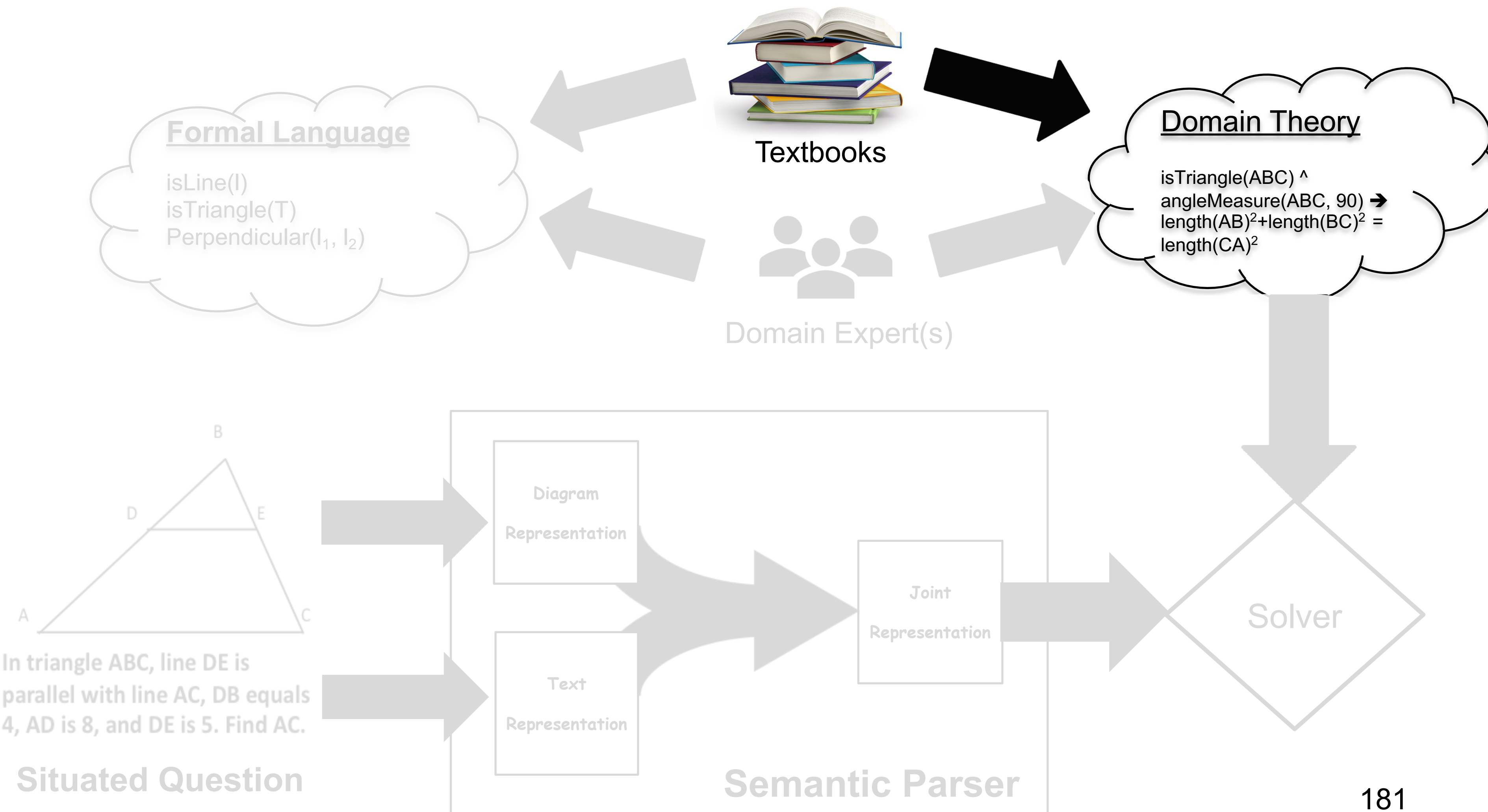


Requires complex reasoning: Cannot understand that the polygon is "hidden"

In the figure at the left, a shaded polygon which has equal angles is partially covered with a sheet of blank paper. If x+y=80, how many sides does the polygon have?

(a) 10 (b) 9 (c) 8 (d) 7 (e) 6

Domain Knowledge



Domain Knowledge

Axiom	Premise	Conclusion
Midpoint Definition	midpoint(M, AB)	length(AM) = length(MB)
Angle Addition	interior(D, ABC)	angle(ABC) = angle(ABD) + angle(DBC)
Supplementary Angles	perpendicular(AB,CD) ∧ liesOn(C,AB)	angle(ACD) + angle(DCB) = 180°
Vertically Opp. Angles	intersectAt(AB, CD, M)	angle(AMC) = angle(BMD)
SSS Congruence	length(AB) = length(DE) ∧ length(BC) = length(EF) ∧ length(CA) = length(FD)	congruent(ABC, DEF)

May be curated by a domain expert or
extracted from textbooks

Axiomatic Solver

Datastructur

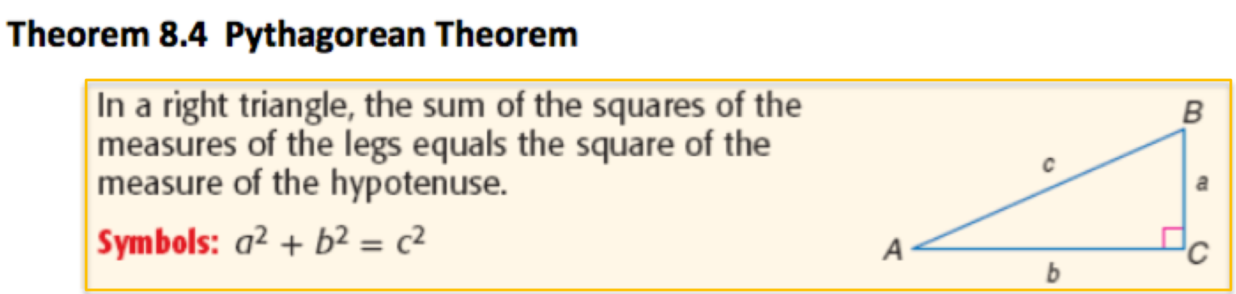
sort point = {A, B, C, D, O, M}
sort line = {AB, BC, CA, BD, DA, OA, OM}
sort angle = {ABC, BCA, CAB, ABD, BDA, DAB, AMO, MOA, OAM, BMO}
sort triangle = {ABC, ABD, AMO}
sort circle = {O}

Where do Axioms come from?

