From Simple to Complex QA

Eduard Hovy

CMU Language Technologies Institute
www.cs.cmu.edu/~hovy

Webclopedia
(Hovy et al. 2001)

Webclopedia QA, 2003

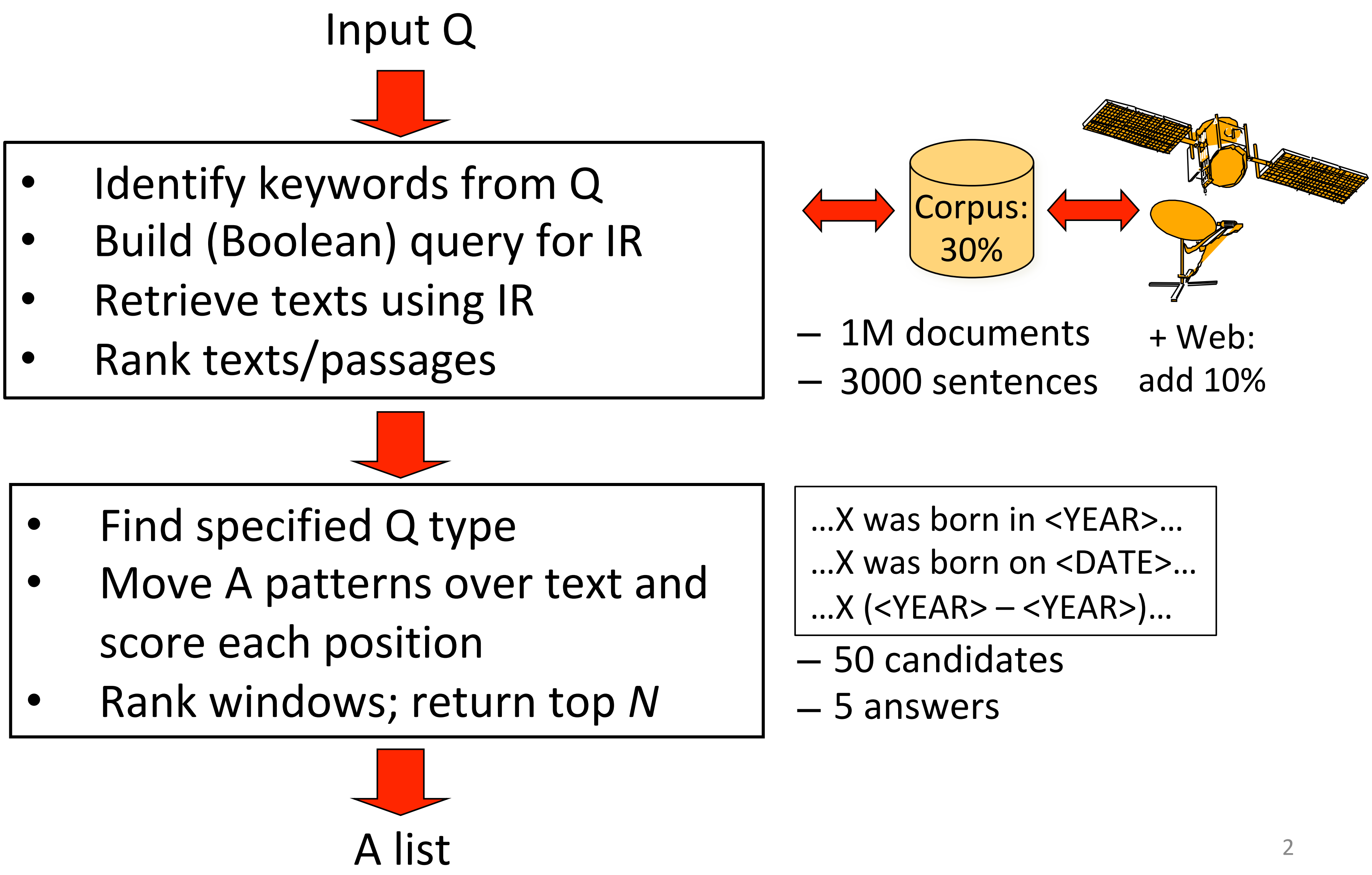
- Where do lobsters like to live?
  — *on the table*

- Where are zebras most likely found?
  — *in the dictionary*

- How many people live in Chile?
  — *nine*

- What is an invertebrate?
  — *Dukakis*

Basic simple factoid QA



Input Q

- Identify keywords from Q
- Build (Boolean) query for IR
- Retrieve texts using IR
- Rank texts/passages

Corpus: 30%

– 1M documents
– 3000 sentences

+ Web: add 10%

- Find specified Q type
- Move A patterns over text and score each position
- Rank windows; return top *N*

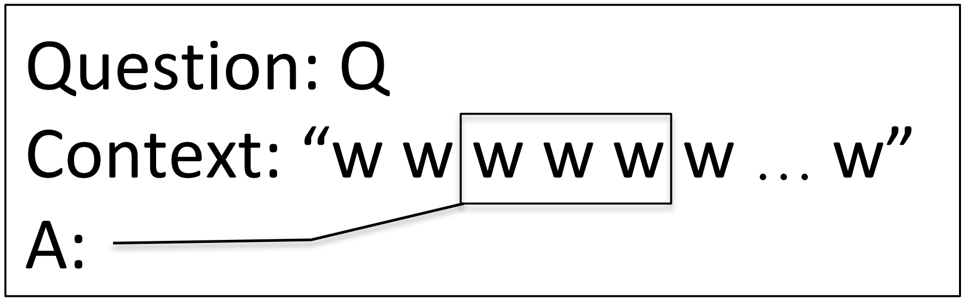
...X was born in <YEAR>...
...X was born on <DATE>...
...X (<YEAR> – <YEAR>)...

– 50 candidates
– 5 answers

A list

Where is the Answer?
— Progress since 2003?

Typical QA format:



Question: Q

Context: “w w w w w w … w”

A:

Either the Q context provides the A

1. nearby (= n-word window) context
2. distant (= doc-level) context

Or not at all…so you have to use background info

3. from the training data
4. from logical derivation/reasoning rules/procedure

Either…

When all info needed to get the A is present in the Q context

…then some form of surface and simple type matching + sub-A composition is enough

—> Ultimately, just do [nested] simple QA

Or…

But when getting the A requires information **not** in the Q context (like background info, calculation, etc.)

…then you are in trouble: this is not standardized, hence impossible to evaluate

—> No complex QA !?

Outline

1. A in the nearby context
2. A in the distant context
3. A hidden in the training data
4. A only by reasoning

Option 1: A in nearby context

- Build and use short patterns or a rich LM
- Tons of work since 2000 on pattern learning and generalization, QA typologies, etc.
- Numerous QA datasets (TREC, SQuAD, CNN…)
- Many QA competitions (SEMEVAL…)

So where’s the limit?

You can do a LOT with patterns
Did you know you are an expert on the Panama Canal?

Blah Panama Canal blah blah Panama blah Pres. Roosevelt blah USA blah blah blah blah blah 10 years blah until 1914 blah blah blah blah 51 miles blah blah blah blah blah blah blah blah blah blah blah blah blah 8 to 10 hours blah blah blah blah Gatun Lake blah

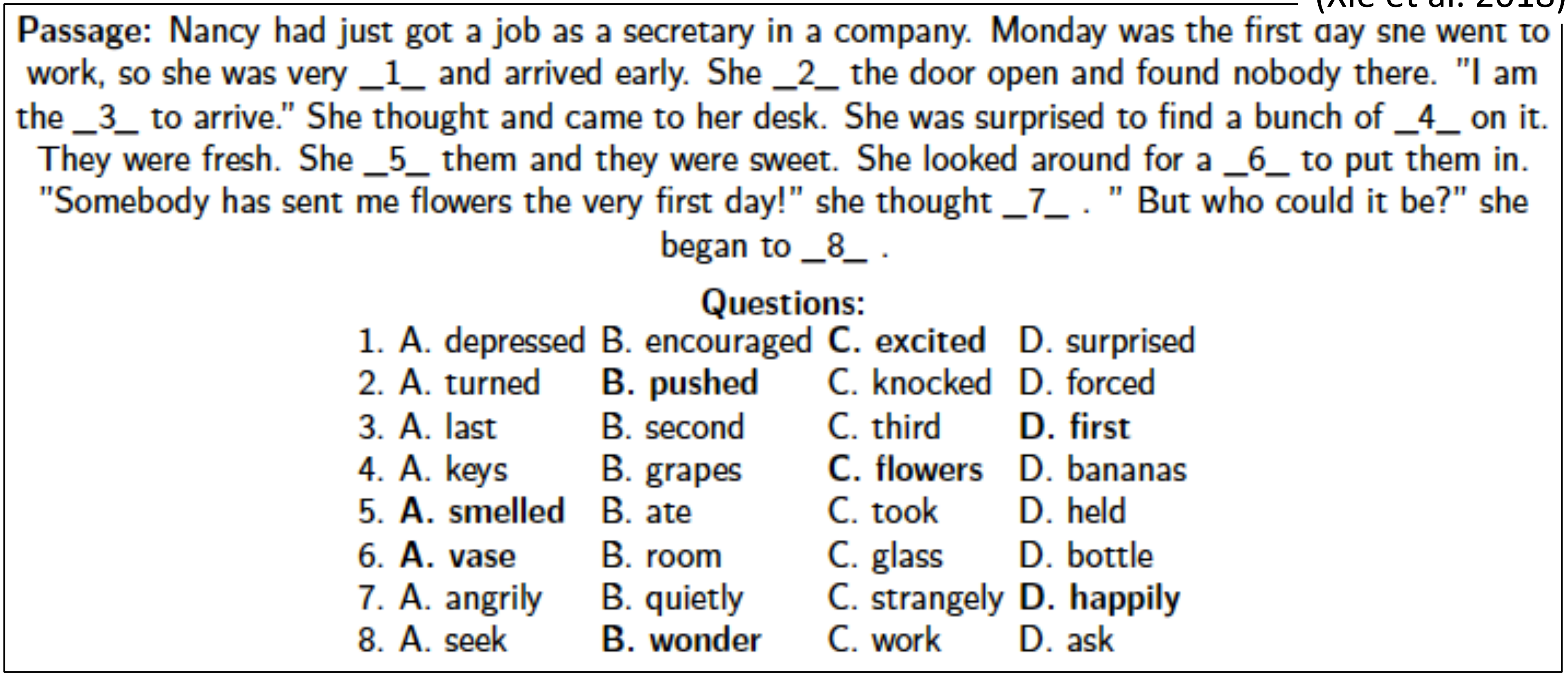
Which oceans does the Panama Canal connect?

