From Simple to Complex QA

Eduard Hovy

CMU

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*

Dukakis unable to uncover the truth on crime charge

For Democrats, what happened in the presidential election last week was criminal — in two ways.

One was that they felt robbed of the White House.

The other involved crime itself. A major tactic of the George Bush campaign was to portray Michael Dukakis and the liberal Democrats as soft on crime.

The fact that Dukakis failed to counter this tactic effectively helped him lose, and it has also left Democrats much weaker, perceived as spineless liberals whose first instinct is to coddle criminals.

This perception outrages many frustrated Democrats — none more than Richard G. Stearns.

Robert L. Turner

In modern times, crime has usually been considered a local issue. Calvin Coolidge gained political capital by putting down the Boston police strike in 1919, but Richard Nixon's law-and-order harangues in 1968 were the first in recent times to raise it as a national issue.

Stearns said he believes then-Vice President Spiro Agnew ratcheted fear of crime up a notch in the congressional elec-

was not much of a leap to minimize the blame attached to criminals for crime.

In some cases, Stearns said, liberals have exhibited "an undue optimism about human nature and its capacity for reform." But nothing in the liberalism philosophy that mandates a soft attitude toward crime.

In fact, in one important sense, the opposite is true. For the victims of crime, overwhelmingly, come from the disadvantaged and minority groups that are the prime constituencies of Democrats, whether liberal or not.

Indeed, Dukakis has favored a number of tough approaches to crime, including heavy sentences for drug violations, sei-

Two ways to make things more complex

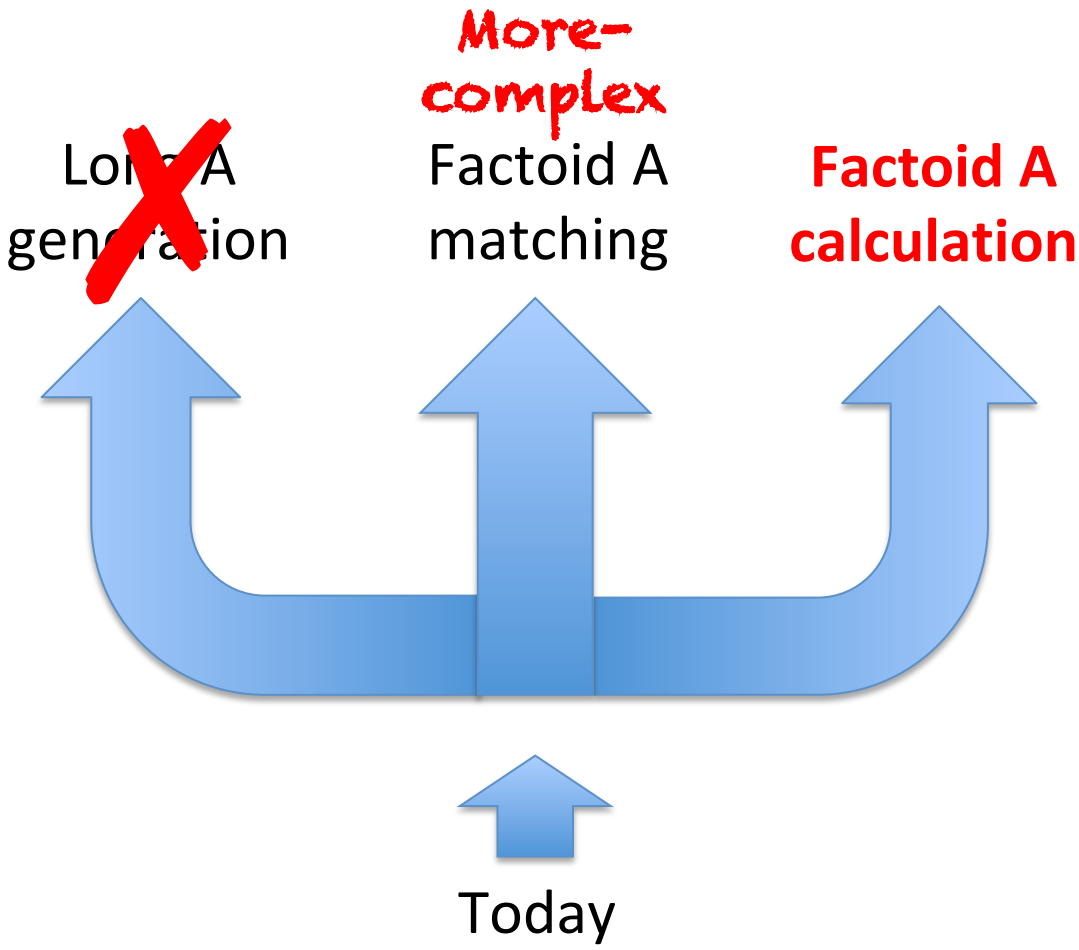
1. **Complex <u>answers</u>**: Long, non-factoid responses. (Q: summarization, text planning, NL generation)

We leave this for another time

2. **Complex <u>answering</u>**: A procedure more than simple atomic pattern matching

So, what to do?

What is ‘complex QA’?



Outline

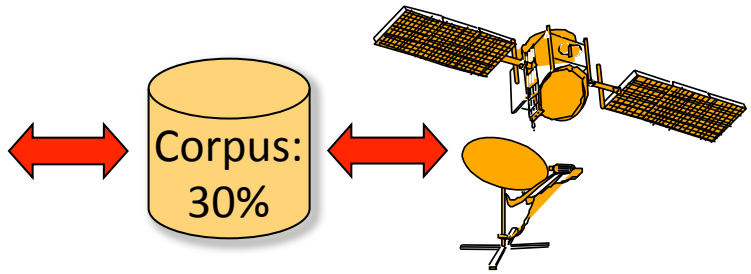
- Simple QA: Matching
- Complex QA: Calculating
- The Conundrum
- Two Paths Forward

SIMPLE QA: MATCHING

Basic simple QA architecture

Input Q

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



– 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

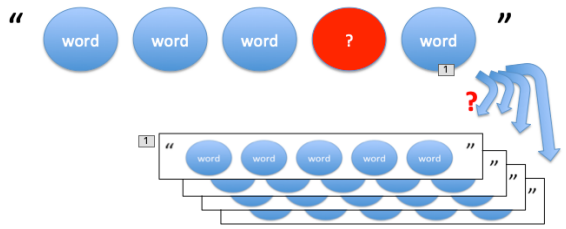
A list

Patterns

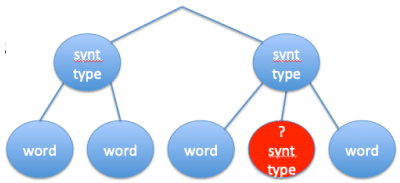
The main research lies in **building/learning**, **specializing**, and **matching** A graphs/patterns:

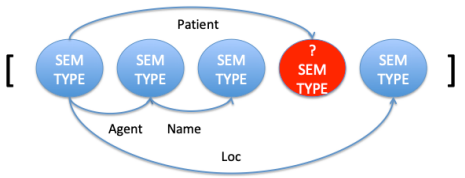
- Manual creation and string matching
- Learned as per Info Extraction: recursive pattern learning, simple matching
- Learned in neural architectures, matching is automatic

Increasingly automated



Pattern strings, trees, frames, graphs…





In fact, all QA processing be seen as matching a Question graph against a set of (normalized / expanded) candidate Answer graphs