From textbooks to Knowledge

Sachan et. al. 2017

- Key Ideas
  - Leverage the **redundancy** and **shared ordering** across multiple textbooks to harvest axioms.
  - Use rich contextual and **typographical** features extracted from textbooks


Theorem 8.4 Pythagorean Theorem

In a right triangle, the sum of the squares of the measures of the legs equals the square of the measure of the hypotenuse.

Symbols: $a^2 + b^2 = c^2$

isTriangle(ABC) ^ measure(ACB, 90) => $BC^2 + AC^2 = AB^2$

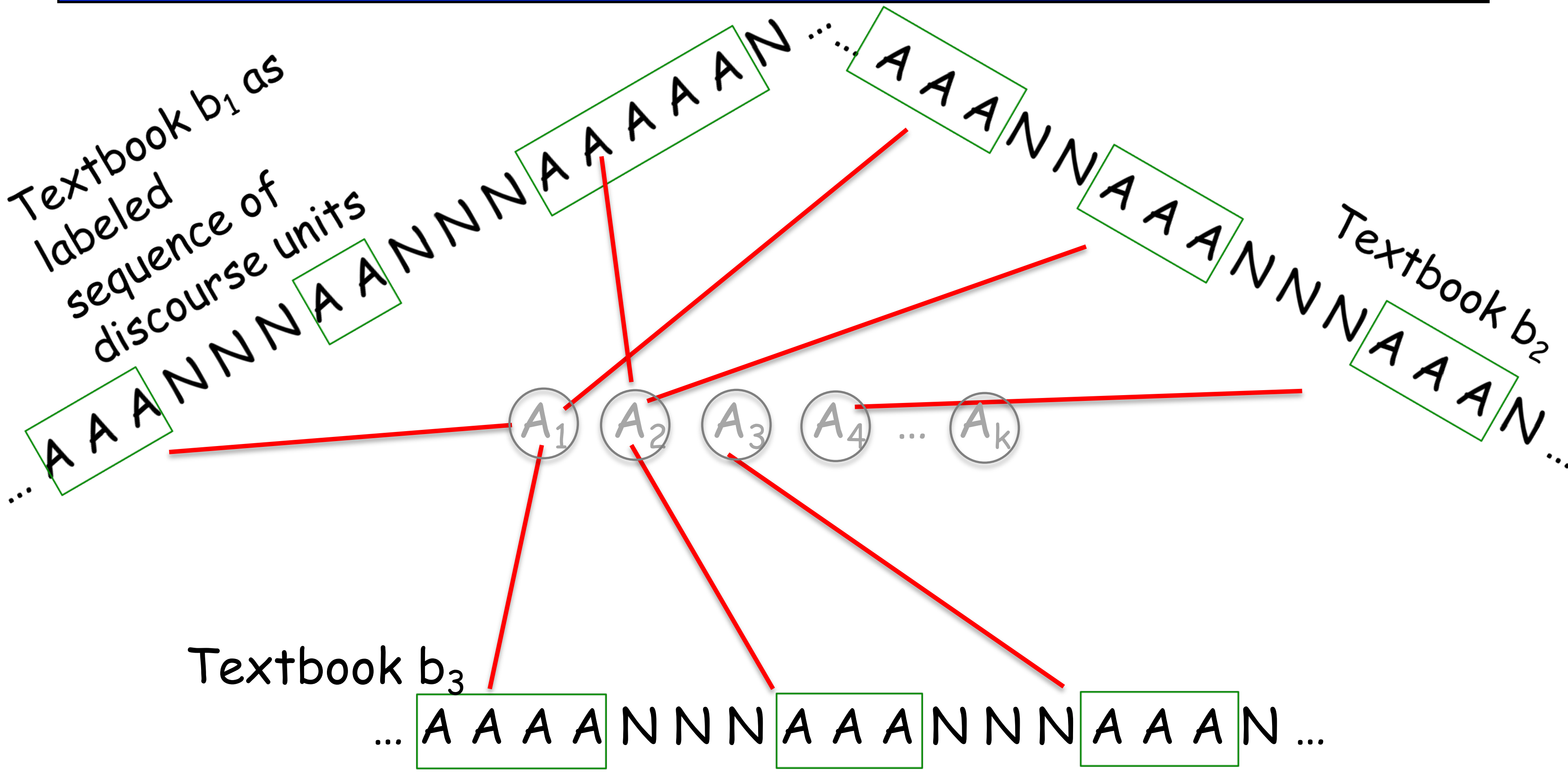
3 stage procedure

- Joint Axiom Identification and Alignment
- Axiom Parsing into horn clause rules
- Horn clause resolver

3 stage procedure

- Joint Axiom Identification and Alignment
- Axiom Parsing into horn clause rules
- Horn clause resolver

Joint Axiom Identification and Alignment



Joint model for Axiom Identification and Alignment

Sequence Labelling CRF

Alignment Model
Log linear model

$$p(\mathbf{Y}, \mathbf{Z} | \{\mathbf{S}_b\}; \boldsymbol{\theta}, \boldsymbol{\phi}) \propto f_{AI}(\mathbf{Y} | \{\mathbf{S}_b\}; \boldsymbol{\theta}) \times f_{AA}(\mathbf{Z} | \mathbf{Y}, \{\mathbf{S}_b\}; \boldsymbol{\phi})$$

