Towards AI-Complete Question Answering : A Set of Prerequisite Toy Tasks

2022.09.08

Hyun Seop Yoon

Thesis Seminar

KOREA UNIVERSITY
Brain and Coginitive Engineering

Brain Signal Processing Laboratory
https://bspl.korea.ac.kr

Overview

- **What is Babl task?**
  - introduced in 2014 / presented in ICLR 2015
  - Argue for the usefulness of a set of proxy tasks that evaluate reading comprehension via question answering.
  - Aim to classify the tasks into skill sets, so that researchers can identify, rectify the failing of their systems.
  - (no task-specific engineering). Because the task are built to make a standard, also milestone for building AI model that understands human language.

> Performing well on all of them is a pre-requisite for any system aiming at understanding language and able to reason

Limitation in AI understanding

- Model with high capacity and representation power(CNNs,RNNs,LSTMs) and lots of data(supervised, semi-supervised, unsupervised) have resulted many breakthrough
- However the model and learning algorithms rely heavily on big data statistics, especially labeled data.
- ⇒ Reasoning is still limited, especially in Q&A systems

Embedding-based model of (Bordes et al. EMNLP14)

Q: What country was Slovakia?

A:austria, A:czech_republic

A: czechoslovakia

Approchments of synthetic data

- Though simple models and a lot of data trump more elaborate models based on less data(Halevy et al., 2009), but are far from being a model that truly understands text.(Yao et al., 2014), Berant et al. (2014)
- AI reasearch become stuck in *local minima problem,* which means the model is not heading to the ultimate goal of AI Model
- The synthetic task covers the ultimate goal: Evaluating performance of an agent in general dialogue, while dissolving into relatively simple tasks in AI field, called question answering(QA)
- Since The QA covers wide range of cognitive capability, enabling researcher to test different ability of learning algorithms, under a common framework

Success in Artificial Problem

- Machine Learning
  - Clustering(Two moons and friends)
  - XOR(Neural Network) Minsky & Papert, 1969; Rumelhart et al., 1985)
  - Regression, Classification(# in UCI) (Bache & Lichman, 2013)
- Artificial Intelligence
  - SHRLDU(Block world)(Terry Winogran, 1971)

    - Basic Nuance(pre, post relationship among words)

    - Combination of location, direction

Artificial task for learning for AI : QA

- It is crucial to build learning algorithm and training condition at the same speed

  QA-based strategies
  • Difficulty of definition of questions
    - Unambiguously answerable by adult human(or children)
    - Still require some thinking

  ⇒ No system has yet been clearly identified capabilities and limitation
  ⇒ No proposal of improvement and modifications

New learning algorithms = Creations of new training conditions are mandatory

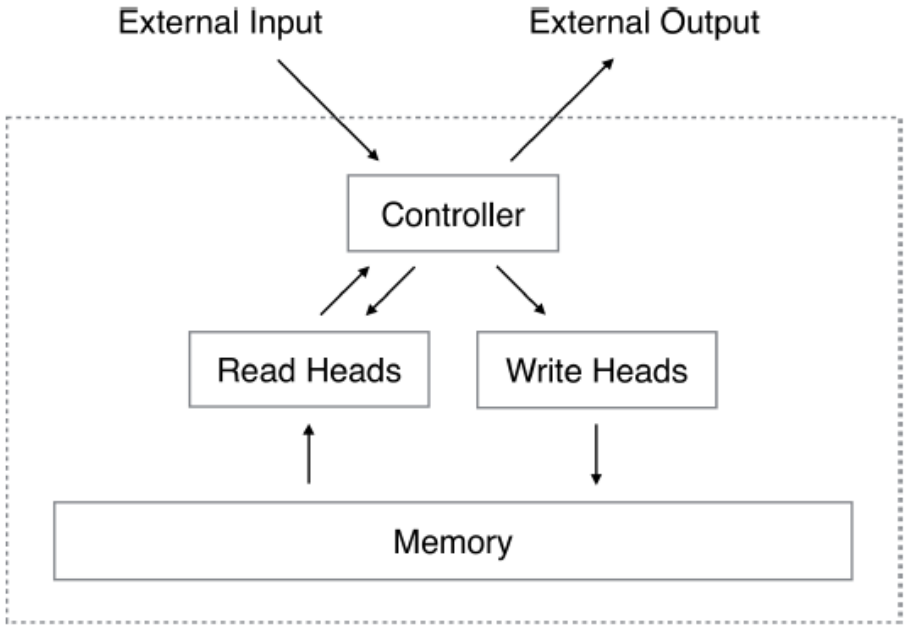
Drawback of Artificial Data and Looking forward