with the best match(es) providing the Answer

With enough pattern engineering…

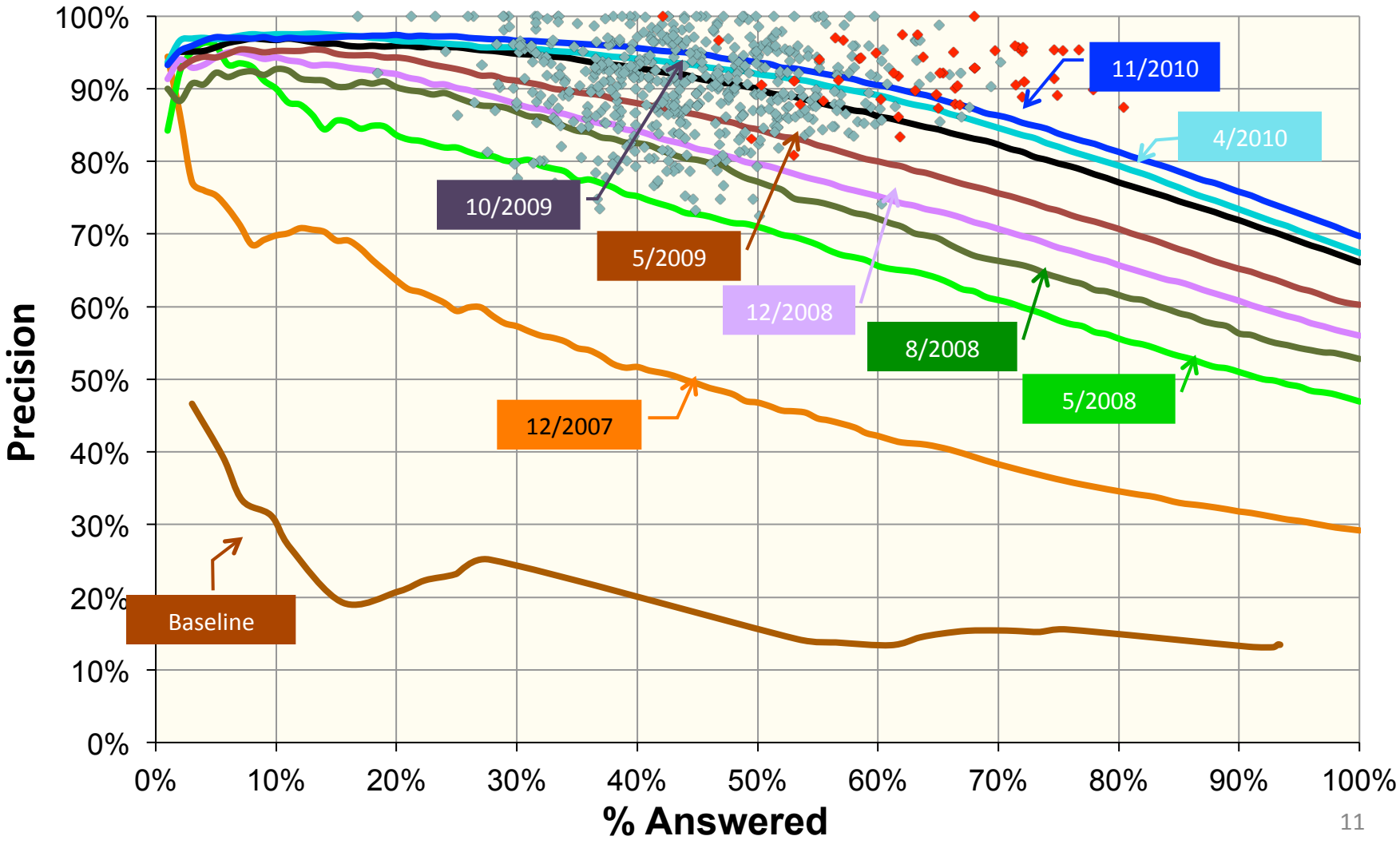
With enough pattern engineering

and some specialized reasoning (rhymes, geospatial and temporal reasoning, etc.)…

…you get IBM's Watson

slide by IBM

Watson's incremental progress in Precision and Confidence



The ultimate pattern technology

- A pattern must capture
  - The sequence of relevant words/types
  - The target word/type [family of synonyms]

- NNs do exactly this
  - Ngram models of word embeddings
  - Embeddings that generalize words

Embeddings make pattern learning and engineering much easier

- A word embedding is a 'local generalized ngram model' for the word
  - captures the word's general context expectations
  - captures (semantic) substitutions of the word
- A sentence embedding is a 'generalized ngram sequence model' for the Question / sentence
  - Ngrams are weighted in different ways depending on NN architecture (direction of LSTM, attention, etc.)
  - People invent architectures to highlight (or smooth out, using adversarial approaches) specific parts
  - The important parts (words for Q parameters, inter-part relations, the A introducer words, etc.) form the 'pattern'

Pattern-level evolution

Method	Resource
• String match	• Surface pattern library
• Synonym substitution	• Synonym dictionary, WN
• Typed pattern match	• QA typology / Type hierarchy
• Parse tree match	• Parser + Tree transform rules
• Shallow semantic match	• Semantic analyzer + Type hierarchy + partial unifier
• Embedding match	• Simple NN

- Attention-focused BiLSTM, etc.

BUT: Patterns that are too good give a false sense of accomplishment

Trained with enough embeddings…

…the NN's word and word-combination models capture world knowledge factoids and relations (incl. their surface-level cues for semantic structure)

So you get what looks like semantic QA

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 blah 8 to 10 hours blah blah blah blah Gatun Lake blah

Which oceans does the Panama Canal connect?

In your training data, you have surely seen “Panama Canal” with only two ocean names…

How powerful are ngram pattern models?

- CLOTH: Large-scale Cloze Test Dataset Created by Teachers (Xie, Lai, Dai, Hovy, EMNLP 2018)
- Large-scale cloze test dataset collected from English exams in China (Middle and High school level)
  - After cleanup: 7k passages; 99k questions (2/3 removed)
- The dropped words and word options were carefully created by teachers:
  - Highly nuanced alternatives
  - Test knowledge of grammar, vocabulary, reasoning
  - How well do state-of-the-art computational models do in comparison to humans? (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

- 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

- 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:

Test data \ Train data: α%	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

So, I conclude:
Given enough training data…

(and assuming the Q provides context text)

…you will always learn good patterns/word combination graphs that connect Q parameters <–> Q context material <–> A…

(If you have not seen the necessary combinations, you won't be able to answer the Q)

…to the degree that…

…you may not even *need* the Q context (or if you have the Q context, you *may not need the Q*!!):

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)

AHEM!!

Kaushik and Lipton EMNLP 2018

- Research goal:
  - How strong are models that see the **question 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 passage needed for engine, randomize its words first
  - If candidate As needed, they are placed in random spots, intervening text filled with gibberish
- **Passage-only** setting:
  - Create corrupted versions of each dataset: assign Q to some passage randomly

(Kaushik and Lipton EMNLP 2018)

Example: Q only

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.

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 _____

Answer: kakuei tanaka

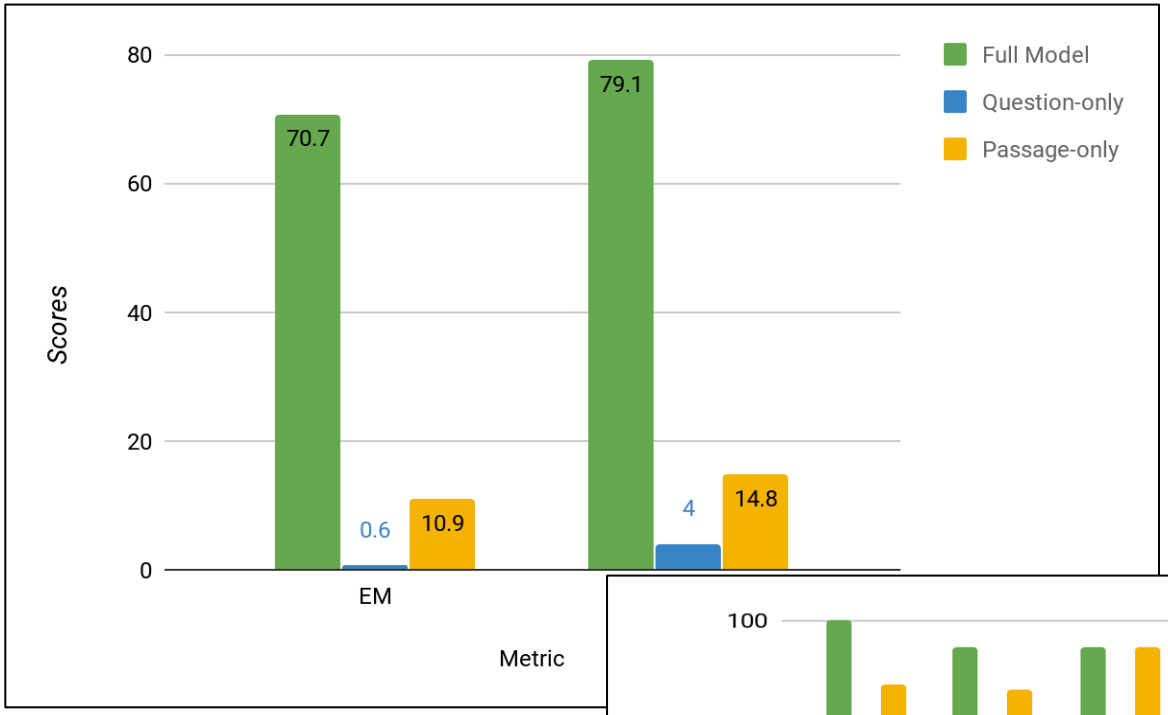
(Kaushik and Lipton EMNLP 2018)

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, Alexander, et al. 2016. Key-Value Memory Networks for Directly Reading Documents. Proceedings of the EMNLP conference.
  - **Gated Attention Readers**:
    Dhingra, Bhuwan, et al. 2017. Gated-Attention Readers for Text Comprehension. Proceedings of the ACL conference (Long Papers).
  - **QANet**:
    Yu, Adams Wei, et al. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. Proceedings of ICLR.