So where's the limit?

A corpus to test the power of ngram/pattern QA models

- **CLOTH** (Xie, Lai, Dai, Hovy, EMNLP 2018)
  - Large-scale Cloze test dataset
  - Created by English teachers in China for English exams (Middle and High school levels)
  - After cleanup: 7k passages; 99k questions (2/3 removed)
  - Dropped words and word options carefully created by teachers: highly nuanced alternatives
  - Tests knowledge of grammar, vocabulary, reasoning
- How well do state-of-the-art computational models do compared to humans?
  - We test using a 1-billion-word language model

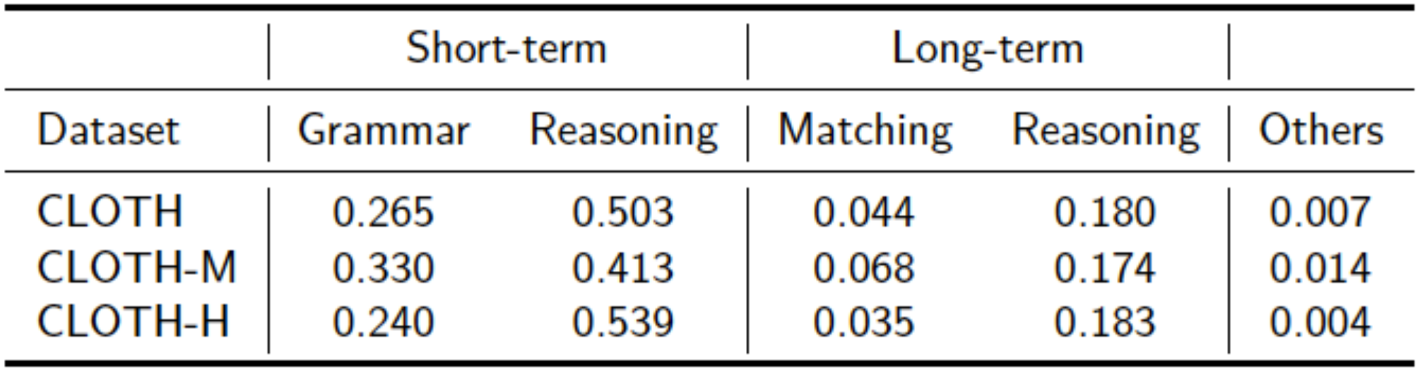
(Xie et al. 2018)



Passage: Nancy had just got a job as a secretary in a company. Monday was the first day she went to work, so she was very _1_ and arrived early. She _2_ the door open and found nobody there. "I am the _3_ to arrive." She thought and came to her desk. She was surprised to find a bunch of _4_ on it. They were fresh. She _5_ them and they were sweet. She looked around for a _6_ to put them in. "Somebody has sent me flowers the very first day!" she thought _7_ . " But who could it be?" she began to _8_ .

Questions:

1. A. depressed B. encouraged **C. excited** D. surprised
2. A. turned **B. pushed** C. knocked D. forced
3. A. last B. second C. third **D. first**
4. A. keys B. grapes **C. flowers** D. bananas
5. **A. smelled** B. ate C. took D. held
6. **A. vase** B. room C. glass D. bottle
7. A. angrily B. quietly C. strangely **D. happily**
8. A. seek **B. wonder** C. work D. ask



	Short-term		Long-term		
Dataset	Grammar	Reasoning	Matching	Reasoning	Others
CLOTH	0.265	0.503	0.044	0.180	0.007
CLOTH-M	0.330	0.413	0.068	0.174	0.014
CLOTH-H	0.240	0.539	0.035	0.183	0.004

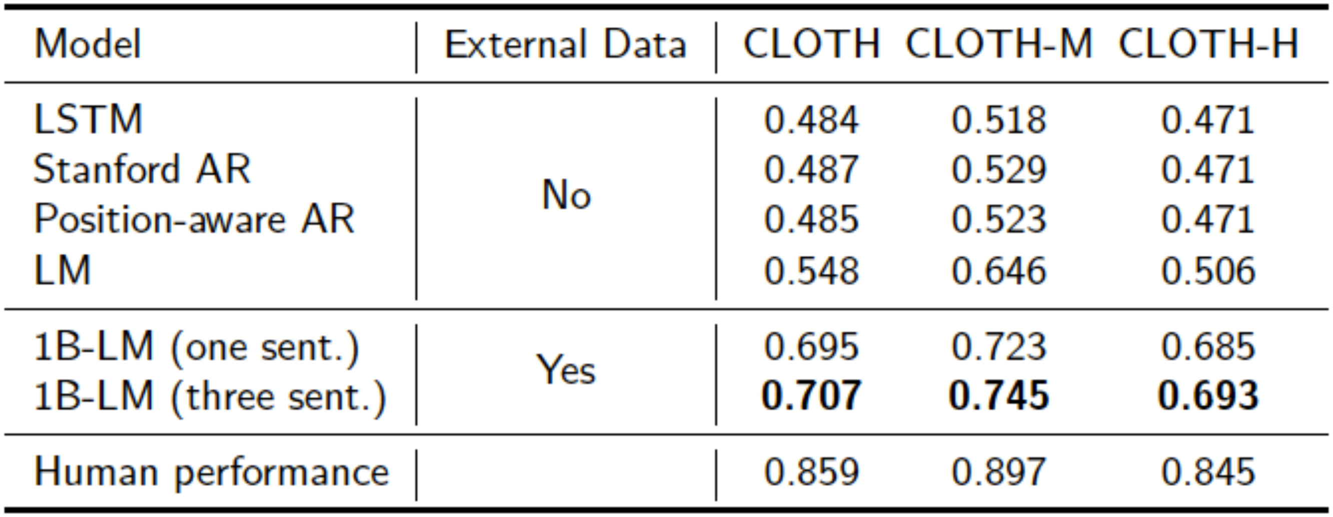
Percentages of test examples, Middle/High school levels

- Tense, voice, preps
- Local content words

- Copy/paraphrase words
- Content words, long-distance dependencies

(Xie et al. 2018)

QA system results

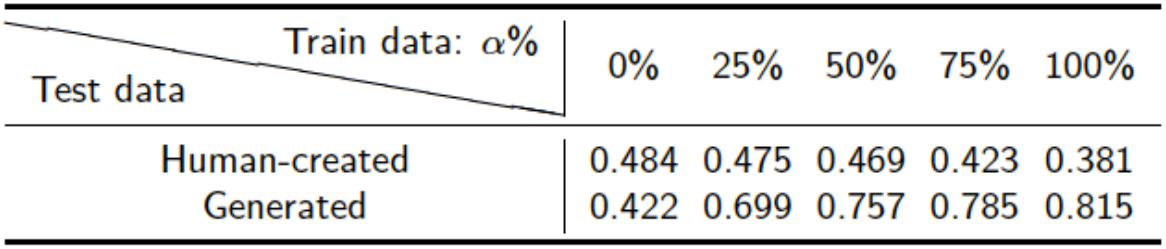


Model	External Data	CLOTH	CLOTH-M	CLOTH-H
LSTM	No	0.484	0.518	0.471
Stanford AR		0.487	0.529	0.471
Position-aware AR		0.485	0.523	0.471
LM		0.548	0.646	0.506
1B-LM (one sent.)	Yes	0.695	0.723	0.685
1B-LM (three sent.)		**0.707**	**0.745**	**0.693**
Human performance		0.859	0.897	0.845

This was pre-BERT!)

(AR: Attention Reader)

- Even a 1B-LM still lags behind human performance
- Increasing the context length for 1B-LM does not help
- However: human-created questions are different:



Train data: α% Test data	0%	25%	50%	75%	100%
Human-created	0.484	0.475	0.469	0.423	0.381
Generated	0.422	0.699	0.757	0.785	0.815

Conclusion for option 1

For factoid QA types that obey patterns,

if the A is close enough, and you have enough training data…

…you will always learn good enough word combination patterns to connect

Q parameters <–> Q context material <–> A

(If you haven’t seen the necessary word combinations, you won’t ever be able to answer the Q)

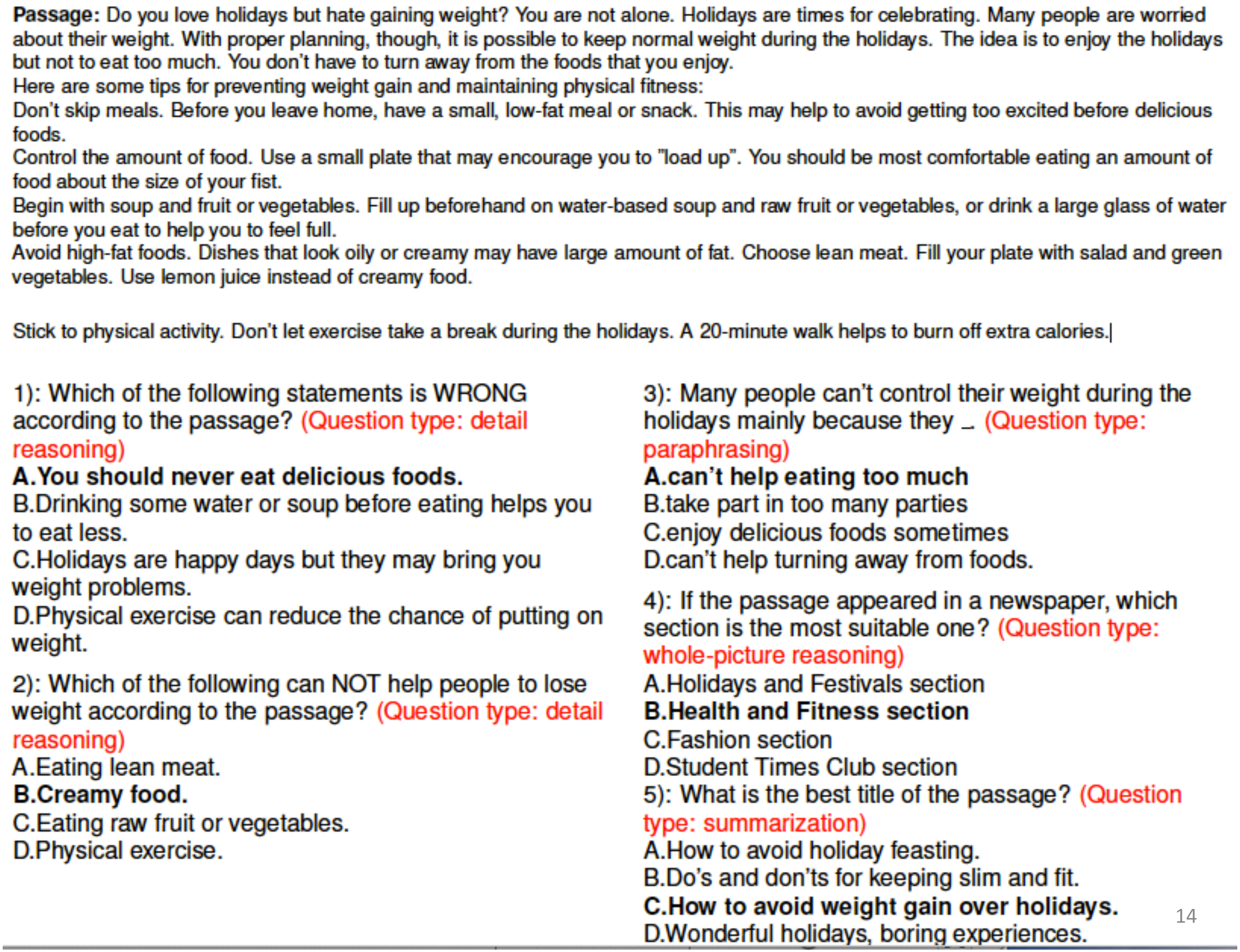
Option 2: A in distant context

- Still use some form of matching Q and A
- Need a more-sophisticated and longer-distance type of ‘pattern’