Soft
Constraint

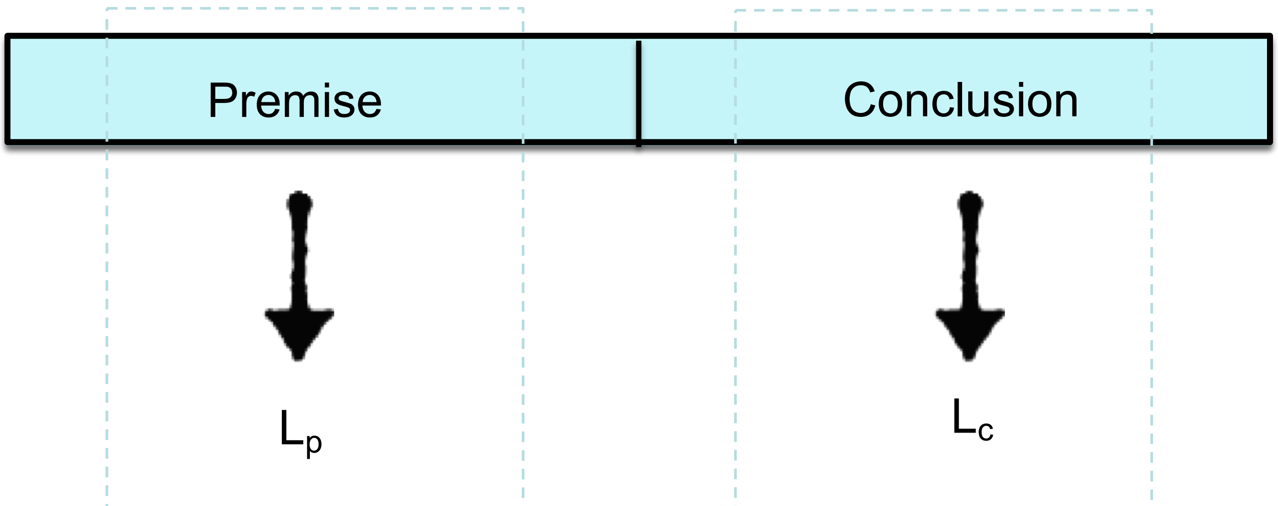
C: Ordering Constraint: If the i^{th} axiom in book b refers to the j^{th} axiom in the global ordering then none of the axioms succeeding the i^{th} axiom can refer to a global axiom preceding j

- EM where we use a Constrained Metropolis Hastings sampler in the E step.
  - Sample Y and Z alternatively
  - For better mixing, we sample Y in blocks

3 stage procedure

- Joint Axiom Identification and Alignment
- Axiom Parsing into horn clause rules
- Horn clause resolver

Base Axiomatic Parser

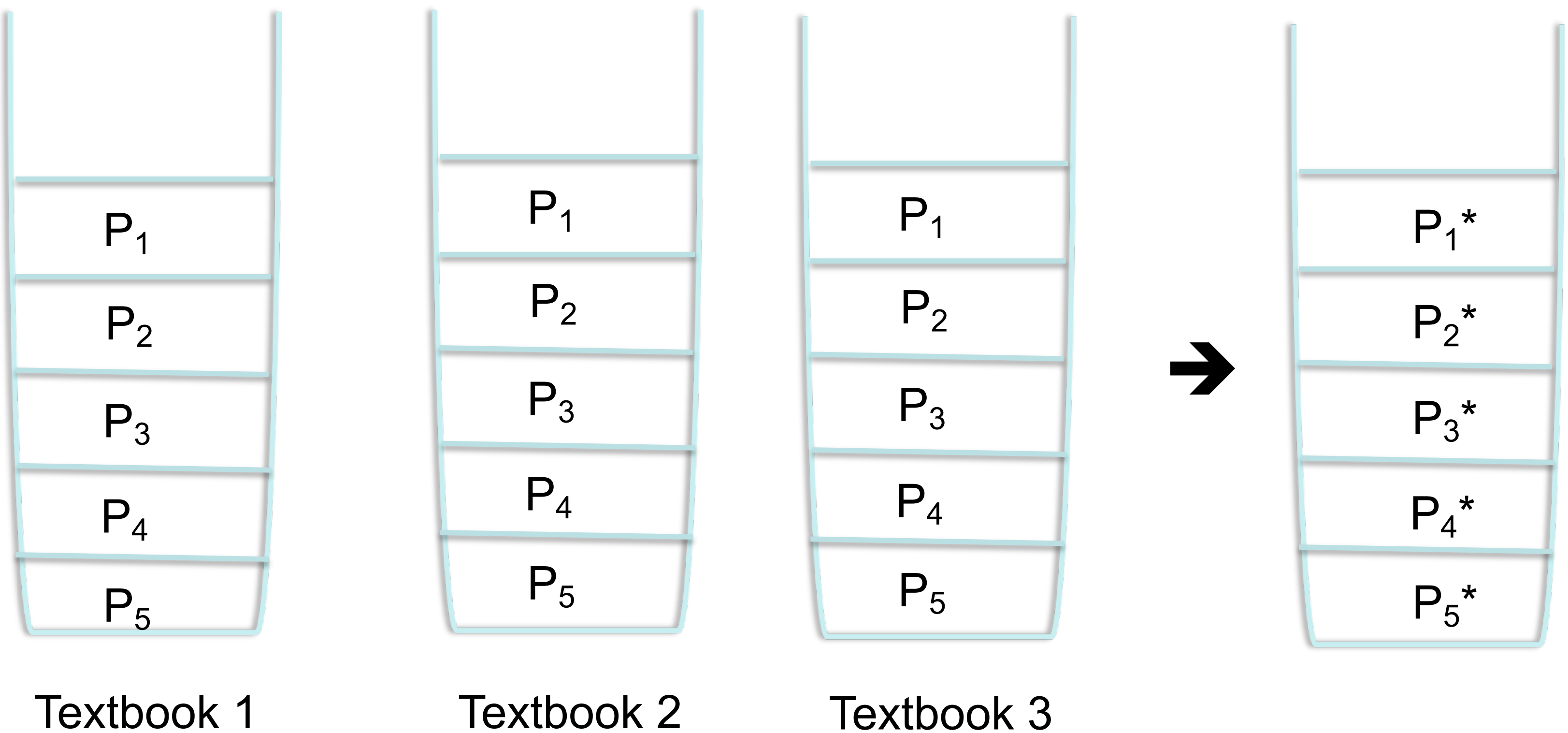


$L_p \Rightarrow L_c$

3 stage procedure

- Joint Axiom Identification and Alignment
- Axiom Parsing into horn clause rules
- Horn clause resolver

Horn Clause Resolver



Majority Voting
Average Score
Learn Source Confidence
Predicate Scoring

Beam of horn clause parses sorted by parse score for each axiom

Dataset for Harvesting Axioms

- Collection of grade 6-10 high school math textbooks by four publishers (20 textbooks) to train our axiom extraction model.