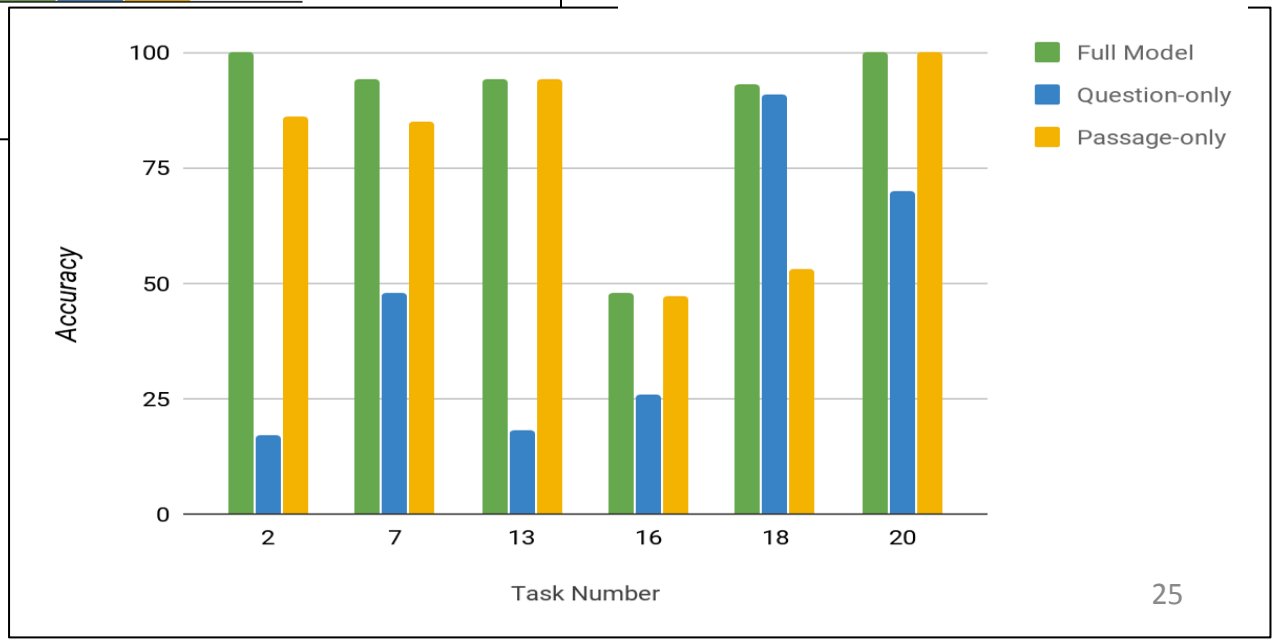
(Kaushik and Lipton EMNLP 2018)

Some results

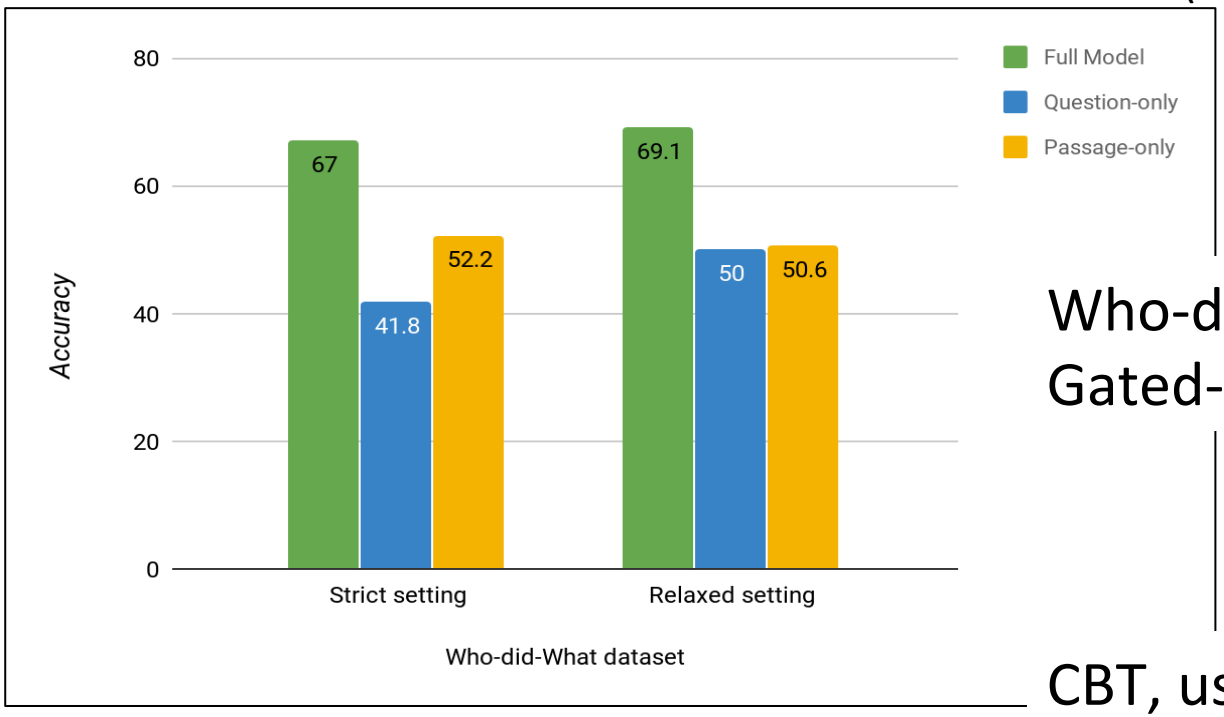


SQuAD, using QANet

bAbI, using Key-Value MemNets

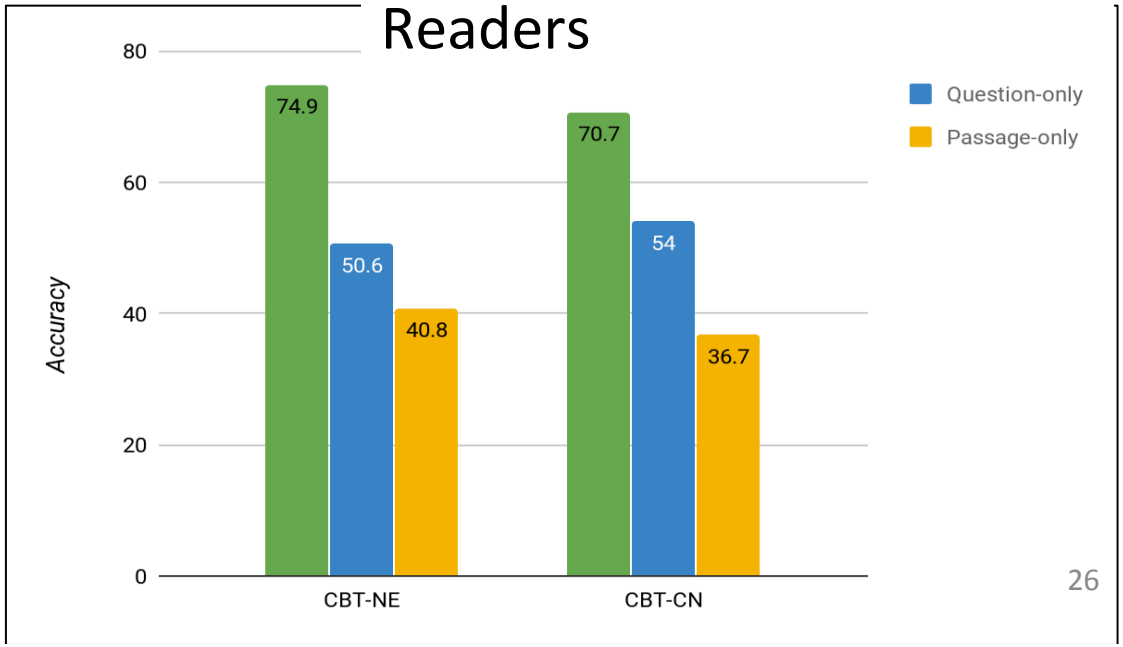


(Kaushik and Lipton EMNLP 2018)



Who-did-What, using Gated-Attention Readers

CBT, using Gated-Attention Readers



t's going on??

Name not in Google

Kanemaru's secretary

Transportation company

Long-term politician

Passage: ... glynis bc-nj-zimmer-profile-2takes-nyt **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** dea internatinoal 90-meter **kakuei tanaka** seven-paragraph 577,610 wendover golf-lpga-jpn partner, ur... ap... mazzei canada-u.s.