(Lai et al. 2018)

Making matching more complex: RACE: A better testbed

- **RACE**: ReAding Comprehension dataset from Examinations (Lai, Xie, Liu, Yang, Hovy, EMNLP 2018)
- Collected from Chinese middle and high school exams that evaluate human students' English reading comprehension ability
  - Designed by human experts: Ensures quality and broad topic coverage
  - Substantially more difficult than existing QA datasets (but RACE-M easier than RACE-H)
  - About 4/5 of source material filtered out to remove duplicates, incorrect format, etc.
    - After cleaning: 27,933 passages; 97,687 questions



Passage: Do you love holidays but hate gaining weight? You are not alone. Holidays are times for celebrating. Many people are worried about their weight. With proper planning, though, it is possible to keep normal weight during the holidays. The idea is to enjoy the holidays but not to eat too much. You don't have to turn away from the foods that you enjoy.

Here are some tips for preventing weight gain and maintaining physical fitness:

Don't skip meals. Before you leave home, have a small, low-fat meal or snack. This may help to avoid getting too excited before delicious foods.

Control the amount of food. Use a small plate that may encourage you to "load up". You should be most comfortable eating an amount of food about the size of your fist.

Begin with soup and fruit or vegetables. Fill up beforehand on water-based soup and raw fruit or vegetables, or drink a large glass of water before you eat to help you to feel full.

Avoid high-fat foods. Dishes that look oily or creamy may have large amount of fat. Choose lean meat. Fill your plate with salad and green vegetables. Use lemon juice instead of creamy food.

Stick to physical activity. Don't let exercise take a break during the holidays. A 20-minute walk helps to burn off extra calories.|

1): Which of the following statements is WRONG according to the passage? (Question type: detail reasoning)

A.You should never eat delicious foods.

B.Drinking some water or soup before eating helps you to eat less.

C.Holidays are happy days but they may bring you weight problems.

D.Physical exercise can reduce the chance of putting on weight.

2): Which of the following can NOT help people to lose weight according to the passage? (Question type: detail reasoning)

A.Eating lean meat.

B.Creamy food.

C.Eating raw fruit or vegetables.

D.Physical exercise.

3): Many people can't control their weight during the holidays mainly because they _ (Question type: paraphrasing)

A.can't help eating too much

B.take part in too many parties

C.enjoy delicious foods sometimes

D.can't help turning away from foods.

4): If the passage appeared in a newspaper, which section is the most suitable one? (Question type: whole-picture reasoning)

A.Holidays and Festivals section

B.Health and Fitness section

C.Fashion section

D.Student Times Club section

5): What is the best title of the passage? (Question type: summarization)

A.How to avoid holiday feasting.

B.Do's and don'ts for keeping slim and fit.

C.How to avoid weight gain over holidays.

D.Wonderful holidays, boring experiences.

(Lai et al. 2018)

Toward ‘reasoning’: types of more-complex matching

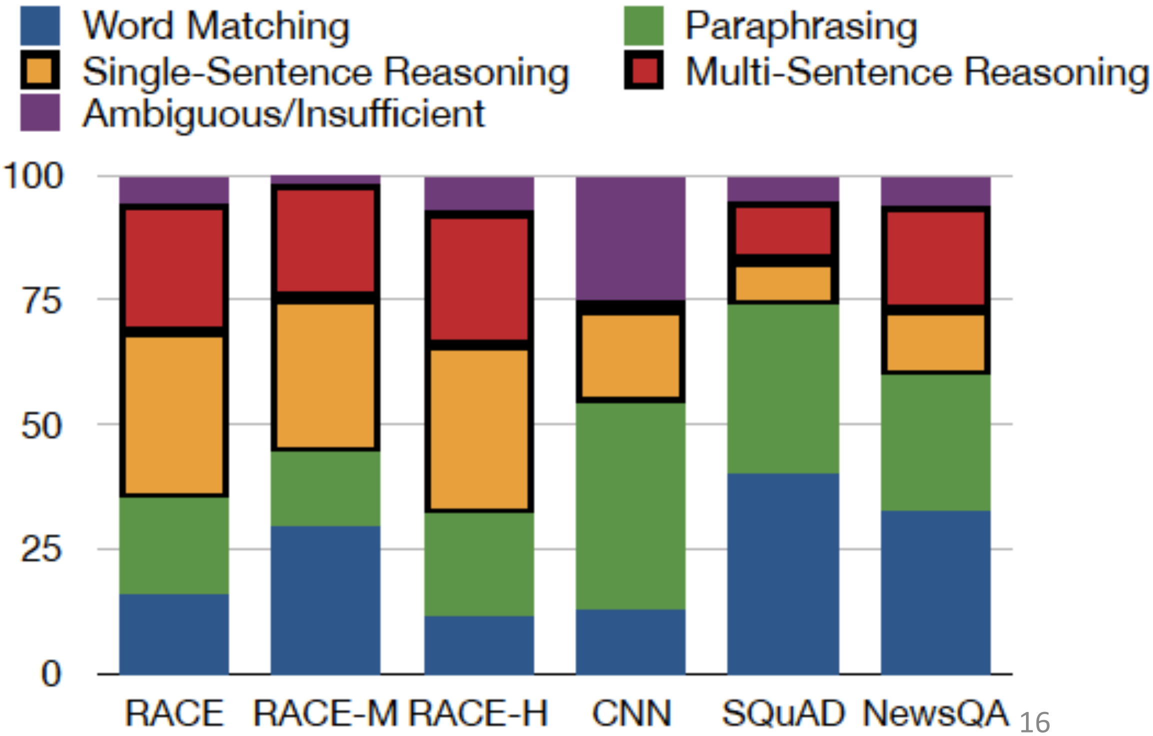
- **Paraphrasing** Qs: test language ability
- **Detail** Qs: identify and match details of a thing
- **Attitude** Qs: find opinions/attitudes of the author towards something (’sentiment’)
- **Whole-picture** Qs: understand the entire story (multi-sentence)
- **Summarization** Qs: understand the point (multi-sentence)

Increasing reasoning

(Lai et al. 2018)

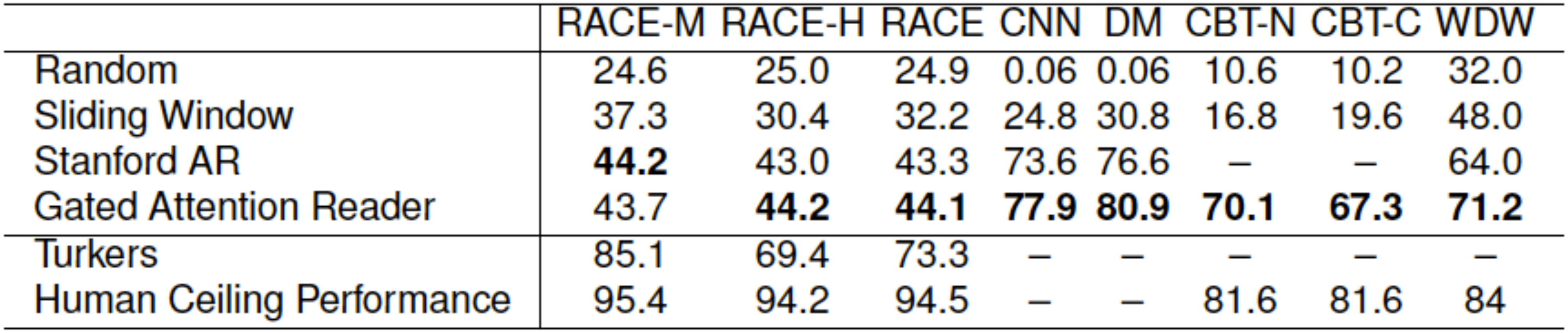
Comparison with other QA datasets

- Reasoning questions: 59.2% of RACE; 20.5% of SQuAD
- Processing types:
  - Word matching: exact match
  - Paraphrasing: paraphrase or entailment
  - Single-sent reasoning: incomplete info or conceptual overlap
  - Multi-sent reasoning: synthesizing information from multiple sentences
  - Insufficient/ Ambiguous: no A, or A is not unique



(Lai et al. 2018)