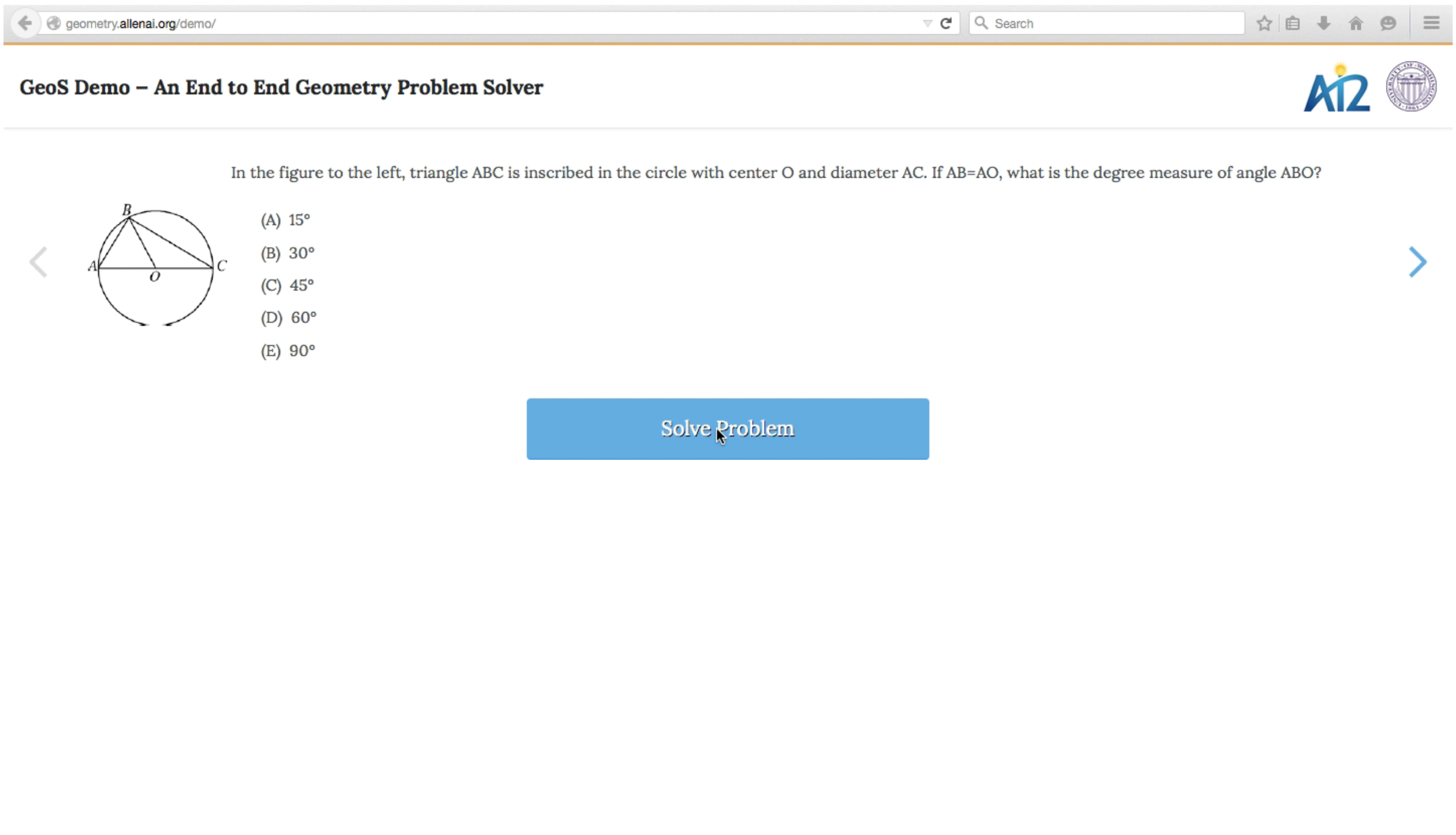
- We manually annotated geometry axioms, alignments and parses
  - We use grade 6, 7 and 8 textbook annotations for development, training, and testing, respectively.

Results

	Textbook	Practice	Official
Avg. Student	44	58	53
Numerical Solver	32	61	49
Axiomatic Solver	51	64	55

Demo

(http://www.cs.cmu.edu/~mrinmays/demo/)