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 _____

Answer: kakuei tanaka

Lessons learned

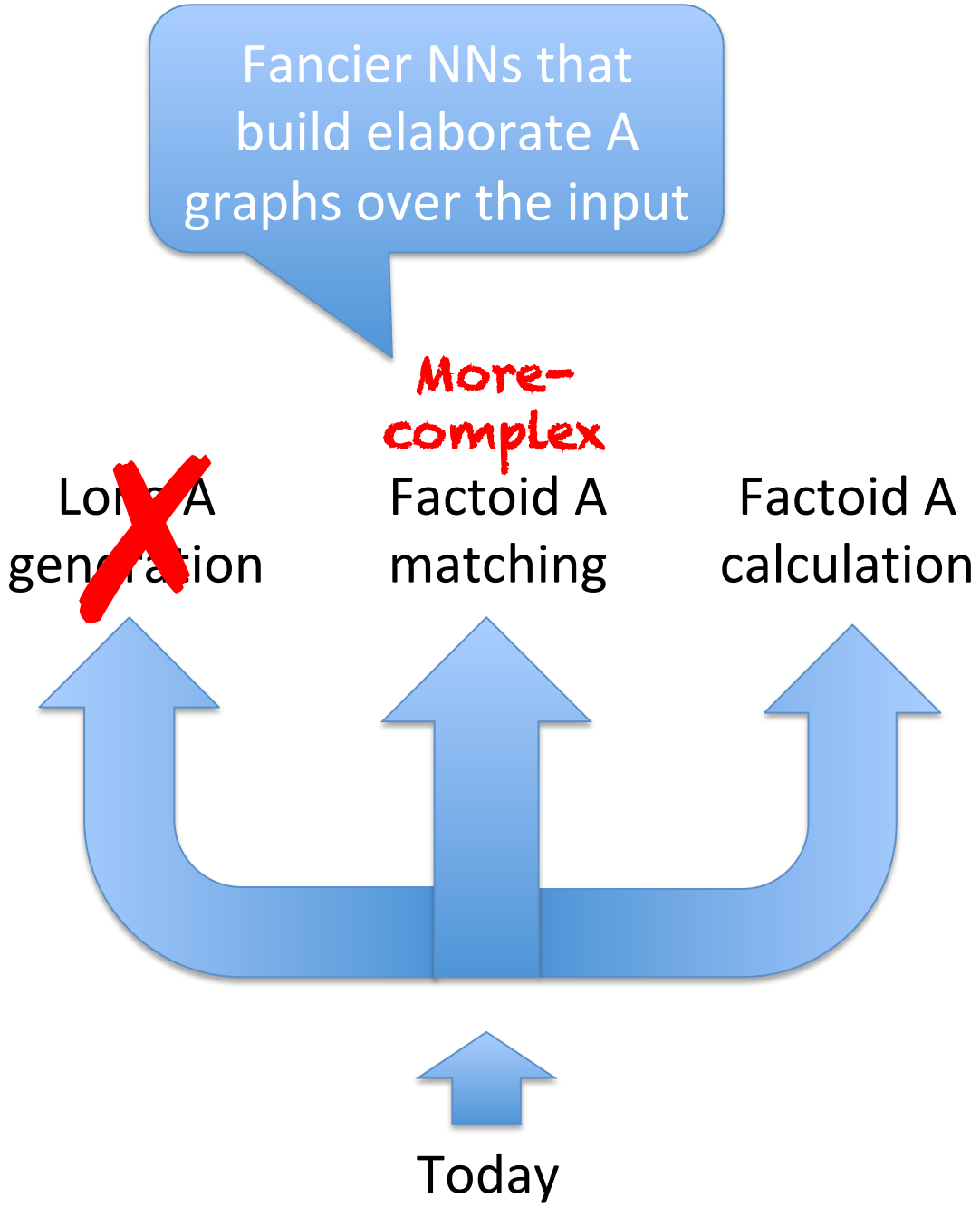
- This study was great but not perfect — it should have checked for pre-existing dependencies among the Q and the candidate As

- Still, it shows: you must provide rigorous baselines, for both **datasets** and **models**
- And you must test that **full context is essential** for the task — no hidden dependencies anywhere!

Where next with this approach?

- Design larger and fancier NN architectures (Feedforward —> BiLSTM —> BiLSTM with Attention to Qtype word —> …) that build increasingly complex generalized 'recognizer graphs'
- In the limit (with enough training data), they will identify all relevant Q parameters in the correct configuration and pinpoint the A

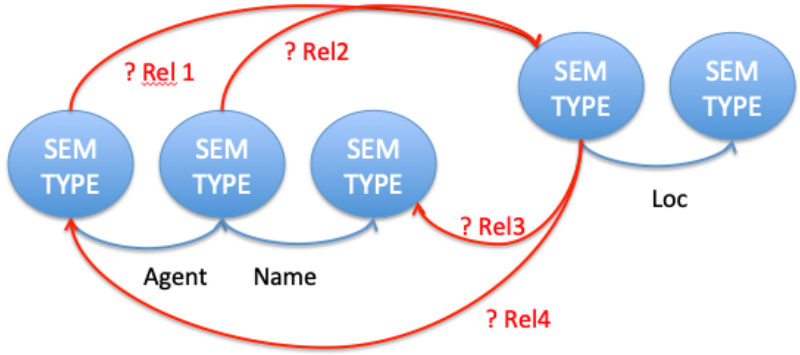
BUT: This works only when all the relevant info is explicitly present


Fancier NNs that build elaborate A graphs over the input
More-complex
Long A generation
Factoid A matching
Factoid A calculation
Today

COMPLEX QA: CALCULATING

QA system development

- Series of specialized subtasks:
  - Produce a Question graph
  - Get multiple candidate Answer texts
  - Produce their graphs
  - Canonicalize them into types
  - Match and get a goodness score
  - Rerank the As and deliver list
- When some info is *missing*…
  - The candidate A graph is disconnected
  - The Q graph cannot match anywhere
  - Need background knowledge (and reasoning?) to connect up the A graph fragments



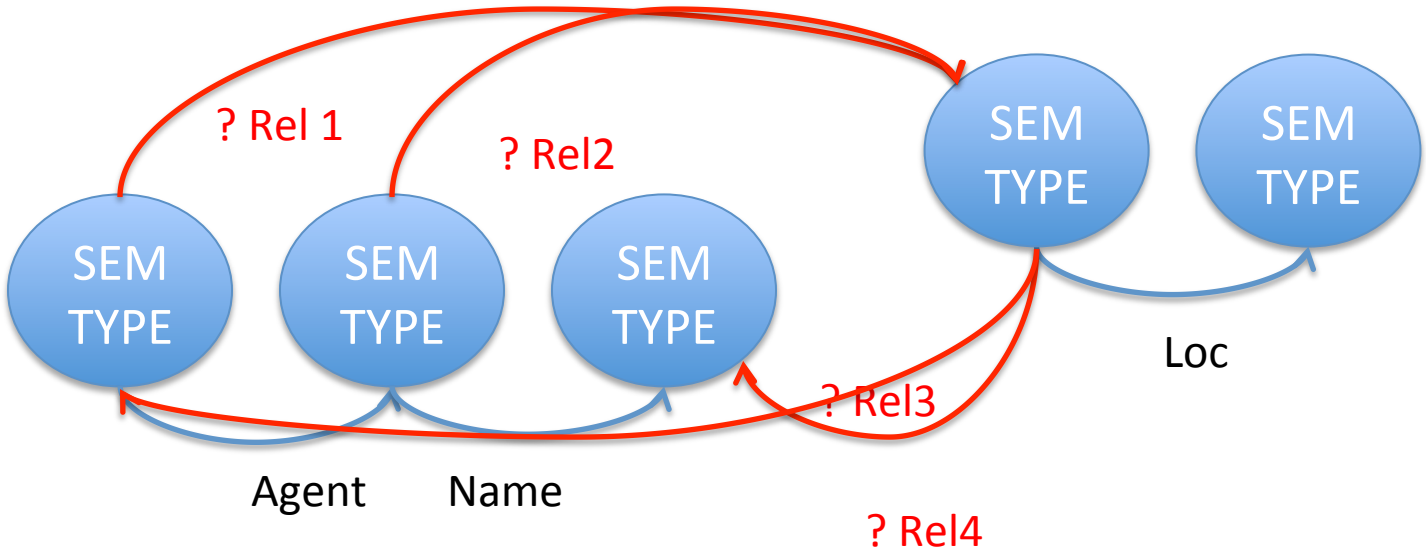
- Modern commercial QA work:
  - Induce Answer types from question logs, and create [thousands of] latent A dimensions
  - Automatically learn patterns associated with answer dimensions
  - Harvest and prepare tons of answers — cover 98% of effective Q space of question logs

QA research has done all the easy cases …but what if some graph info is *missing*?

The candidate A graph is disconnected

The Q graph cannot match anywhere

How to connect up the A graph fragments?



Need background knowledge!

In every case, you need background knowledge

- Easy knowledge: Transform A string into matchable form:
  - Synonym substitution
  - Semantic type normalization
  - Parse tree normalization (tenses, passive–>active, etc.)
- Hard: Add new nodes and links between subgraphs:
  - Provide new relations
  - Provide additional nodes

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

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)

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:

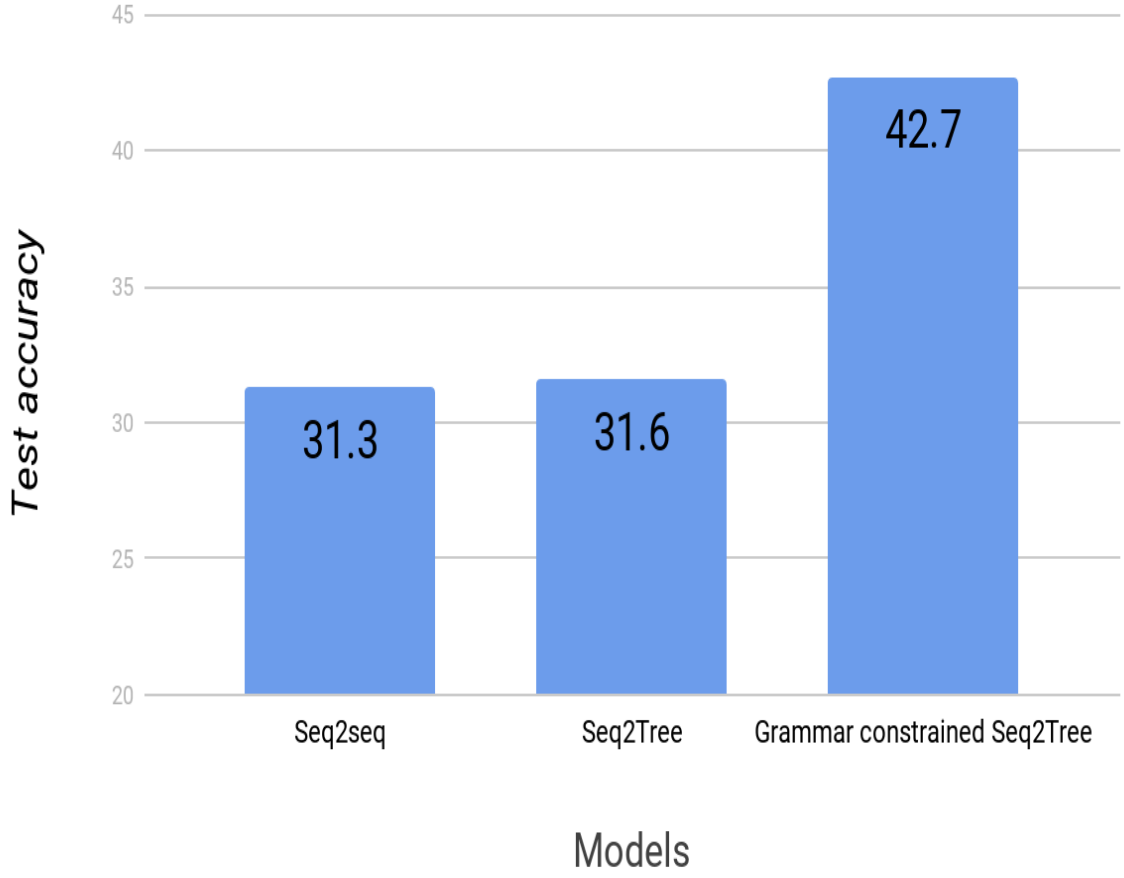
```
((reverse athlete) (and
      (nation south_korea)
      (year ((reverse date)
                  (>= 2010-mm-dd)))
```