Comparing QA algorithms

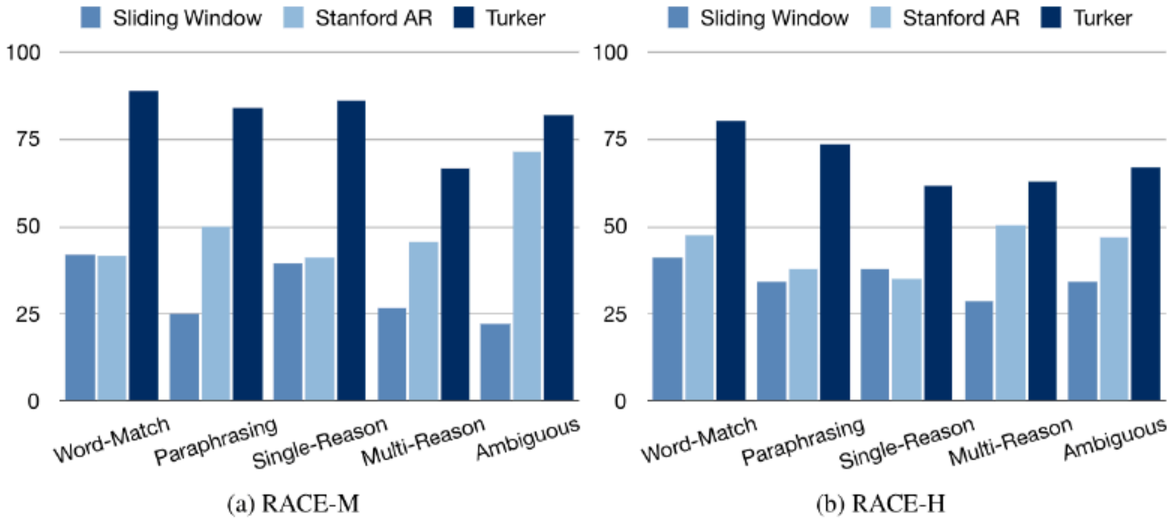


	RACE-M	RACE-H	RACE	CNN	DM	CBT-N	CBT-C	WDW
Random	24.6	25.0	24.9	0.06	0.06	10.6	10.2	32.0
Sliding Window	37.3	30.4	32.2	24.8	30.8	16.8	19.6	48.0
Stanford AR	**44.2**	43.0	43.3	73.6	76.6	–	–	64.0
Gated Attention Reader	43.7	**44.2**	**44.1**	**77.9**	**80.9**	**70.1**	**67.3**	**71.2**
Turkers	85.1	69.4	73.3	–	–	–	–	–
Human Ceiling Performance	95.4	94.2	94.5	–	–	81.6	81.6	84

- Baselines:
  - Sliding Window: TF-IDF based matching algorithm
  - Stanford Attention Reader (AR) and Gated Attention Reader (early-2018 state-of-the-art neural models)
- RACE has more ‘semantics’ (= requires more ‘reasoning’) than other corpora:
  - higher human ceiling
  - harder for neural models

(Lai et al. 2018)

Matching type performance



- Turkers and Sliding Window are good at simple matching questions
- Surprisingly, Stanford AR does not have better performance on matching questions

Conclusion for option 2

When the A is distant, or requires more-sophisticated matching/’reasoning’ (not just simple word-string / language model),

then attention-based neural models can do some of it, but still fail with the harder parts

Option 3: A ‘hidden’ in training data

Sometimes the Q context does not contain the A at all

…but you can STILL get the right A!

(And even get it without the Q itself!)

Corrupted ngrams and other SQuAD perturbations (Jia and Liang, EMNLP 2017)

Necessity of Q context or even of Q itself (Kaushik and Lipton, EMNLP 2018, Best Short Paper award)

(Kaushik and Lipton, EMNLP 2018)

Example: Q only

Question: shin kanemaru , the gravel-voiced back-room boss who died on thursday aged 81 , goes down in history as japan’s most corrupt post-war politician after ____________

Passage: ... glynis bc-nj-zimmer-profile-2takes-nyt rahane **fumio yasuhiro** dragnea lhadon bjorkman/max ... seventh-largest embarrased jeopardy hilariously **masahisa haibara** bajram 8-to-24 duke/meredith acceding ... koidu iraqs 2:32:21 //www.ironmanlive.com/ **sagawa kyubin** dean internatinoal 90-meter **kakuei tanaka** seven-paragraph 577,610 wendover golf-lpga-jpn partner, un-appointed ue mazzei canada-u.s.

Answer: kakuei tanaka

(Kaushik and Lipton EMNLP 2018)

Do you actually need the context?

- Research goal:
  - How strong are models that see the **Q only**?
  - What about models that see the **Q context passage only**?
  - How do we know models are really “reading” the **whole passage**?

- **Question-only** setting:
  - If the QA system needs the passage, randomize its words first
  - If just candidate As needed, place them in random spots, fill intervening text with gibberish
- **Passage-only** setting:
  - ‘Ignore’ the Qs: assign each Q to some random passage

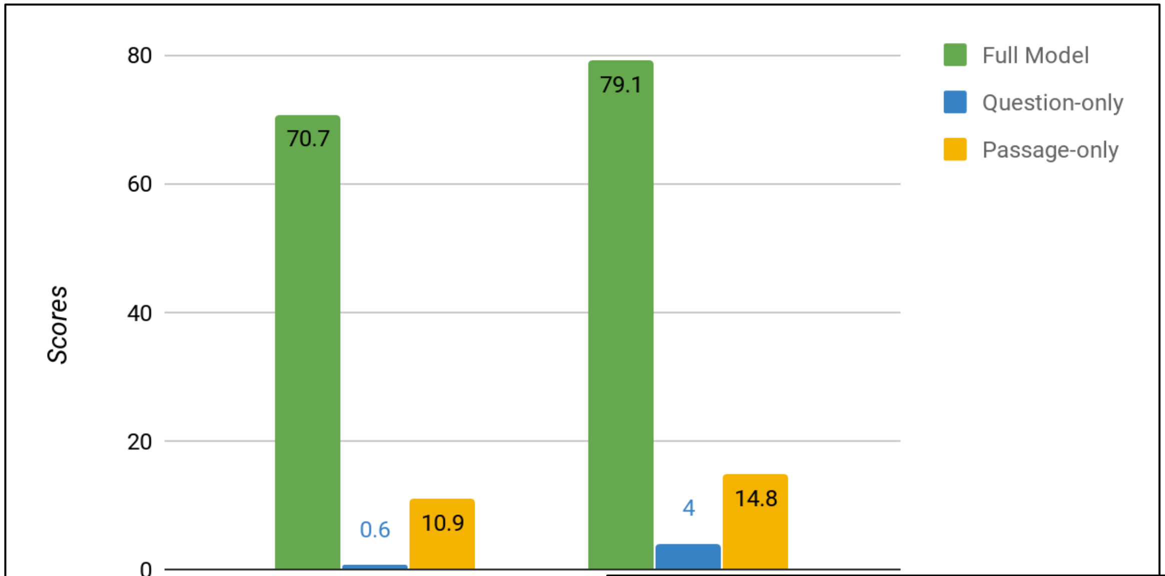
(Kaushik and Lipton EMNLP 2018)

Experiments

- Datasets / tests:
  - Span selection: SQuAD, TriviaQA
  - Cloze queries: Childrens Book Test (CBT), CNN, CLOTH, Who-did-What, DailyMail
  - Multi-class classification (implicit): bAbI (20 tasks)
  - Multiple-choice question answering: RACE, MCTest
  - Answer generation: MS MARCO
- Algorithms:
  - **Key-Value Memory Networks**:
    Miller et al. 2016: Key-Value Memory Networks for Directly Reading Documents. *Proceedings of EMNLP*
  - **Gated Attention Readers**:
    Dhingra et al. 2017: Gated-Attention Readers for Text Comprehension. *Proceedings of ACL*
  - **QANet**:
    Yu et al. 2018: QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *Proceedings of ICLR*

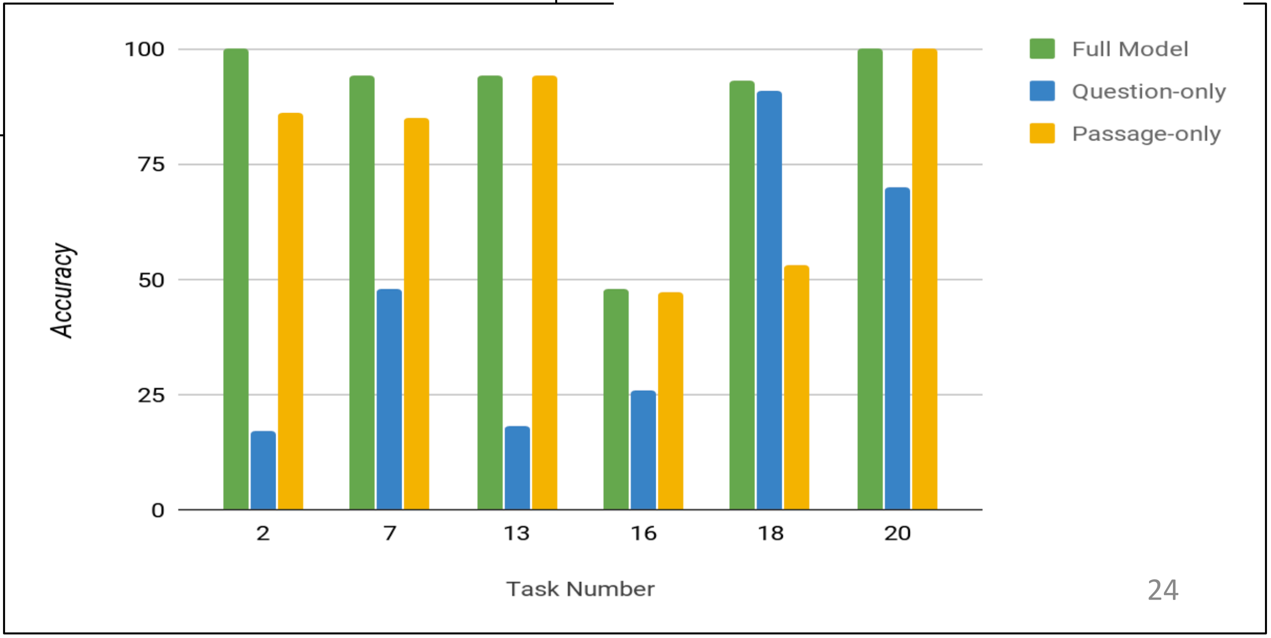
(Kaushik and Lipton EMNLP 2018)