Newtonian Physics Problems

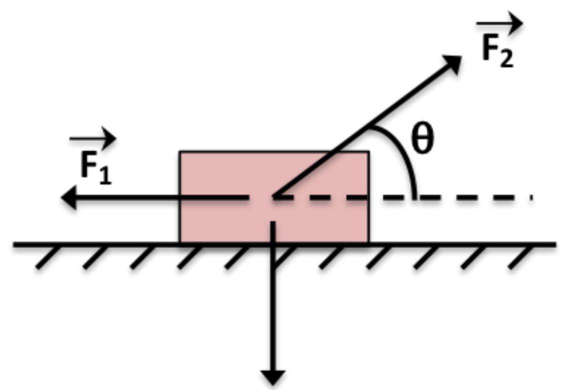
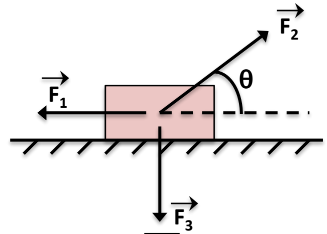
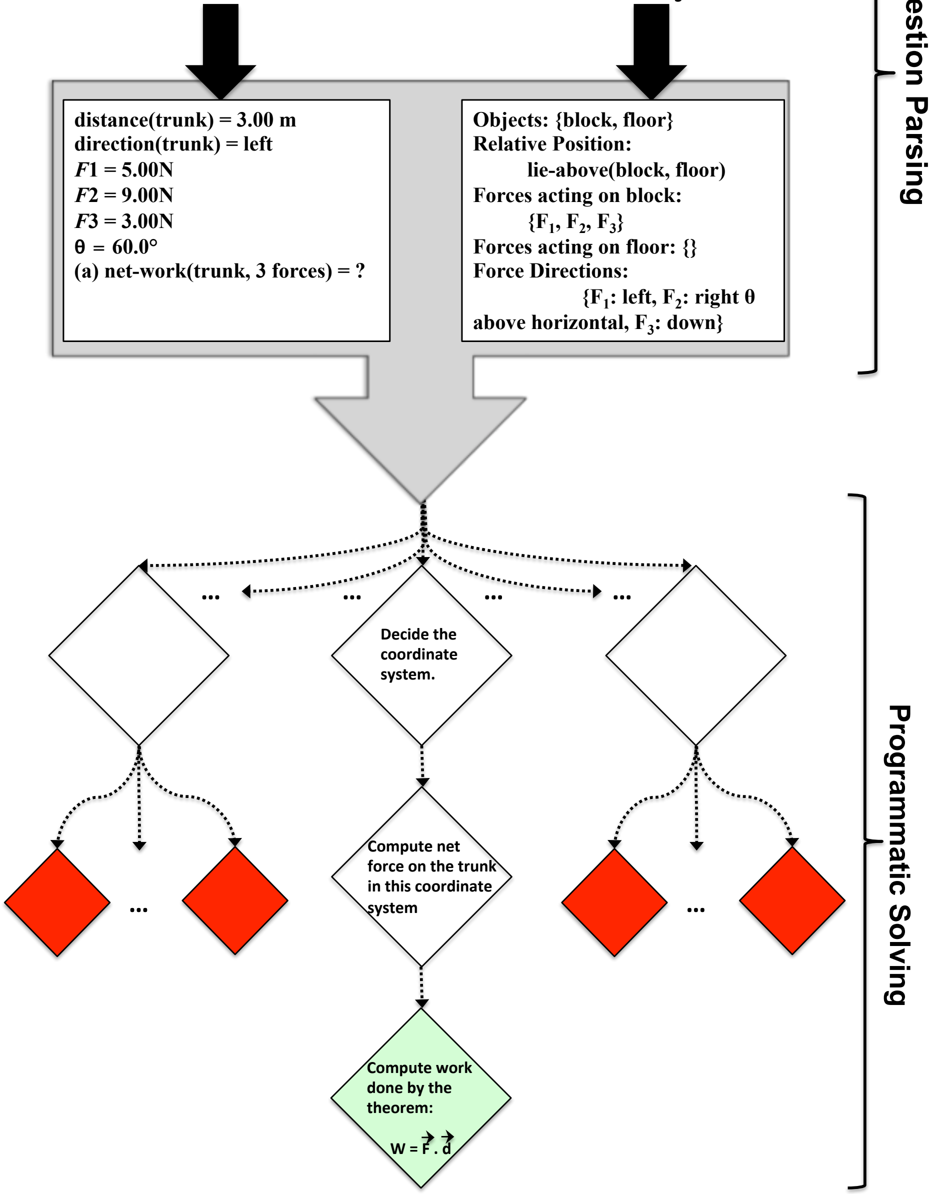


Figure above shows three forces applied to a trunk that moves leftward by 3.00 m over a frictionless floor. The force magnitudes are F_1 = 5.00N, F_2 = 9.00N, and F_3 = 3.00N, and the indicated angle is θ = 60.0° . During the displacement, what is the net work done on the trunk by the three forces?

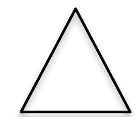
Figure 7-27 shows three forces applied to a trunk that moves leftward by 3.00 m over a frictionless floor. The force magnitudes are $F1$ = 5.00N, $F2$ = 9.00N, and $F3$ = 3.00N, and the indicated angle is θ = 60.0°. During the displacement, what is the net work done on the trunk by the three forces?

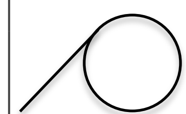


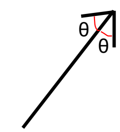


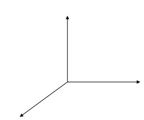
Domain Knowledge as Rules

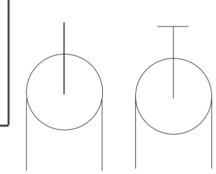
Geometry	**Polygon**: A geometric shape consisting of a number of points and an equal number of line segments, namely a cyclically ordered set of points where no three successive points are collinear and line segments join consecutive pairs of the points. **Circle, Centre**: A set of points that are equidistant from a given point. The point is called the centre. **Tangent, Secant**: Tangent is a line that touches the circle at exactly one point. Secant is a line that intersects the circle in two distinct points.
Physics	**Arrow**: The central (stem) line is the longest of the three lines, the two arrowhead lines are roughly of the same length, and the two angles subtended by the arrowhead lines with the arrow stem line must be roughly equal **Dotted line**: The various lines should be in a straight line, roughly the same sized lines and equi-spaced **Ground**: A solid line is in contact with a number of smaller parallel lines which subtend roughly the same angle with it, their end-point lies on the solid line and the smaller lines are on the same side with respect to the solid line **Coordinate System**: Three arrows where the arrow tails are incident on the same point. Two lines are mutually perpendicular (i.e. angle=90°) and the third roughly bisects the complementary (270°) angle **Block**: Four lines which form a rectangle **Wedge**: Three lines which form a triangle **Pulley**: A circle with two lines tangent to it. An end-point of the two lines lies on the circle