Example: Dasigi thesis 2018

- Approach for learning to build access routines:
  - Parse Q, build dependency tree
  - Convert into Logical Form
  - Translate into candidate table access routine
  - Test composition by repeated trial and error
- Essentially, learning is a search in 'operator combination space' to build logical form. Speed up 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')

(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: strongs, trees, etc.
- Efficient search followed by pruning using human annotations

The problem is learning to construct the answer script

- Weak supervision is not enough:
  - Incorporating knowledge of grammar constraints helps
  - Handling spurious examples (right answer for wrong reasons)
  - Considering coverage of cases: Use overlap as a measure to guide search
- Combine into single Objective: Minimize expected value of cost (Goodman, 1996; Goel and Byrne, 2000; Smith and Eisner, 2005)

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

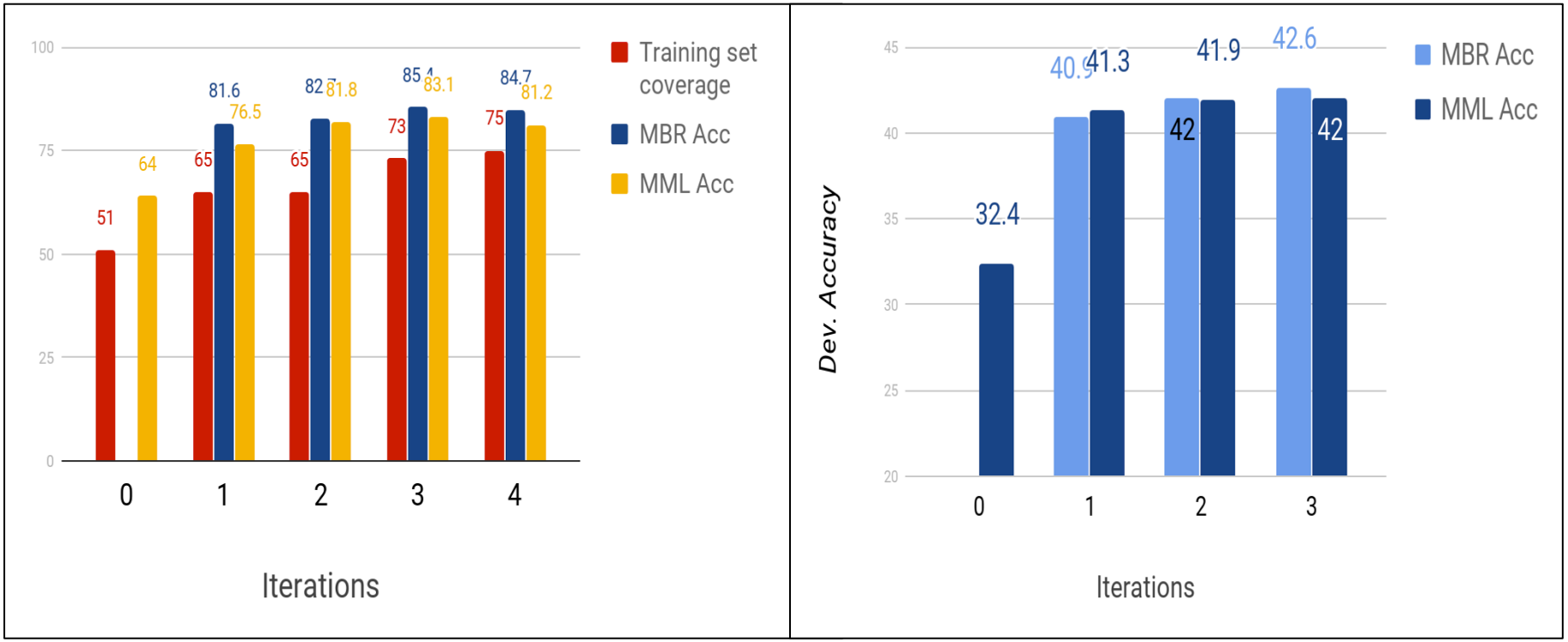
with $\mathcal{C}$ a linear combination of coverage and denotation costs

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

- Implement iterative search (from simpler to more complex)

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

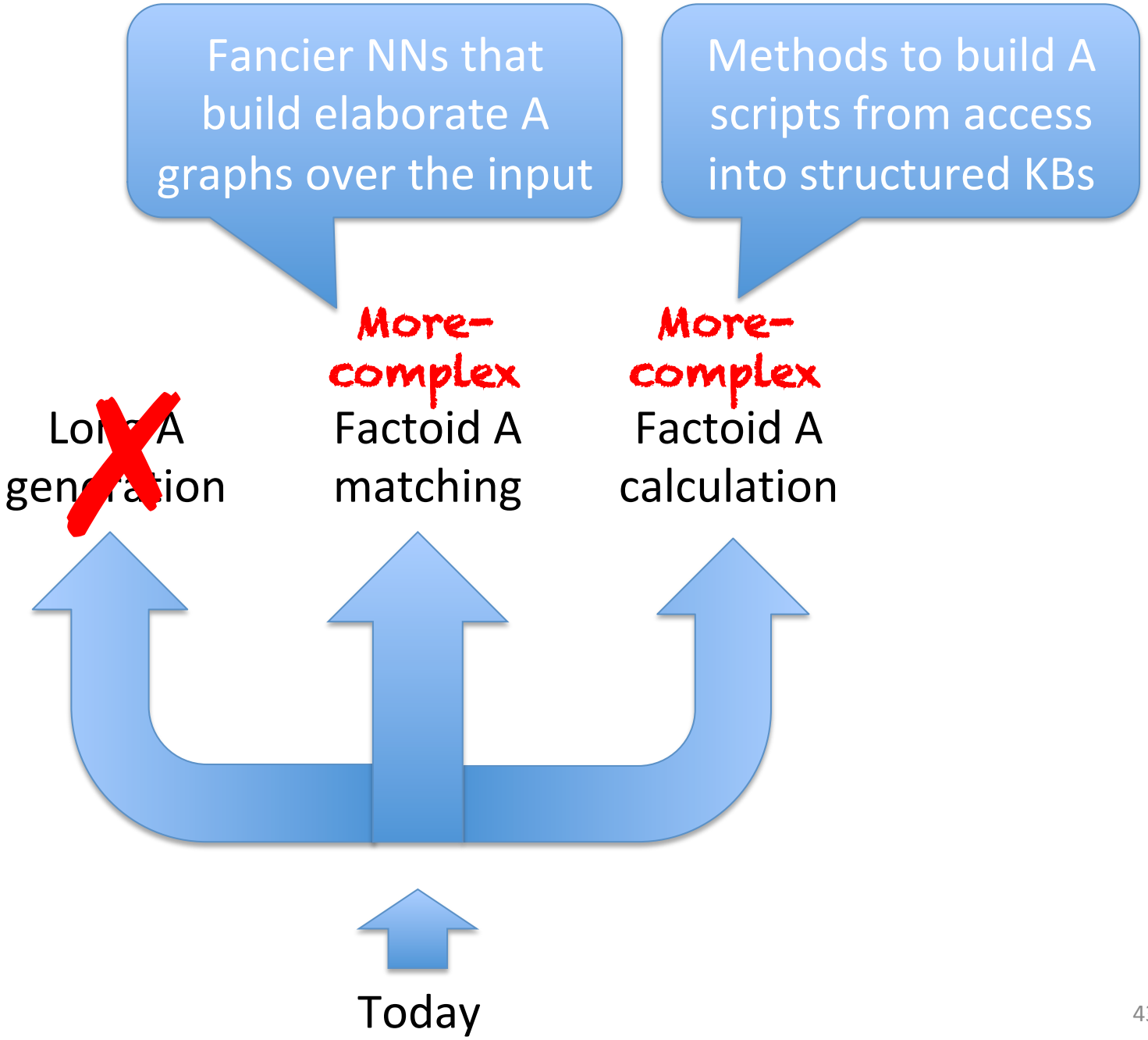
Results of training with iterative search on WikiTableQuestions



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

Where next with this approach?