Some results

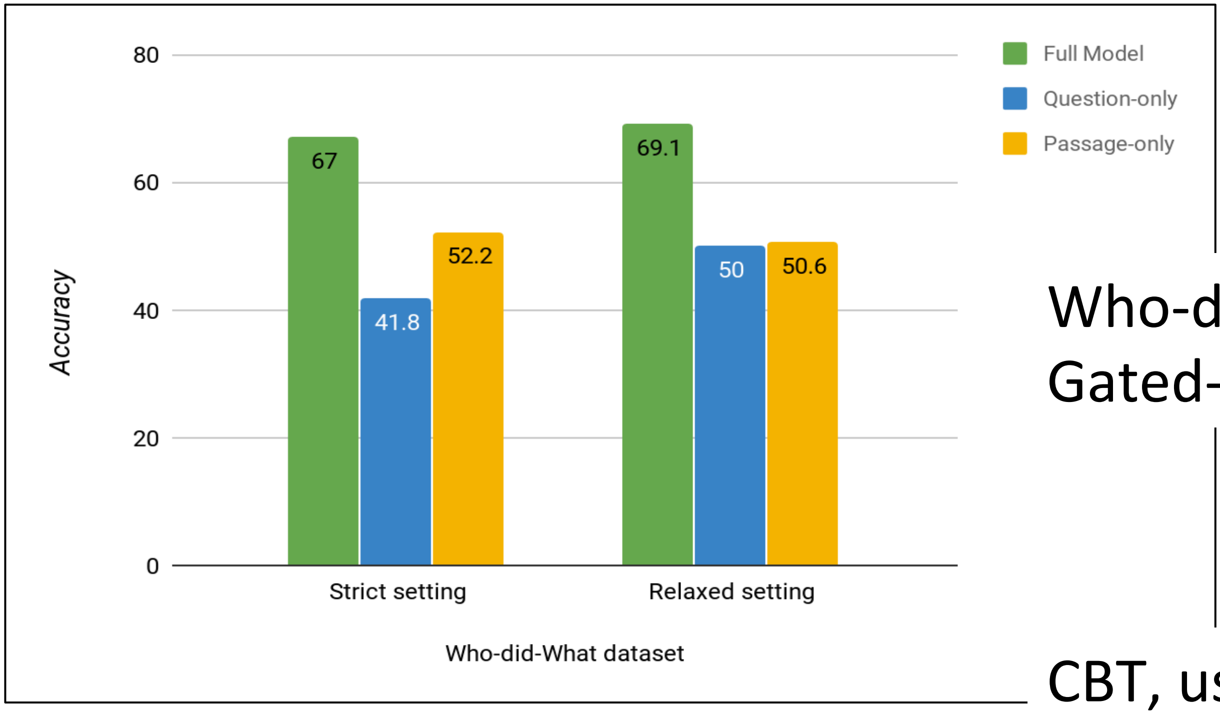


SQuAD, using QANet

bAbI, using Key-Value MemNets

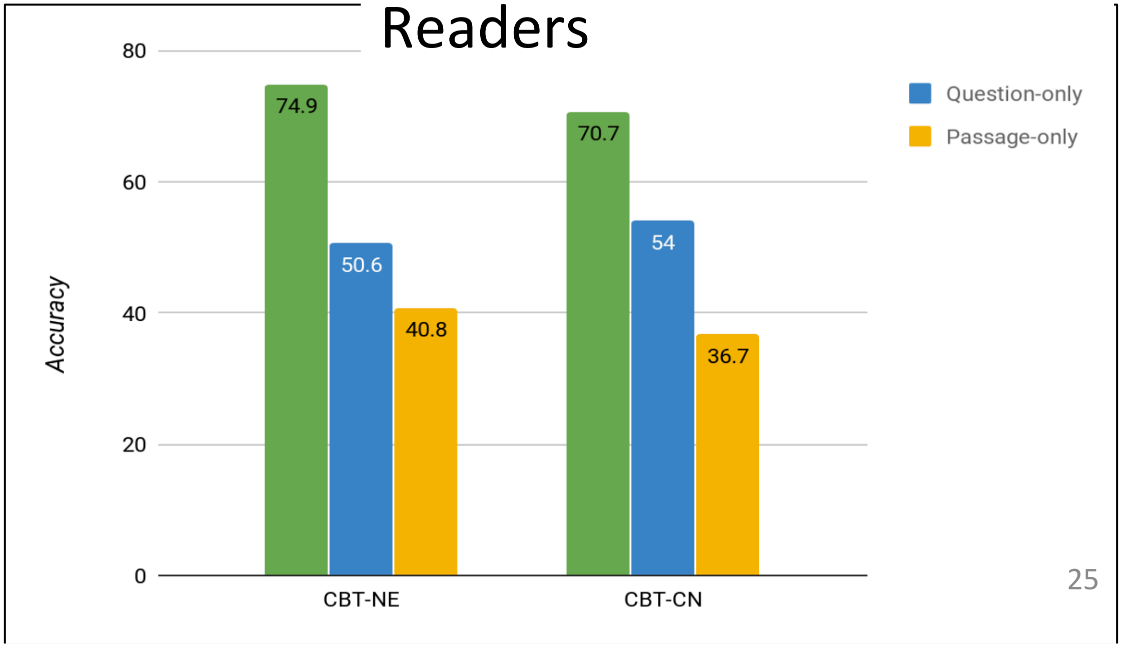


(Kaushik and Lipton EMNLP 2018)



Who-did-What, using Gated-Attention Readers

CBT, using Gated-Attention Readers



Why? What's going on??

Question: shin kanemaru , the gravel-voiced back-room boss who died on thursday aged 81 , goes down in history as japan's most corrupt post-war politician after ___________

Passage: ... glynis bc-nj-zimmer-profile-2takes-nyt ra **fumio yasuhiro** dragnea lhadon bjorkman/max ... seventh largest embarrased jeopardy hilariously **masahisa haibara** bajram 3-to-24 duke/meredith acceding ... koidu iraqs //www.ironmanlive.com/ **sagawa kyubin** dea oal 90-meter **kakuei tanaka** seven-paragrap vendover golf-lpga-jpn partner, un mazzei canada-u.s.

Kanemaru's secretary

Name not in Google

Transportation company

Long-term politician

Answer: kakuei tanaka

Conclusion for option 3

- Don’t trust QA datasets!
- Don’t trust QA system claims!
- First, check if
  - any pre-existing (= training data) dependencies among the Q and candidate As?
  - full context predicts the A without even the Q?

Option 4: A only through reasoning

For truly complex QA:

1. Identify the individual steps/pieces needed to derive the A
2. Figure out how to compute/find them
   – From the Q context and/or from elsewhere
3. Compose (and check?) them
   – Build an A finding ‘script’

Possible sources of this knowledge

- External search:
  - Query something like the web and hope to be lucky

- Entailments: “sentence” –> “sentence”
  - Operate at surface form (in RTE formulation)
  - Allow one **surface form** to be stated when another is given
  - New surface form may provide Answer
  - Need: entailment rules + entailment applier

- Axioms: $A \vee B$ –> C
  - Operate at deeper level
  - Connect **representation subgraphs**, even providing new nodes
  - Expanded graph may provide Answer
  - Need: axioms / composition rules + theorem prover

Type 1: A popular task today: QA over structured data

- **Data**: database, table, etc.
- **Task**: ask Qs that require (1) finding various bits of data and (2) composing them to make the A
- The missing information is the script governing the sequence of access and composition
- **Research**: how to [learn to] build this script?
- **Evaluation**: did the system produce the right A?
- Examples:
  - U.S. geography database of 800 facts (Zelle & Mooney, 1996)
  - Wikitable questions (Pasupat and Liang, 2015; Dasigi 2018)
  - Other domains' tables (several AI2 projects)

(Dasigi LTI PhD thesis, 2018)

Wikitable dataset

Athlete	Nation	Olympics	Medals
Gillis Grafström	Sweden (SWE)	1920–1932	4
Kim Soo-Nyung	South Korea (KOR)	1988-200	6
Evgeni Plushenko	Russia (RUS)	2002–2014	4
Kim Yu-na	South Korea (KOR)	2010–2014	2
Patrick Chan	Canada (CAN)	2014	2

WikiTableQuestions, Pasupat and Liang, 2015

Question: Which athlete was from South Korea after the year 2010?

Answer: Kim Yu-Na

Reasoning:

1) Get rows where *Nation* column contains *South Korea*
2) Filter rows where *Olympics* has a value greater than *2010*.
3) Get value from *Athlete* column from filtered rows.

Program:

```
((reverse athlete) (and
      (nation south_korea)
      (year ((reverse date)
                  (>= 2010-mm-dd)))
```

(Dasigi LTI PhD thesis, 2018)

Example: Dasigi

- Approach for learning to build access routines:
  1. Parse Q, build dependency tree
  2. Convert into Logical Form
  3. Translate into candidate table access routine
  4. (try all kinds of mappings from words to query operators/structure)
  5. Test composition by repeated trial and error
- Essentially, learning is a search in ‘operator combination space’ to build the logical form
- Weak supervision is not enough. Speed up the learning/search by:
  - Learning to associate **table access parameters** with parts of the tree (Q variables)
  - Learning to associate **nesting and access operators** with parts of the tree (‘operator’ words: “the most”, “last”, etc.)
  - Predefining some lexicon-to-operation mappings
  - Paying attention to grammatical construction of the tree
  - Implementing heuristics to guide exploration (‘short Qs first’)

Dasigi approach

- Strategies:
  - Incorporate knowledge of grammatical constraints
  - ‘Lucky’ examples: remove right A with wrong query logic
  - Question coverage: how many Q words mapped?
  - Complex queries (denotation): how large is the query?
  - Do iterative search, from simpler to more complex Qs

- Combine into single Objective: Minimize expected value of cost (Goodman, 1996; Goel and Byrne, 2000; Smith and Eisner, 2005)

$$\min_{\theta} \sum_{I=1}^{N} \mathbb{E}_{p(y_i|x_i;\theta)} \mathcal{C}(x_i, y_i, w_i, d_i)$$

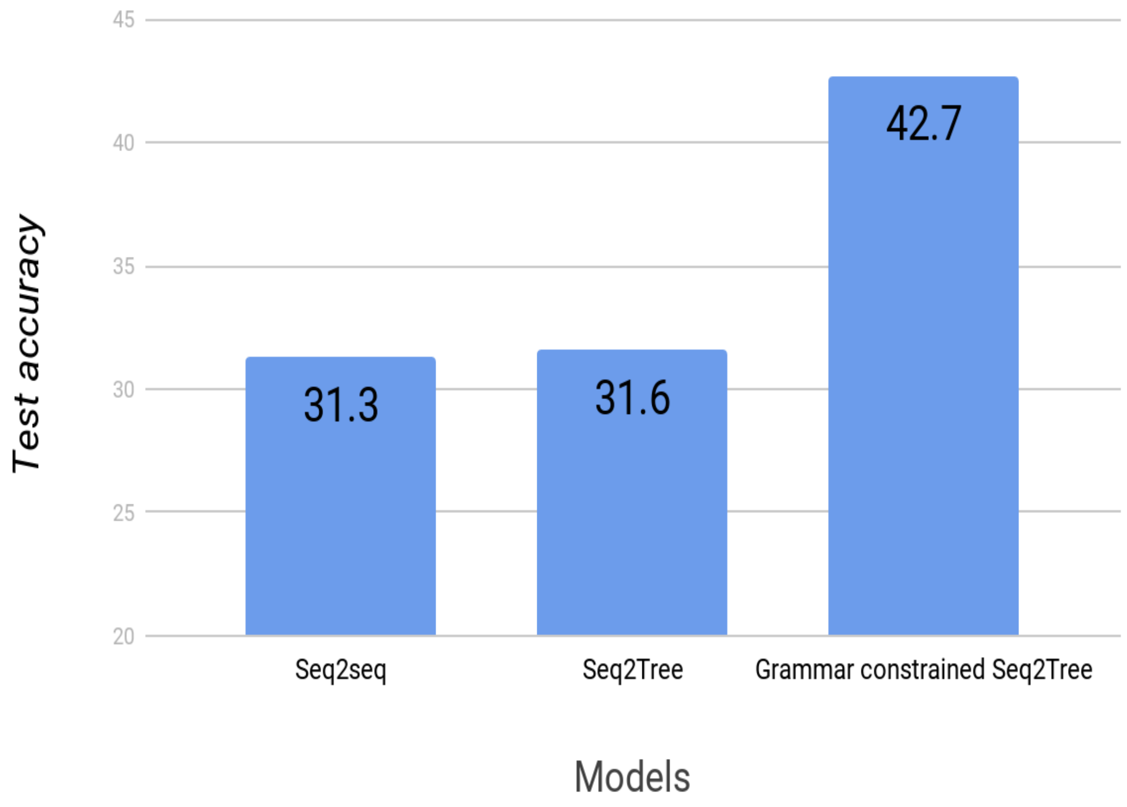
x: NL term
y: script term
d: denotation

with$\mathcal{C}$ a linear combination of coverage and denotation costs

$$\mathcal{C}(x_i, y_i, w_i, d_i) = \lambda \mathcal{S}(y_i, x_i) + (1 - \lambda) \mathcal{T}(y_i, w_i, d_i)$$