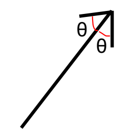






Domain Knowledge as Rules

Arrow: The central (stem) line is the longest of the three lines, the two arrowhead lines are roughly of the same length, and the two angles subtended by the arrowhead lines with the arrow stem line must be roughly equal



$$C_1 = isLine(line1) \land isLine(line2) \land isLine(line3)$$

$$C_2 = length(line1) > length(line2) \land length(line1) > length(line3)$$ *i.e. line1 is stem*

$$C_3 = roughly\_equal(angle(line1, line2), angle(line1, line3))$$

$$C_1 \land C_2 \land C_3 \rightarrow H_{line}$$

Sachan et al. 2018

Domain Theory as Programs

```
def vector_addition(Vectors vectors):
    result = zero_vector()
    for vector in vectors:
        result = result + vector
    return result

def angle_bw_vectors(Vector vec1, Vector vec2):
    return cos_inv(dot(vec1, vec2)/(norm(vec1)*norm(vec2)))

def project_vector(Vector vec, Direction theta):
    return (vec*cos(theta), vec*sin(theta))

def implicit_g_force(Mass m, Forces forces):
    if not forces.contains((-mg i + 0 j)):
        forces.append((mg i + 0 j))

def Newton_II_law(Mass m, Forces forces, Accelerations accs):
    net_force = vector_additon(forces)
    net_acceleration = vector_addition(accs)
    return Constraint(net_force = m * net_acceleration)

def conservation_of_momentum(Mass m1, Velocity v1_initial, Mass m2, Velocity v2_initial, Velocity
v1_final, Velocity v2_final):
    preconditions = [external_force_on_system() == None]
    return Constraint(m1*v1_initial+m2*v2_initial = m1*v1_final+m2*v2_final)
```

Datasets

- Newtonian Physics questions taken from popular pre-university physics textbooks and few AP Physics C: Mechanics courses.
  - Training set: Questions taken from three popular pre-university physics textbooks: *Resnick Halliday Walker*, *D. B. Singh* and *NCERT*.
    - Millions of students in India study physics from these books every year and these books are available online.

- 4941 questions (1019 w/ associated diagrams)
  - 1000 training, 500 development and 3441 test questions.

- We annotated ground truth logical forms for the training and dev question texts and diagrams.
- Evaluated datasets: Section 1 of three AP Physics C Mechanics tests:
  - AP Physics C Mechanics practice test - 10 questions
  - AP Physics C Mechanics official tests (1998) – 75 questions
  - AP Physics C Mechanics official tests (2012) – 35 questions

Results

	Textbook	Practice	1998	2012
Avg. Student	63	52	44	48
P2P	68	50	42	54

Conclusion and Takeaways

- Standardized tests can serve as drivers for AI.
  - They can provide us with interesting challenges that can help us make progress towards the general goals of linguistic and visual understanding and reasoning.
  - Issues with this: adversarial examples, interpretability, …
- (Domain/Background) Knowledge, Common Sense and Reasoning are important
  - Non-Symbolic Methods (e.g. Deep Learning) + Symbolic Methods

References

Standardized Tests as Benchmarks in AI

- R. Brachman et al. (2005). “Selected Grand Challenges in Cognitive Science,” MITRE Technical Report 05-1218
- *Peter Clark and Oren Etzioni,* AI Magazine2016, My Computer is an Honor Student — but how Intelligent is it? Standardized Tests as a Measure of AI
- *Peter Clark, IAAI 2015,* Elementary School Science and Math Tests as a Driver for AI: Take the Aristo Challenge!
- Ernest Davis, The Limitations of Standardized Science Tests as Benchmarks for Artificial Intelligence Research: Position Paper
- K. Barker et al. (2004). “A Question-Answering System for AP Chemistry: Assessing KR&R Technologies,” KR-2004.
- S. Ohlsson, R.H. Sloan, G. Turan, A. Urasky (2013), “Verbal IQ of a Four-Year Old Achieved by an AI System.” Commonsense-2013.
- H. Levesque, E. Davis, L. Morgenstern, (2012). “The Winograd Schema Challenge,” AAAI-12.
- S. Gaudin (2013). “Top Artificial Intelligent system is as smart as a 4-year old”, Computerworld, July 15, 2013.

References

Historical Work on Reading Comprehensions:

- Eugene Charniak. Toward a model of children's story comprehension. PhD thesis, Massachusetts Institute of Technology, 1972.
- Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. Deep read: A reading comprehension system. In ACL, 1999.
- Ellen Riloff and Michael Thelen. A rule-based question answering system for reading comprehension tests. In Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems, 2000.
- Eric Breck, Marc Light, Gideon S Mann, Ellen Riloff, Brianne Brown, Pranav Anand, Mats Rooth, and Michael Thelen. Looking under the hood: Tools for diagnosing your question answering engine. In Proceedings of the workshop on Open-domain question answering, ACL, 2001.
- Jochen L Leidner, Tiphaine Dalmas, Bonnie Webber, Johan Bos, and Claire Grover. Automatic multi-layer corpus annotation for evaluating question answering methods: Cbc4kids. In In Proceedings of the 3rd International Workshop on Linguistically Interpreted Corpora, 2003.
- Eugene Grois and David C Wilkins. Learning strategies for story comprehension: a reinforcement learning approach. In ICML, 2005.
- Sanda M. Harabagiu, Steven J. Maiorano, and Marius A. Pasca. Open domain textual question answering techniques. Natural Language Engineering, 2003.
- Ben Wellner, Lisa Ferro, Warren Greiff, and Lynette Hirschman. Reading comprehension tests for computer-based understanding evaluation. Natural Language Engineering, 2006.