- Better ways to learn to build the A retrieval+composition script (= mapping from English Q to nested operators and variables):
  - Learning from corpora of Qs and scripts (like Mooney et al.)
  - Learning by guided search through 'operator space' (like Dasigi and others)
- Ways to *prove* correctness of the script


Fancier NNs that build elaborate A graphs over the input
Methods to build A scripts from access into structured KBs
More-complex
More-complex
Factoid A matching
Factoid A calculation
Today

THE CONUNDRUM

Either…

All info needed to get the A is present in the Q context

…so some form of surface matching + sub-A composition suffices

—> Ultimately, nested simple QA (…still OK?)

Or…

Getting the A requires information **not** in the Q context: background knowledge, calculation, etc.

…but this is not standardized, hence impossible to evaluate

—> No complex QA !?

TWO PATHS FORWARD

Two types of ‘complex’ QA

Matching
(Fact[oid]-oriented)

How to deepen this?

- Create Qs needing multiple matches + dynamic composition
- Build this dataset
- Evaluate the composition

Computing
(Procedure-oriented)

How to deepen this?

- Identify general background knowledge all QA systems should have
- Build the dataset(s) of Qs and As requiring it

(Lai et al. 2018)

1. 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)

Kinds of more-complex matching

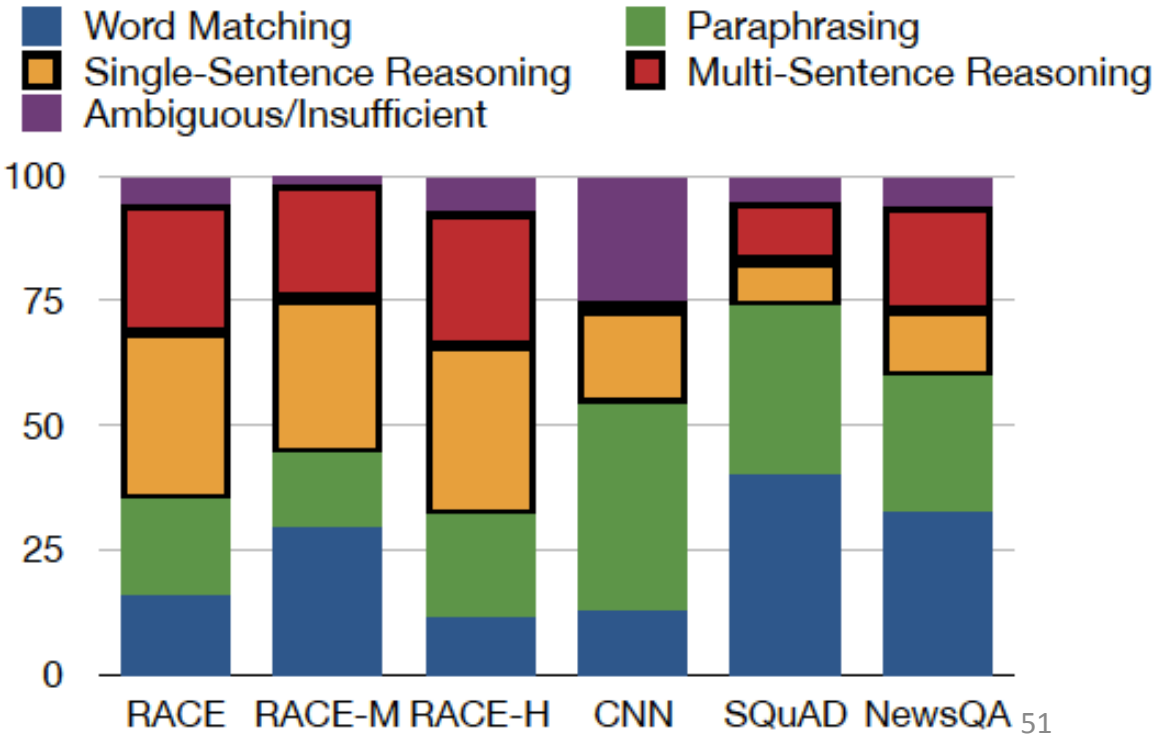
- **Detail** matching: understand details of the passage
- **Paraphrasing** Qs: test language ability
- **Whole-picture** matching: comprehend the entire story
- Passage **summarization** Qs: understand the point
- **Attitude** matching: find opinions/attitudes of the author towards something

- **World knowledge** Qs: use external knowledge such as simple arithmetic

(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)

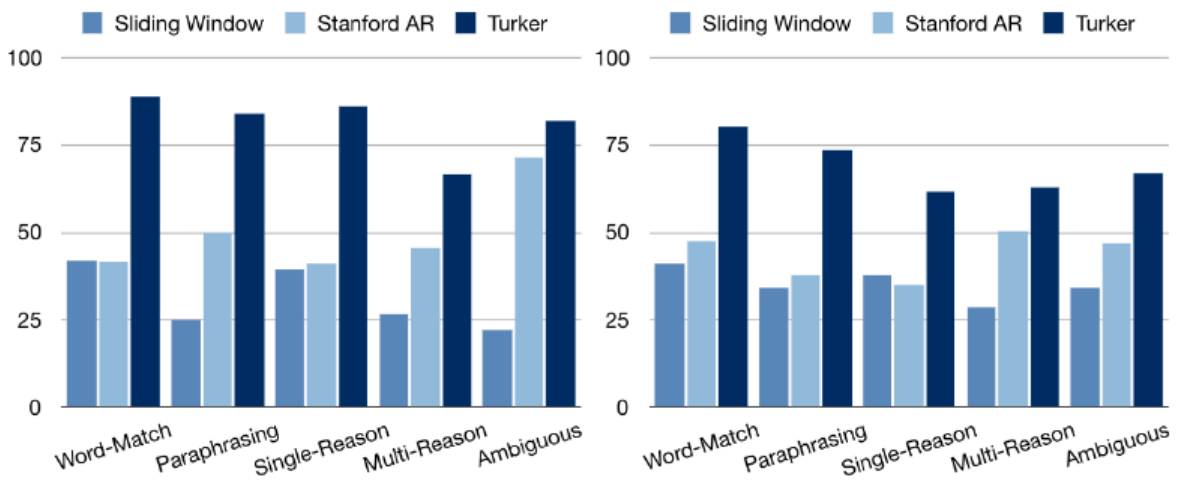
Comparing QA algorithms

	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: state-of-the-art neural models
- RACE has higher human ceiling performance, which shows the data is quite clean
- RACE is harder for AR models, proving a significant gap remains

(Lai et al. 2018)

Matching type performance



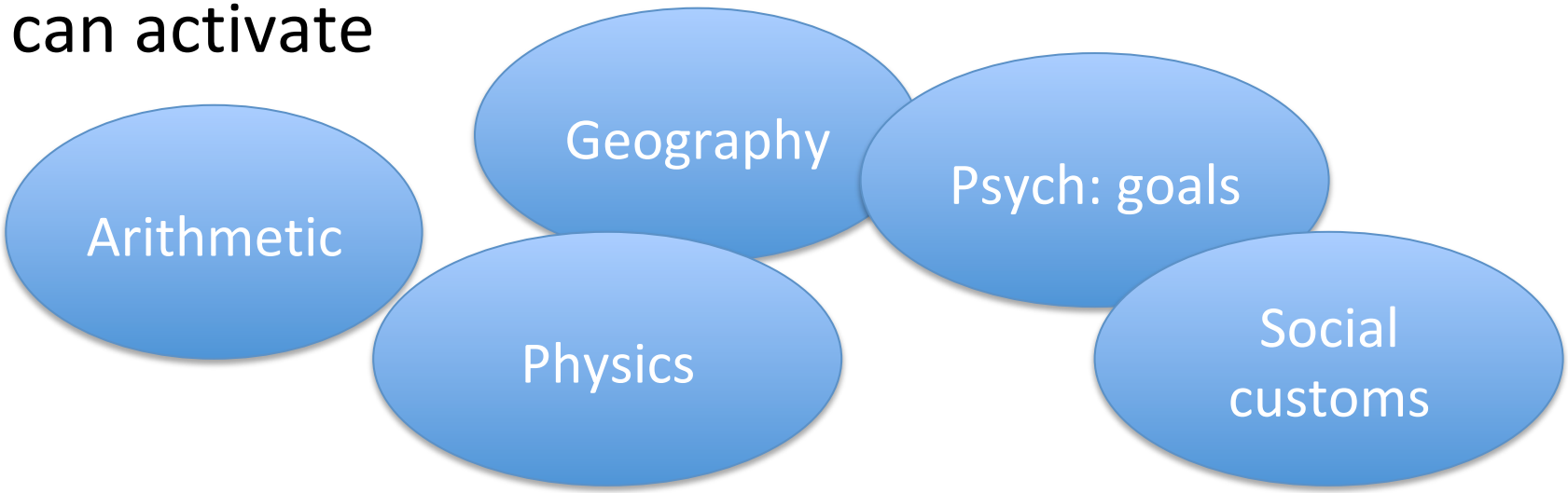
- Turkers and Sliding Window are good at simple matching questions
- Surprisingly, Stanford AR does not have a better performance on matching questions

Moving ahead on Matching QA

- Identify the types of more-complex matching, match composition/nesting, etc.
- Build datasets like RACE that
  - Come from the real world
  - Do not support matching on simple multi-fact presence
  - Show a real gap between human and system
- Evaluate **correctness** AND **composition** (require the answer trace plus its component factoids)

2. Making **computation** more complex: Inference of various kinds

- Define *N* self-contained standardized 'domain specialists' (KBs+reasoners) that any QA engine can activate



- At run-time, analyze the Q, build the A script, activate the specialists as needed, compute the A

Examples

- **What is the largest capital city south of Santiago de Chile?**
  - Geographic knowledge (lat-long, population)
  - Numerical ability (sorting, etc.)