(Krishnamurthy, Dasigi and Gardner, 2017)

Empirical comparison on WikiTableQuestions

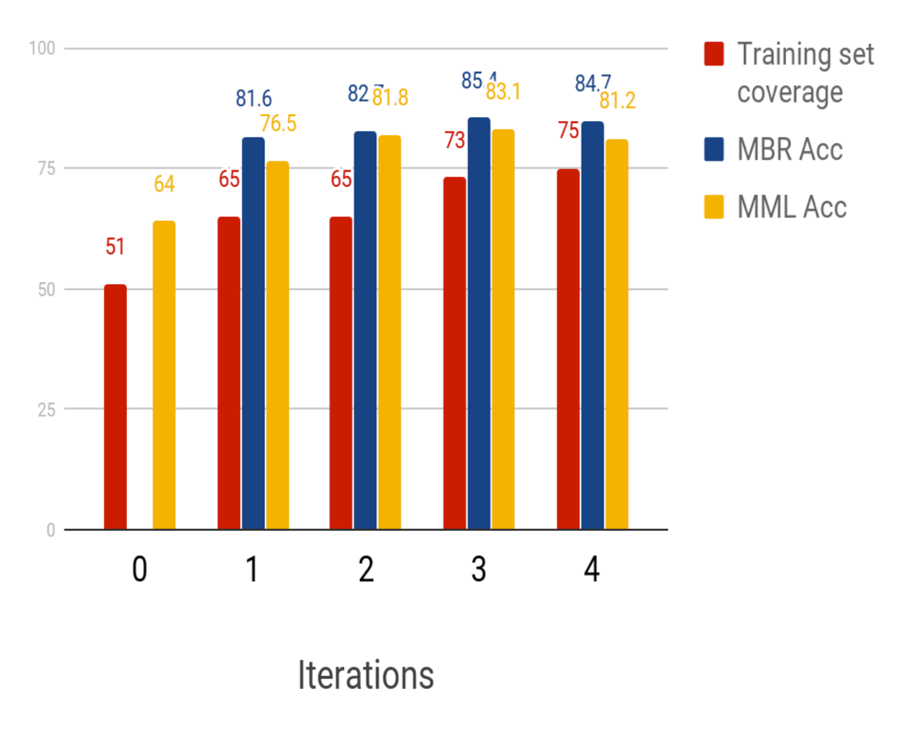


- Requires approximate set of logical forms during training
- Used output from Dynamic Programming on Denotations (Pasupat and Liang, 2016)
- Various models: strings, trees, etc.
- Efficient search followed by pruning using human annotations

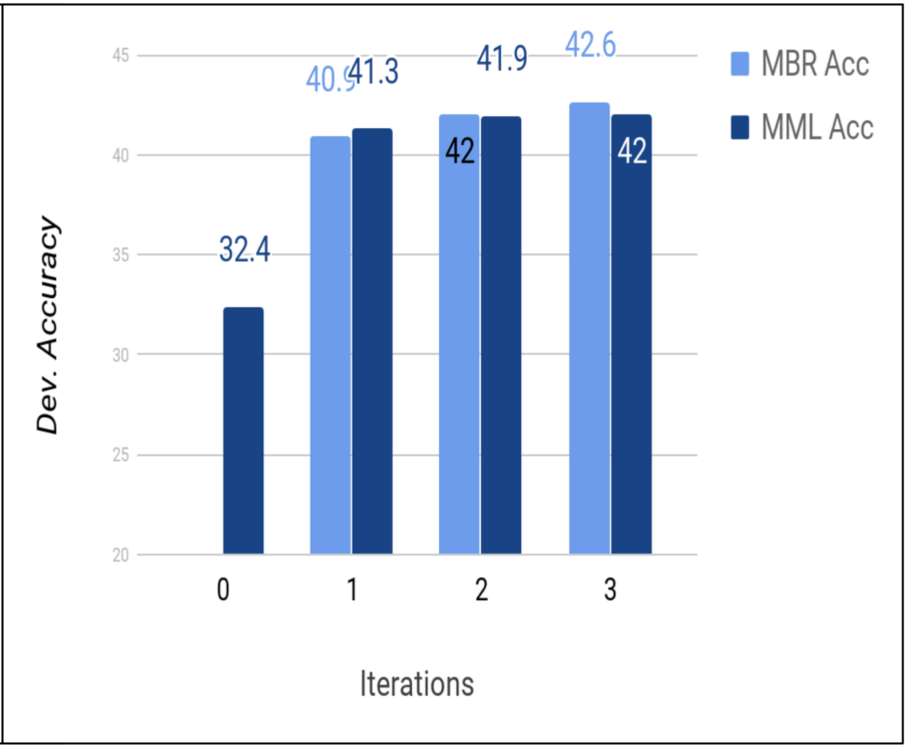
(Dasigi, Gardner, Murty, Zettlemoyer, Hovy 2018)

Dasigi results using iterative search

NLVR



WikiTableQuestions



- Similar trend in 2 domains
- Used functional query language (Liang et al., 2018)

Conclusion for option 4.1

Interesting idea to ‘operationalize’ the Q and test its ‘truth’ by running the script for the A

But works only with structured A sources where such operationalization is possible

Can we ‘operationalize’ other, typical kinds of Qs?

Type 2: A new QA task: Multi-domain knowledge

Q: What is the largest capital city south of Santiago de Chile?

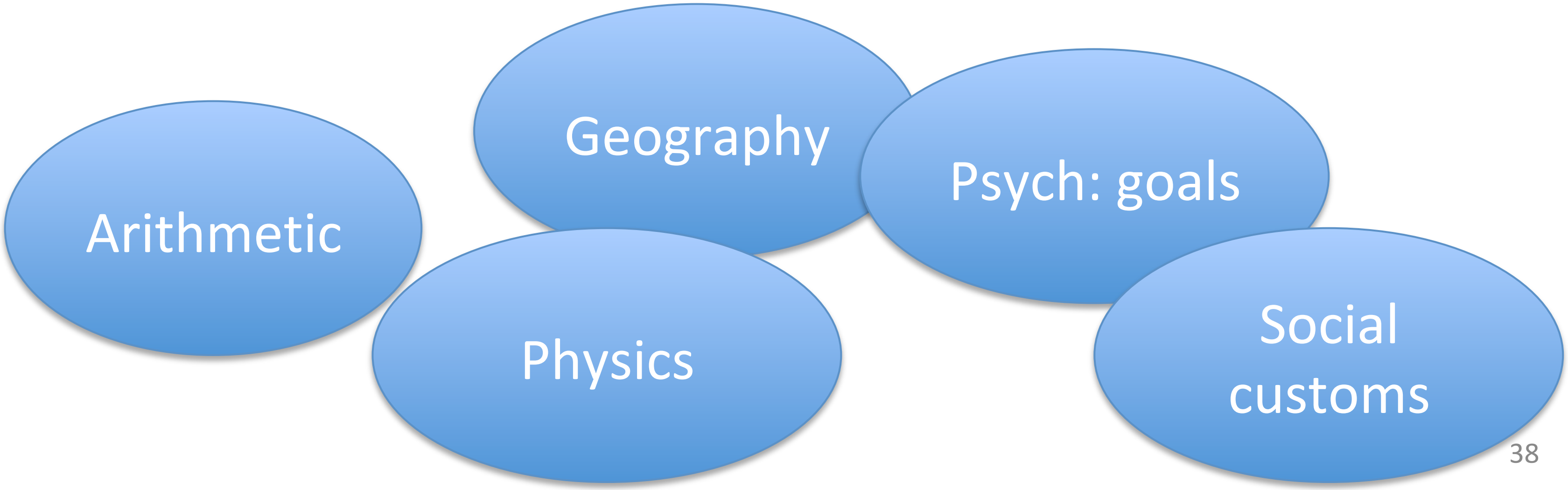
- Geographic knowledge (lat-long, population)
- Numerical ability (sorting, etc.)

Q: Which of the leaders of the XYZ enterprise are well-liked, and why?