References

MCTest:

- *Matthew Richardson, Christopher J.C. Burges, Erin Renshaw, MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In EMNLP 2013.*
- *Mrinmaya Sachan, Avinava Dubey, Eric P. Xing and Matthew Richardson. Learning Answer-Entailing Structures for Machine Comprehension. In ACL 2015.*
- *Ellery Smith, Nicola Greco, Matko Bosnjak and Andreas Vlachos. A Strong Lexical Matching Method for the Machine Comprehension Test. In EMNLP 2015.*
- *Mrinmaya Sachan, Eric P. Xing. Machine Comprehension using Rich Semantic Representations. In ACL 2016*
- *Hai Wang, Mohit Bansal, Kevin Gimpel, David McAllester. Machine comprehension with syntax, frames, and semantics. In ACL 2015.*
- *Karthik Narasimhan, Regina Barzilay, Machine Comprehension with Discourse Relations, in ACL 2015.*
- *Adam Trischler, Zheng Ye, Xingdi Yuan, Jing He, Philip Bachman, A Parallel-Hierarchical Model for Machine Comprehension on Sparse Data, in ACL 2016.*

References

SQuAD:

- *Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP 2016.*
- *Shuohang Wang and Jing Jiang. Machine Comprehension Using Match-LSTM and Answer Pointer. ICLR 2017.*
- *Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. ICLR 2017.*
- *Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic Coattention Networks For Question Answering. ICLR 2017.*
- *Christopher Clark and Matt Gardner. Simple and Effective Multi-Paragraph Reading Comprehension. 2017.*
- *Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to Answer Open-Domain Questions. ACL 2017.*
- *Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. EMNLP 2017.*
- *Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in Translation: Contextualized Word Vectors. NIPS 2017.*
- *Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. ICLR 2018.*
- *Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. NAACL 2018.*
- *Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. ACL 2018.*
- *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.*

References

Other Reading Comprehension Datasets:

- Yi Yang, Wen-tau Yih and Christopher Meek WIKIQA: A Challenge Dataset for Open-Domain Question Answering
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In NIPS
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev and Percy Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, In EMNLP 2016
- Denis Paperno, German Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda and Raquel Fernandez, The LAMBADA dataset: Word prediction requiring a broad discourse context, In ACL 2016
- Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman and Kaheer Suleman, NEWSQA: A MACHINE COMPREHENSION DATASET, In NEWSQA: A MACHINE COMPREHENSION DATASET 2nd Workshop on Representation Learning for NLP 2016
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder and Li Deng, MS MARCO: A Human Generated MAchine Reading COmprehension Dataset
- Mandar Joshi, Eunsol Choi, Daniel Weld, Luke Zettlemoyer. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, In ACL 2017
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang and Eduard Hovy, RACE: Large-scale ReAding Comprehension Dataset From Examinations, In EMNLP 2017
- Tomas Kovcisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis and Edward Grefenstette, The NarrativeQA Reading Comprehension Challenge, In TACL 2018
- Pranav Rajpurkar, Robin Jia and Percy Liang, Know What You Don't Know: Unanswerable Questions for SQuAD, In ACL 2018

References

Other Reading Comprehension Datasets:

- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay and Dan Roth, NAACL 2018, Looking Beyond the Surface:A Challenge Set for Reading Comprehension over Multiple Sentences
- *Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang and Luke Zettlemoyer, EMNLP 2018,* QuAC: Question Answering in Context
- *Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord,* Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge
- Siva Reddy, Danqi Chen and Christopher D. Manning. CoQA: A Conversational Question Answering Challenge . In EMNLP 2018
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov and Christopher D. Manning. HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In EMNLP 2018

References

Science QA:

- Mrinmaya Sachan, Avinava Dubey, Eric P. Xing. Science Question Answering using Instructional Materials. In ACL 2016
- *Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, Daniel Khashabi.* AAAI 2016, Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions.
- Bhavana Dalvi Mishra Niket Tandon Peter Clark. TACL'17, Domain-Targeted, High Precision Knowledge Extraction
- Bhavana Dalvi, Sumithra Bhakthavatsalam, Chris Clark, Peter Clark, Oren Etzioni, Anthony Fader, Dirk Groeneveld, AKBC'16, IKE - An Interactive Tool for Knowledge Extraction
- Peter Clark, Niranjan Balasubramanian, Sumithra Bhakthavatsalam, Kevin Humphreys, Jesse Kinkead, Ashish Sabharwal, Oyvind Tafjord, AKBC'14, Automatic Construction of Inference-Supporting Knowledge Bases
- Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth, AAAI 2018, Question Answering as Global Reasoning over Semantic Abstractions
- Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth, CONLL 2017, Learning What is Essential in Questions
- Ashish Sabharwal, Peter Clark, Oren Etzioni and Dan Roth, IJCAI 2016, Question Answering via Integer Programming over Semi-Structured Knowledge
- *Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Marco A. Valenzuela-Escárcega, Peter Clark, and Michael Hammond, CONLL 2017,* Tell Me Why: Using Question Answering as Distant Supervision for Answer Justification
- *Sujay Kumar Jauhar, Peter D. Turney, Eduard Hovy, ACL 2016,* Tables as Semi-structured Knowledge for Question Answering
- *Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark, COLING 2016,* What's in an Explanation? Characterizing Knowledge and Inference Requirements for Elementary Science Exams
- *Yang Li and Peter Clark, EMNLP 2015,* Answering Elementary Science Questions by Constructing Coherent Scenes using Background Knowledge

References

Diagram QA:

- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi, ECCV 2016, A Diagram Is Worth A Dozen Images
- Jayant Krishnamurthy, Oyvind Tafjord, and Aniruddha Kembhavi, EMNLP 2016, Semantic Parsing to Probabilistic Programs for Situated Question Answering

Texbook QA:

- *Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Hannaneh Hajishirzi, and Ali Farhadi, CVPR 2017,* Are You Smarter Than A Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension
- *Todor Mihaylov, Peter Clark, Tushar Khot, Ashish Sabharwal, EMNLP 2018,* Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering

References

Math Word Problems:

- *Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang, TACL 2015,* Parsing Algebraic Word Problems into Equations
- Hosseini, M. J., Hajishirzi, H., Etzioni, O., and Kushman, N. (2014). Learning to solve arithmetic word problems with verb categorization. In EMNLP.
- Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., and Hajishirzi, H. (2016). Mawps: A math word problem repository. In NAACL.
- Kushman, N., Zettlemoyer, L., Barzilay, R., and Artzi, Y. (2014). Learning to automatically solve algebra word problems. In ACL, pages 271–281.
- Mitra, A. and Baral, C. (2016). Learning to use formulas to solve simple arithmetic problems. In ACL
- Mukherjee, A. and Garain, U. (2008). A review of methods for automatic understanding of natural language mathematical problems. Artificial Intelligence Review, 29(2):93–122.
- Roy, S. and Roth, D. (2015). Solving general arithmetic word problems. In EMNLP.
- Roy, S. and Roth, D. (2017). Unit dependency graph and its application to arithmetic word problem solving. In AAAI.
- Roy, S., Vieira, T., and Roth, D. (2015). Reasoning about quantities in natural language. In TACL
- Annotating Derivations: A New Evaluation Strategy and Dataset for Algebra Word Problems, Shyam Upadhyay and Ming-Wei Chang, In *EACL*, 2017
- Learning from Explicit and Implicit Supervision Jointly For Algebra Word Problems, Shyam Upadhyay, Ming-Wei Chang, Kai-Wei Chang, and Wen-tau Yih, In *EMNLP*, 2016
- Wang Ling, Dani Yogatama, Chris Dyer, Phil Blunsom. Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems. In ACL 2017
- Yan Wang Xiaojiang Liu Shuming Shi, Deep Neural Solver for Math Word Problems, In EMNLP 2017

References

Geometry Problems:

- ***Minjoon Seo*, Hannaneh Hajishirzi, Ali Farhadi and Oren Etzioni. Diagram understanding in geometry questions. In AAAI 2014**
- ***Minjoon Seo*, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni and Clint Malcolm. Solving geometry problems: combining text and diagram interpretation. In EMNLP 2015**
- ***Mrinmaya Sachan, Avinava Dubey and Eric P. Xing.* From Textbooks to Knowledge: A Case Study in Harvesting Axiomatic Knowledge from Textbooks to Solve Geometry Problems*. In EMNLP 2017***
- ***Mrinmaya Sachan and Eric P. Xing.* Learning to Solve Geometry Problems from Natural Language Demonstrations in Textbooks*. In *SEM 2017***
- **Mrinmaya Sachan, Minjoon Seo, Hannaneh Hajishirzi and Eric P. Xing. Parsing to Programs: A Framework for Situated Question Answering**
- **Mrinmaya Sachan, Eduard Hovy and Eric P. Xing. Discourse in Multimedia: A Case Study in Information Extraction.**
- T. Davis. Geometry with computers. *2006*
- E. A. Feigenbaum and J. Feldman. Computers and thought. The AAAI Press, 1963
- D. Schattschneider and J. King. Geometry Turned On: Dynamic Software in Learning, Teaching, and Research. Mathematical Association of America Notes, 1997
- W. Wen-Tsun. Basic principles of mechanical theorem proving in elementary geometries. Journal of automated Reasoning, 1986
- D. Kapur. Using gröbner bases to reason about geometry problems. Journal of Symbolic Computation, 1986
- S.-C. Chou, X.-S. Gao, and J.-Z. Zhang. Machine proofs in geometry: Automated production of readable proofs for geometry theorems, World Scientific, 1994.
- X.-S. Gao and Q. Lin. Mmp/geometer–a software package for automated geometric reasoning. In International Workshop on Automated Deduction in Geometry, 2002.
- S. Wilson and J. D. Fleuriot. Combining dynamic geometry, automated geometry theorem proving and diagrammatic proofs. In Workshop on User Interfaces for Theorem Proving (UITP), 2005
- C. Alvin, S. Gulwani, R. Majumdar, and S. Mukhopadhyay. Synthesis of geometry proof problems. In AAAI 2014
- S. Gulwani, V. A. Korthikanti, and A. Tiwari. Synthesizing geometry constructions. In ACM SIGPLAN Notices, 2011.
- S. Itzhaky, S. Gulwani, N. Immerman, and M. Sagiv. Solving geometry problems using a combination of symbolic and numerical reasoning. In International Conference on Logic for Programming Artificial Intelligence and Reasoning, 2013

References

Physics Problems:

- ***Mrinmaya Sachan, Eric P. Xing.* Parsing to Programs: A Framework for Situated QA*. In KDD 2018***
- **Mrinmaya Sachan, Minjoon Seo, Hannaneh Hajishirzi and Eric P. Xing. Parsing to Programs: A Framework for Situated Question Answering**
- M. D. Chang, J. W. Wetzel, and K. D. Forbus. Spatial reasoning in comparative analyses of physics diagrams. In International Conference on Spatial Cognition, 2014
- M. Klenk and K. Forbus. Cognitive modeling of analogy events in physics problem solving from examples. Technical report, NORTHWESTERN UNIV, 2007
- M. Klenk and K. Forbus. Measuring the level of transfer learning by an ap physics problem-solver. In AAAI/IAAI 2007
- M. Klenk and K. Forbus. Analogical model formulation for transfer learning in AP physics. Artificial intelligence, 173(18):1615–1638, 2009. 2
- M. Klenk and K. Forbus. Exploiting persistent mappings in cross-domain analogical learning of physical domains. Artificial intelligence, 2013
- M. Klenk, K. D. Forbus, E. Tomai, H. Kim, and B. Kyckelhahn. Solving everyday physical reasoning problems by analogy using sketches. In Proceedings of National Conference on Artificial Intelligence, 2005.
- Bundy, A., Byrd, L., Luger, G., Mellish, C., Milne, R., and Palmer, M. (1979). MECHO: A program to solve mechanics problems. Department of Artificial Intelligence, University of Edinburgh.
- Novak, G. S. (1976). Computer understanding of physics problems stated in natural language. PhD thesis, The University of Texas at Austin.
- Oberem, G. (1987). Albert: a physics problem solving monitor and coach. In Proceedings of the first international conference on computer assisted learning ICCAL 1987.