- **Which of the leaders of the XYZ enterprise are well-liked, and why?**
  - Discovery of social role by actions
  - Sentiment judgments attached to actions

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 to determine parameters and need
  - Decompose the need into 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
2. Psych goals for sentiment justification
3. Social roles for group activity support

(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

- Impact/need:
  - Contradiction pairs from Wikipedia and Google News: 29% from numeric discrepancies (de Marneffe et al. ACL 2008)
  - Several Recognizing Textual Entailment datasets: numeric contradictions are 8.8% of contradictory pairs (Dagan et al. RTE 2006)

(Ravichander et al., 2019)

EQUATE

- Models of quantitative reasoning should:
  - Interpret quantities expressed in language
  - Perform basic arithmetic calculations
  - Justify quantitative claims by combining verbal and numeric reasoning
- Current inference datasets do not test this
- Our work EQUATE:
  - A **corpus** of numerical pairs with entailment labels
  - A **benchmark evaluation framework** to test model ability to perform quantitative reasoning for natural language inference, combining 9 previous approaches

(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)

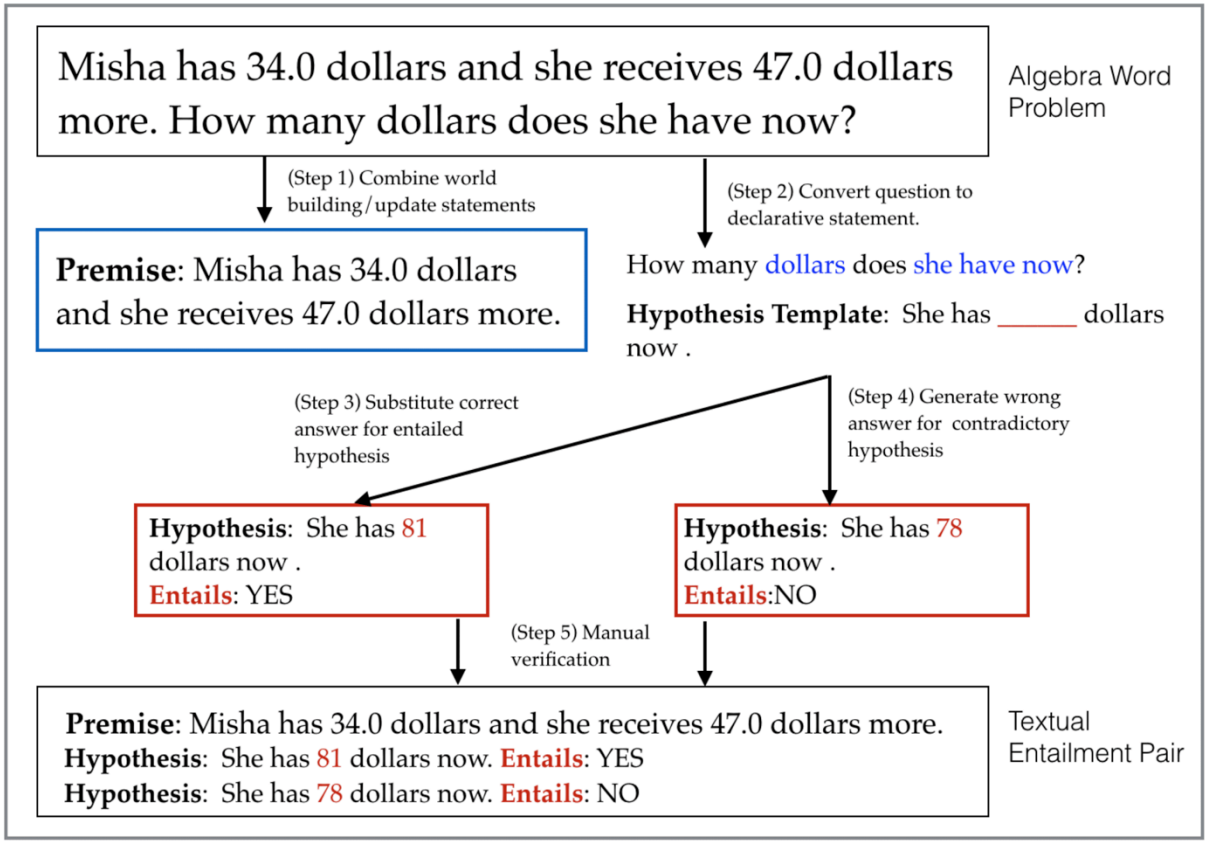
Baselines (SOTA methods)

- 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



- 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 goals and sentiment
  - Finding sentiment Holder, Topic+Facet, and Valence are relatively easy
  - Example: "I loved the hotel's price but the room was noisy" —> [price +] [room -]
- **Task: sentiment justification**: WHY does the Holder have that sentiment valence for that facet?
- Approach: Classify each clause into a list of human (psychological and social) goals
  - Input: Sentiment-bearing clause
  - Output: Sentiment valence label, facet, human goal(s) that justify sentiment valence

(Otani and Hovy, 2019)

Psych goals

- Technical approach: Automated classification into taxonomies of psych (not social) goals
  - Initial set: Maslow hierarchy (see Wikipedia)
  - Currently: about 110 human goals, identified and taxonomized (Talevich et al.)
- Domains: Reviews of Hotels, Cameras, Movies
- Data: Crowdsourcing training material; kappa agreement ≈ 0.55
- Results: traditional and neural classifiers do better on objects and poorer on events/complex things like movies

(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 Z2
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			

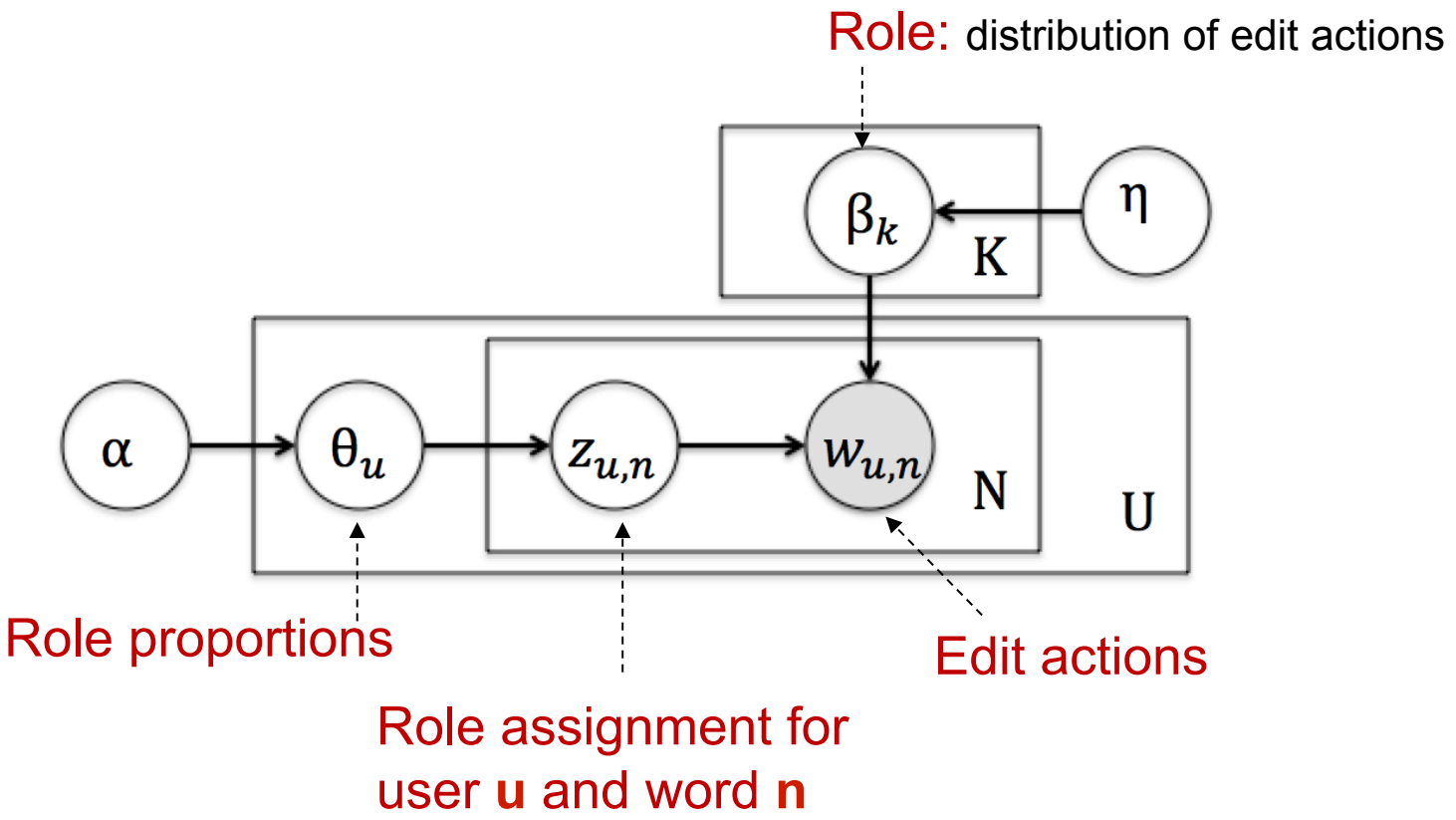
(Yang, Kraut, Hovy, 2019)

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, 2019)

Latent role model in Wikipedia



(Yang, Kraut, Hovy, 2019)

Roles

Information_insertion 0.4
Reference_insertion 0.2
….