- Discovery of social role by actions
- Sentiment judgments attached to actions

Multi-domain knowledge

- Define *N* self-contained standardized ‘domain specialists’ (KBs+reasoners) that <u>any</u> QA engine can run
- At run-time, analyze the Q, build the A script, activate the specialists as needed, compute the A



Arithmetic

Geography

Physics

Psych: goals

Social customs

Research needed

- For each domain specialist:
  - Define its ‘knowledge service’
  - Create the underlying knowledge
  - Define the I/O APIs for the QA engine to use
  - Build the specialist
- For each QA engine:
  - Analyze the Q —> determine parameters and need
  - Decompose need, build a script of specialist queries plus their result composition
  - Execute

Some specialist areas we are currently working on in my group

1. Arithmetic / numerical reasoning for entailment (Ravichander, Naik, Rosé, Hovy, CoNLL 2019, ACL 2019)
2. Psych goals for sentiment justification (Otani and Hovy, ACL 2019)
3. Social roles for group activity support (Yang, Kraut, Hov,y EMNLP, HCI, and others 2017–18)

(Ravichander, Naik, Rose, Hovy, 2019)

Topic 1. Numerical calculation

- Task: Entailment problem
- Input: clauses containing numbers
- Output: entailed / not-entailed

> P: 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

- Results:
  - EQUATE dataset extracted from ~8 existing QA and Entailment resources, with As added
  - Baseline numerical reasoner scores on the dataset

(Ravichander et al., 2019)

EQUATE corpus

Dataset	**Size**	**Classes**	**Synthetic**	**Data Source**	**Annotation Source**	**Quantitative Phenomena**
Stress Test	7500	3	✓	AQuA-RAT	Automatic	Quantifiers
RTE-Quant	166	2	✗	RTE2-RTE4	Expert	Arithmetic, World knowledge, Ranges, Quantifiers
AwpNLI	722	2	✓	Arithmetic Word Problems	Automatic	Arithmetic
NewsNLI	1000	2	✗	CNN	Crowd-sourced	Ordinals, Quantifiers, Arithmetic, World Knowledge, Magnitude, Ratios
RedditNLI	250	3	✗	Reddit	Expert	Range, Arithmetic, Approximation, Verbal

(Ravichander et al., 2019)

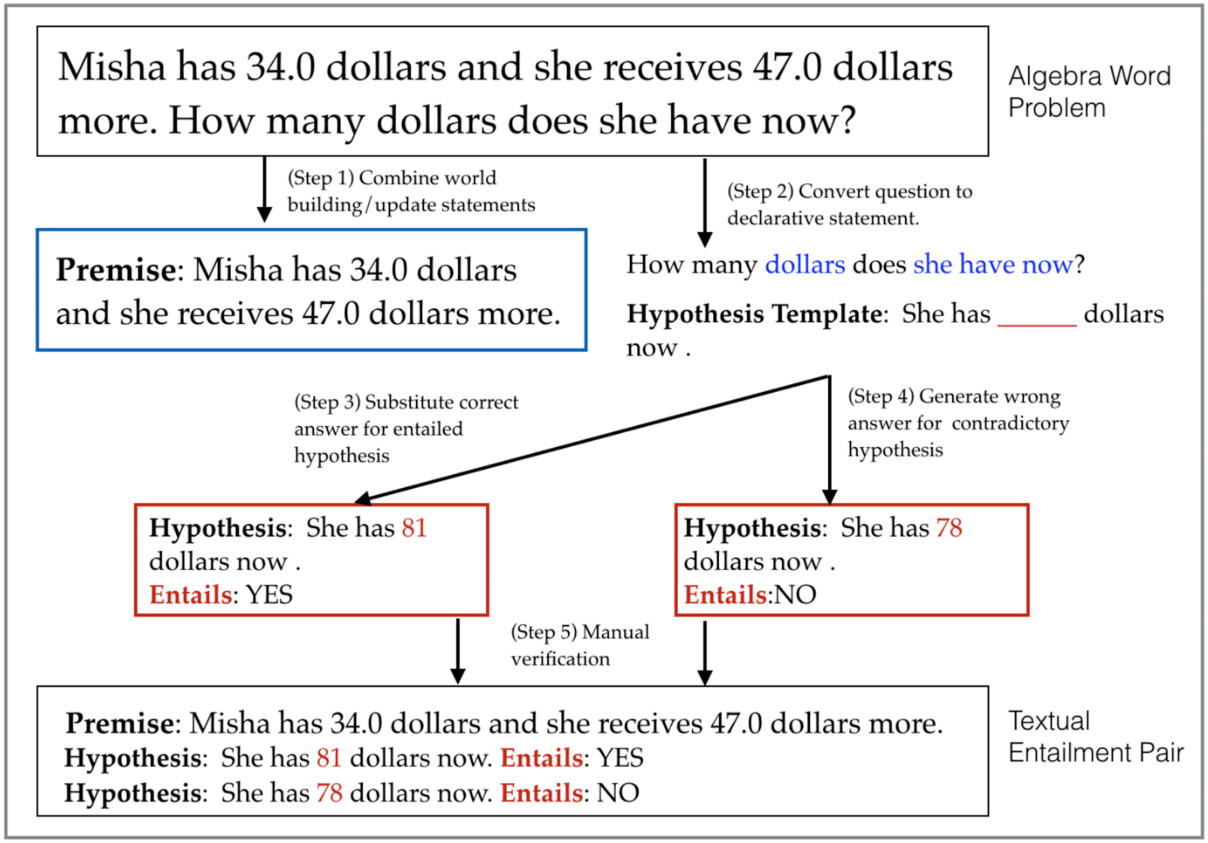
Baselines (SOTA methods)

- Majority Class (MAJ): Simple baseline always predicts the majority class in test set.
- Hypothesis-Only (HYP): FastText classifier trained on only hypotheses to predict the entailment relation (Gururangan et al. 2018)
- ALIGN: A bag-of-words alignment model inspired by MacCartney (2009)
- NB (Nie and Bansal 2017): Sentence encoder consisting of stacked BiLSTM-RNNs with shortcut connections and fine-tuning of embeddings. Achieves top non-ensemble result in the RepEval-2017 shared task
- CH (Chen et al. 2017): Sentence encoder consisting of stacked BiLSTM-RNNs with shortcut connections, character-composition word embeddings learned via CNNs, intra-sentence gated attention and ensembling. Achieves best overall result in the RepEval-2017 shared task
- RC (Balazs et al. 2017): Single-layer BiLSTM with mean pooling and intra-sentence attention
- IS (Conneau et al. 2017): Single-layer BiLSTM-RNN with max-pooling, shown to learn robust universal sentence representations that transfer well across inference tasks
- BiLSTM: We reimplement the simple BiLSTM baseline model of Nangia et al. (2017). Our reimplementation achieves slightly better results on the MultiNLI devset
- CBOW: Bag-of-words sentence representation from word embeddings passed through a tanh non-linearity and a softmax layer for classification.

(Ravichander et al., 2019)

Constructing entailment inferences

- Generate a report for each premise-hypothesis pair, consisting of:
  - Extracted NUMSETS for premise and hypothesis
  - Which NUMSETS were combined and by what operation
  - Which NUMSETS were justified and which weren't



Misha has 34.0 dollars and she receives 47.0 dollars more. How many dollars does she have now?

Algebra Word Problem

(Step 1) Combine world building/update statements

(Step 2) Convert question to declarative statement.

Premise: Misha has 34.0 dollars and she receives 47.0 dollars more.

How many dollars does she have now?

Hypothesis Template: She has _____ dollars now .

(Step 3) Substitute correct answer for entailed hypothesis

(Step 4) Generate wrong answer for contradictory hypothesis

Hypothesis: She has 81 dollars now .
Entails: YES

Hypothesis: She has 78 dollars now .
Entails:NO

(Step 5) Manual verification

Premise: Misha has 34.0 dollars and she receives 47.0 dollars more.
Hypothesis: She has 81 dollars now. **Entails**: YES
Hypothesis: She has 78 dollars now. **Entails**: NO

Textual Entailment Pair

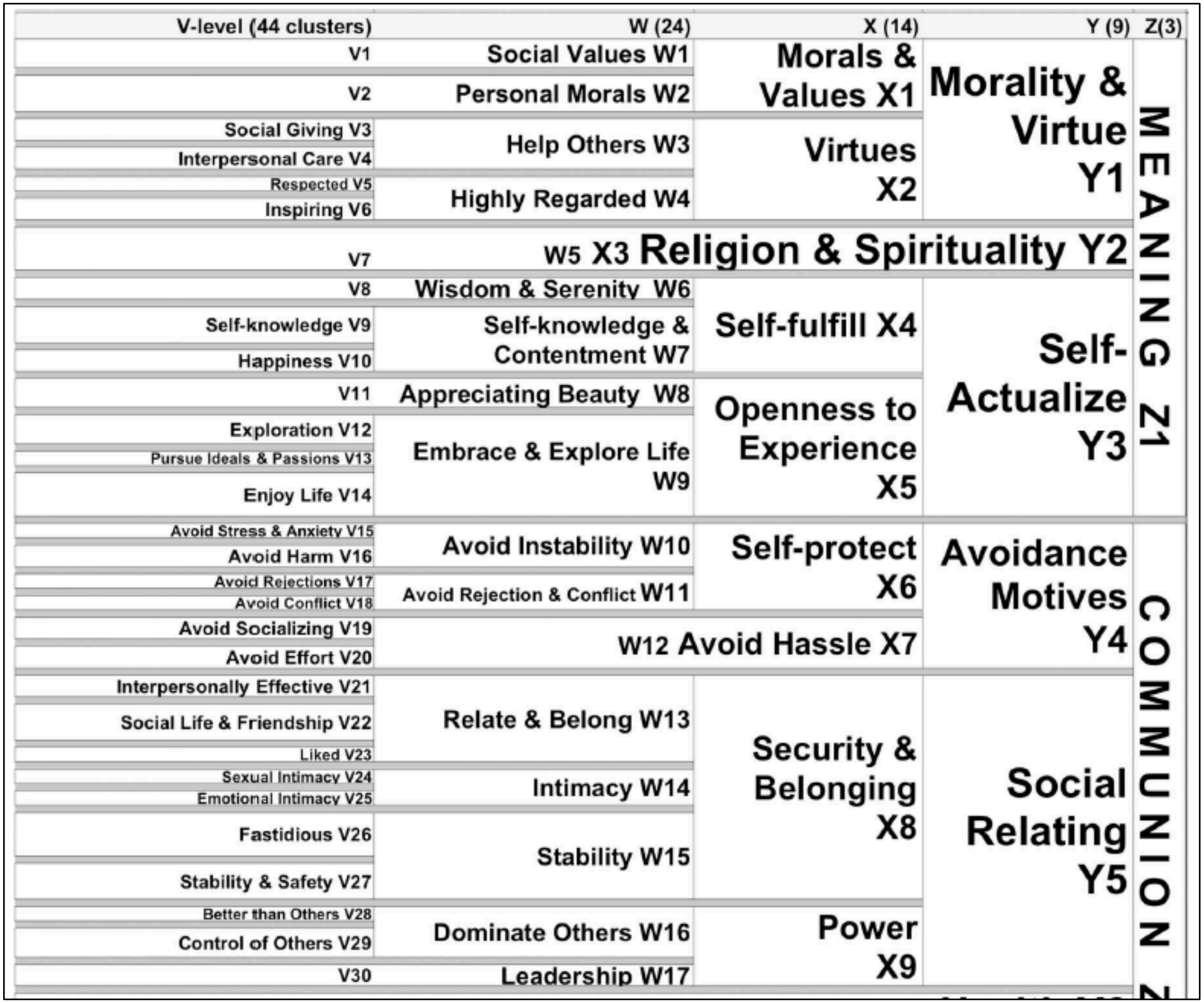
- Combines neural and symbolic programs
  - Some submodules are neural; overall framework is symbolic
  - Lightweight supervision