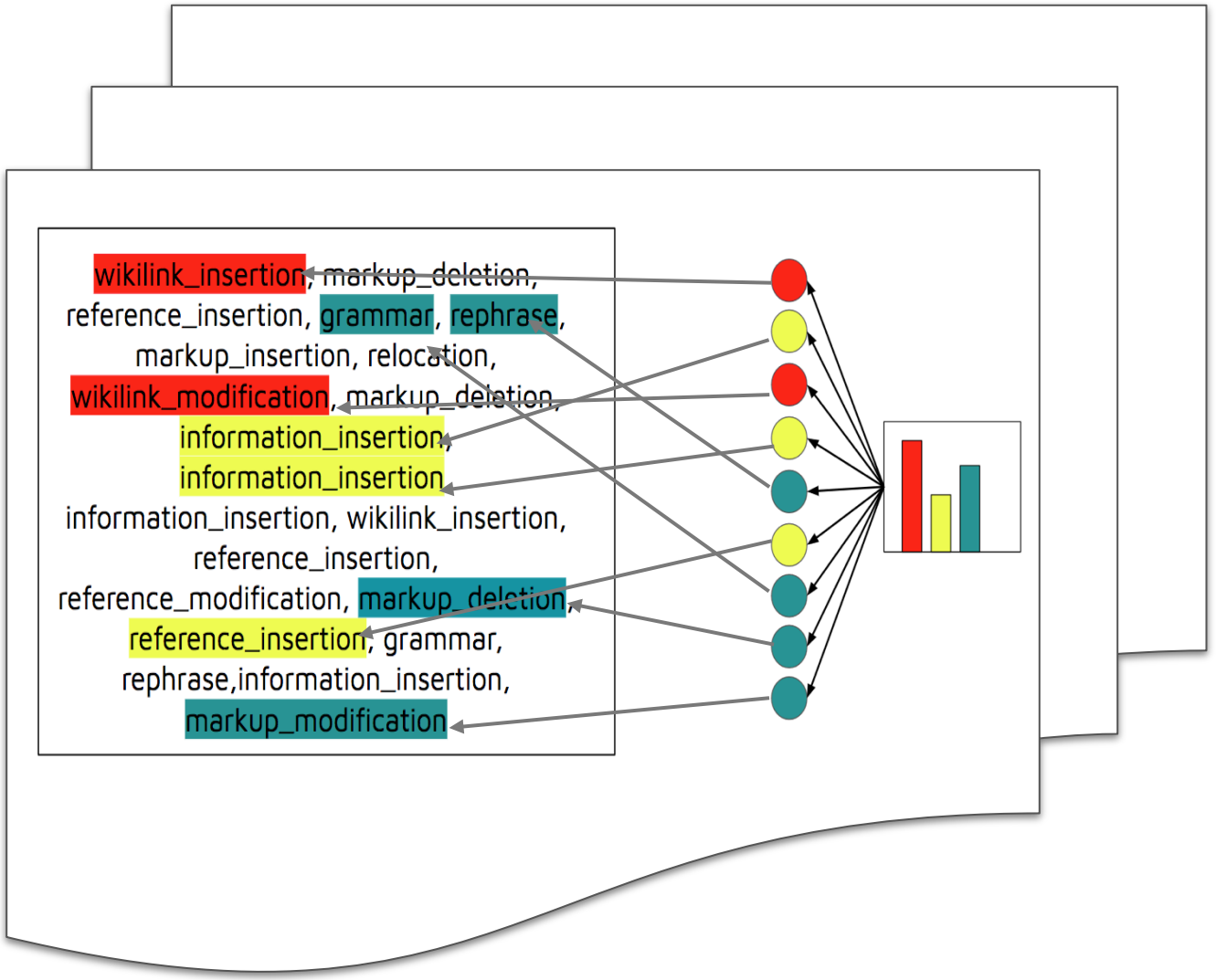
Grammar 0.2
Markup_deletion 0.1
Rephrase 0.1
….

Wikilink_insertion 0.2
Wikilink_deletion 0.1
….

User edit history

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

Role assignments



(Yang, Kraut, Hovy, 2019)

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?

Physics: Some Big Problems!

- *"Salt (Na^+Cl^-) is a white powder with a salty taste. As you can see, it is an ionic compound. You will see the powder dissolve when you put it into water."*
  - Does the formula Na^+Cl^- have a salty taste?
  - Is the powder the formula? Can you write a powder?
  - Does the taste dissolve? Or the whiteness?

- A lot of information is hidden, and a lot assumed:
  - Knowledge gaps : explicit links between one term and another
  - Omissions : missing (assumed known?) information

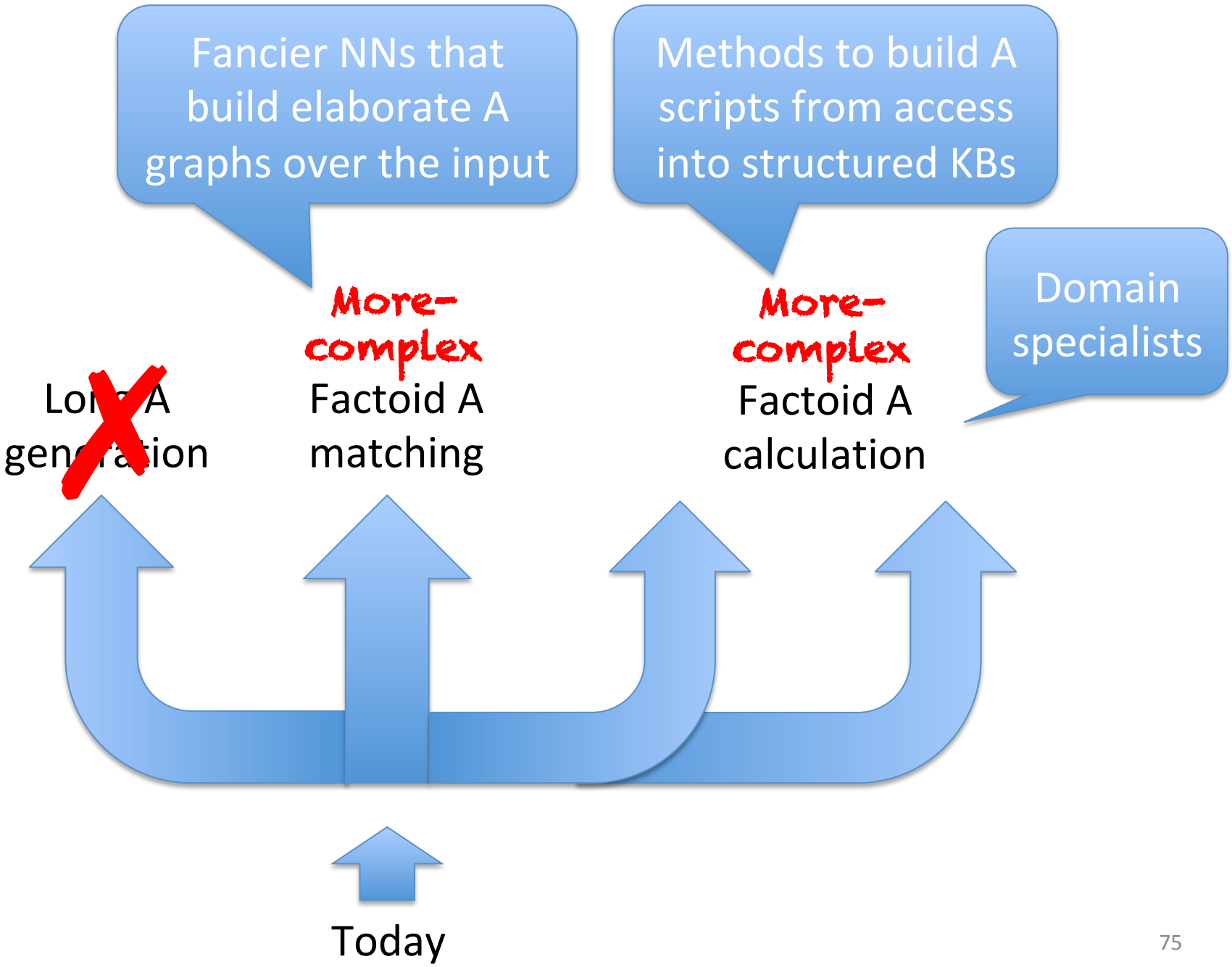
Language is full of what Peter Clark calls '*loosespeak*'

Where next with Calculation 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')


Fancier NNs that build elaborate A graphs over the input
Methods to build A scripts from access into structured KBs
Domain specialists
More-complex
More-complex
Long A generation
Factoid A matching
Factoid A calculation
Today

THANK YOU