(Otani and Hovy, 2019)

Topic 2. Human goals

- Complex QA domain: human goal for sentiment
  - *I loved the hotel's price but the room was noisy* —> [price +] [room -]
- **Task: sentiment justification**: WHY does the Holder have the sentiment value for the facet?
- Approach: Classify each clause into a list of human (psychological and social) goals
  - Initial set: Maslow hierarchy
  - Currently: ~110 human goals from USC (Talevich et al.)
- Data: Crowdsourced; κ ≈ 0.55

(Talevich et al. 2017)



V-level (44 clusters)	W (24)	X (14)	Y (9)	Z(3)
V1	Social Values W1	Morals & Values X1	Morality & Virtue Y1	MEANING Z1
V2	Personal Morals W2			
Social Giving V3	Help Others W3	Virtues X2		
Interpersonal Care V4				
Respected V5	Highly Regarded W4			
Inspiring V6				
V7	W5	X3	Religion & Spirituality Y2	
V8	Wisdom & Serenity W6	Self-fulfill X4	Self-Actualize Y3	
Self-knowledge V9	Self-knowledge & Contentment W7			
Happiness V10				
V11	Appreciating Beauty W8	Openness to Experience X5		
Exploration V12	Embrace & Explore Life W9			
Pursue Ideals & Passions V13				
Enjoy Life V14				
Avoid Stress & Anxiety V15	Avoid Instability W10	Self-protect X6	Avoidance Motives Y4	COMMUNION Z
Avoid Harm V16				
Avoid Rejections V17	Avoid Rejection & Conflict W11			
Avoid Conflict V18				
Avoid Socializing V19	W12	Avoid Hassle X7		
Avoid Effort V20				
Interpersonally Effective V21	Relate & Belong W13	Security & Belonging X8	Social Relating Y5	
Social Life & Friendship V22				
Liked V23				
Sexual Intimacy V24	Intimacy W14			
Emotional Intimacy V25				
Fastidious V26	Stability W15			
Stability & Safety V27				
Better than Others V28	Dominate Others W16	Power X9		
Control of Others V29				
V30	Leadership W17			

(Yang, Kraut, Hovy, 2018)

Topic 3. Social roles

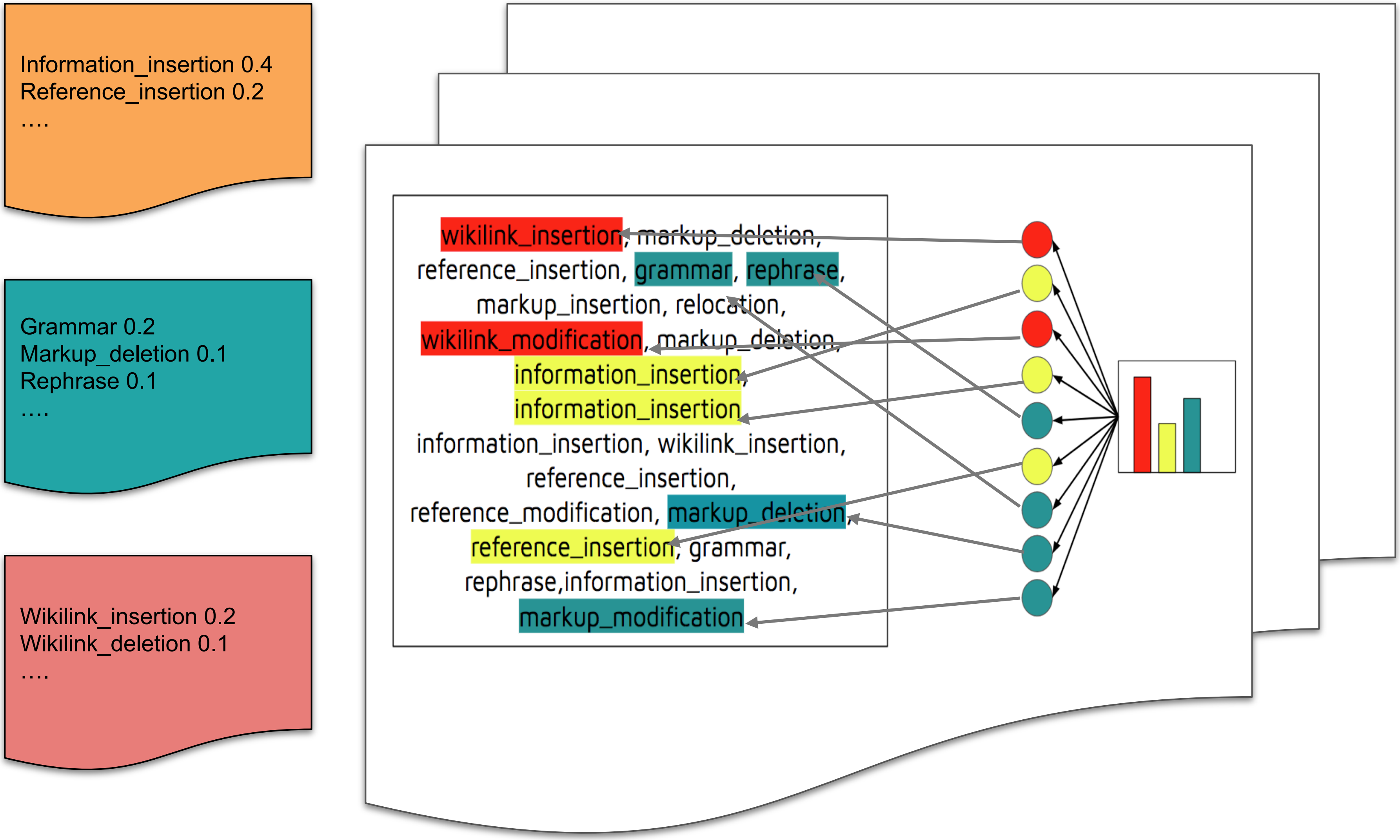
- Complex QA domain: Human interactions in groups
- Task: Automated social role discovery
  - Input: Discussions in a social media platform
  - Output: Role list, and assignment for each user
- Data:
  - Wikipedia editors: our role taxonomy conforms to Wikipedia’s internal set
  - Cancer Survivor Network discussion groups

(Yang, Kraut, Hovy, 2018)

Roles

User edit history

Role assignments



Information_insertion 0.4
Reference_insertion 0.2
….

Grammar 0.2
Markup_deletion 0.1
Rephrase 0.1
….

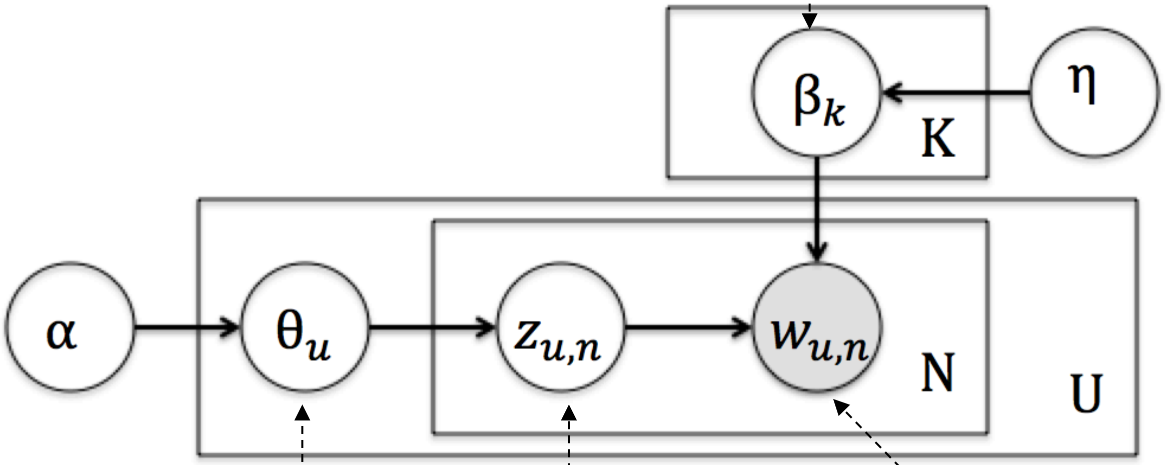
Wikilink_insertion 0.2
Wikilink_deletion 0.1
….

wikilink_insertion, markup_deletion,
reference_insertion, grammar, rephrase,
markup_insertion, relocation,
wikilink_modification, markup_deletion,
information_insertion,
information_insertion
information_insertion, wikilink_insertion,
reference_insertion,
reference_modification, markup_deletion,
reference_insertion, grammar,
rephrase,information_insertion,
markup_modification

(Yang, Kraut, Hovy, 2018)

Latent role model in Wikipedia

Role: distribution of edit actions



Role proportions

Edit actions

Role assignment for
user **u** and word **n**

(Yang, Kraut, Hovy, 2018)

Discovered editor roles (naming by expert)

Expert's role name	Discovered representative behavior
Substantive Expert	Information insertion, wikilink insertion, reference insertion
Social Networker	Main talk namespace, user namespace
Vandal Fighter	Reverting, user talk namespace
Quality Assurance	Wikilink insertion, wikipedia namespace, template namespace
Fact Checker	Information deletion, wikilink deletion, reference deletion
Cleanup Worker	Wikilink modification, template insertion, markup modification
Fact Updater	Template modification, reference modification
Copy Editor	Grammar, paraphrase, relocation

Topics 4–. Other inference specialists

- Geography and Time… (see (Allen, CACM 1983) and (Davis, JAIR 2017))
  - E.g.: *north-of, area-included-in-region*…
- Physics, Biology… (see the HALO project)
  - Recent work on aspects of Physics at AI2 (Clark et al.)
- Emotions

Physics: noun-noun compounds

Where is…

- …the kitchen table
- …the coffee table
- …the wood table
- …the teacher’s table
- …the data table

- Need to know the relation and the noun types to infer additional info:
- LOC
- FUNCTION ➔ LOC
- MATERIAL
- ?FUNCTION ➔ LOC ?
- TYPES ➔ CONTENT ➔ LOC?

Conclusion for option 4.2
Where next with Complex QA?

- Identify and build the most useful domain specialists
  - Find basic knowledge primitives
  - Develop reasoning logics, models, and implementations
  - Develop / find QA datasets that exercise this sort of specialist knowledge and reasoning
    - Great overview in (Davis, JAIR 2018)
- Create a common library for all to share
- Evaluate **correctness** AND Answer production **scripts** (traces, as 'explanation')

Open-source and general-purpose (not just scientific/ political) version of Wolfram Alpha

THANK YOU