- Artificial data can have a probability of overfitting, since the task does not fully apply the real world.
- So to eventually scale up to real language
  - No model should be tailored for a task alone, nor for the tasks only.

  ⇒ The task is a sample guideline, which upgraded version should keep this in mind
  - Models should be able to learn new things incrementally and faster
  - The simulation should be parameterized to ramp up complexity, create more tasks.

Recent approach and Follow-up ideas

- Shapeset
  - Following steps → Improved to answer Questions



- Sequences
  - Basic tasks(Copying, Sorting, Associative recall, Dynamic n-grams)

1. Neural Turing Machine(Grave et al. 14)

→ Neural Network using explicit memory, compared to implicit memory in LSTM, RNN

2. Stack-augmented RNNs, Weston et al. (2014)

→ Stacked RNNs which enable to cope with long sequential data.

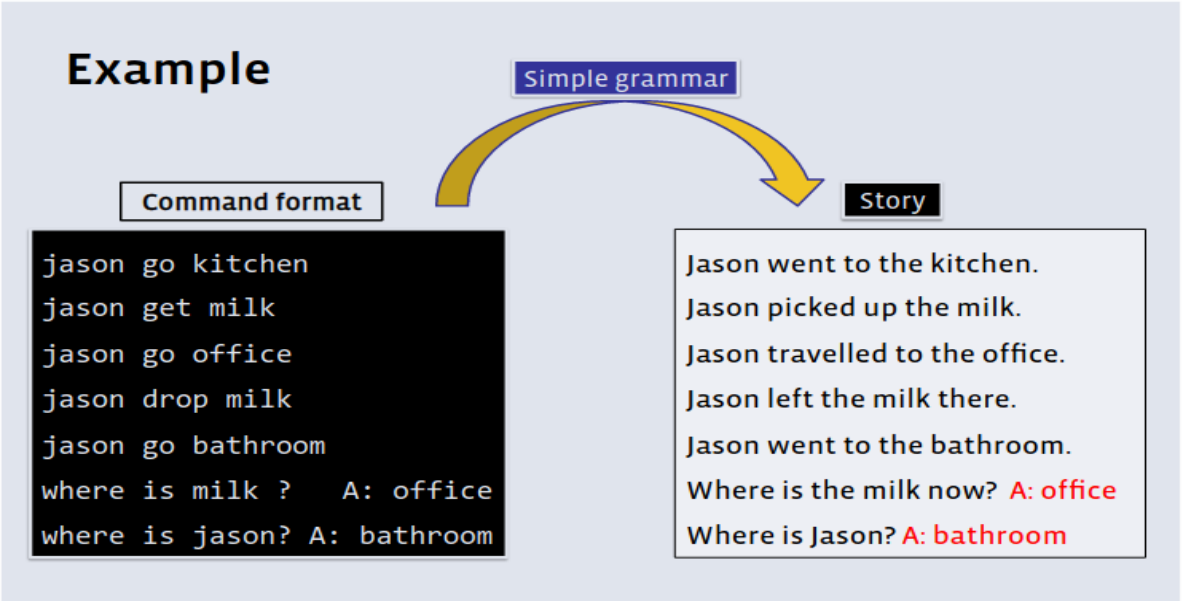
Ability - Tasks

- The goal is to categorize different kinds of questions into skills sets, which become tasks.
- While the analysis of performance: failure or success of a system on any of them can unequivocally provide feedback on its capabilities.
- Motivate new algorithm design that alleviate the weaknesses
- One task - One Ability each

Ability - Tasks

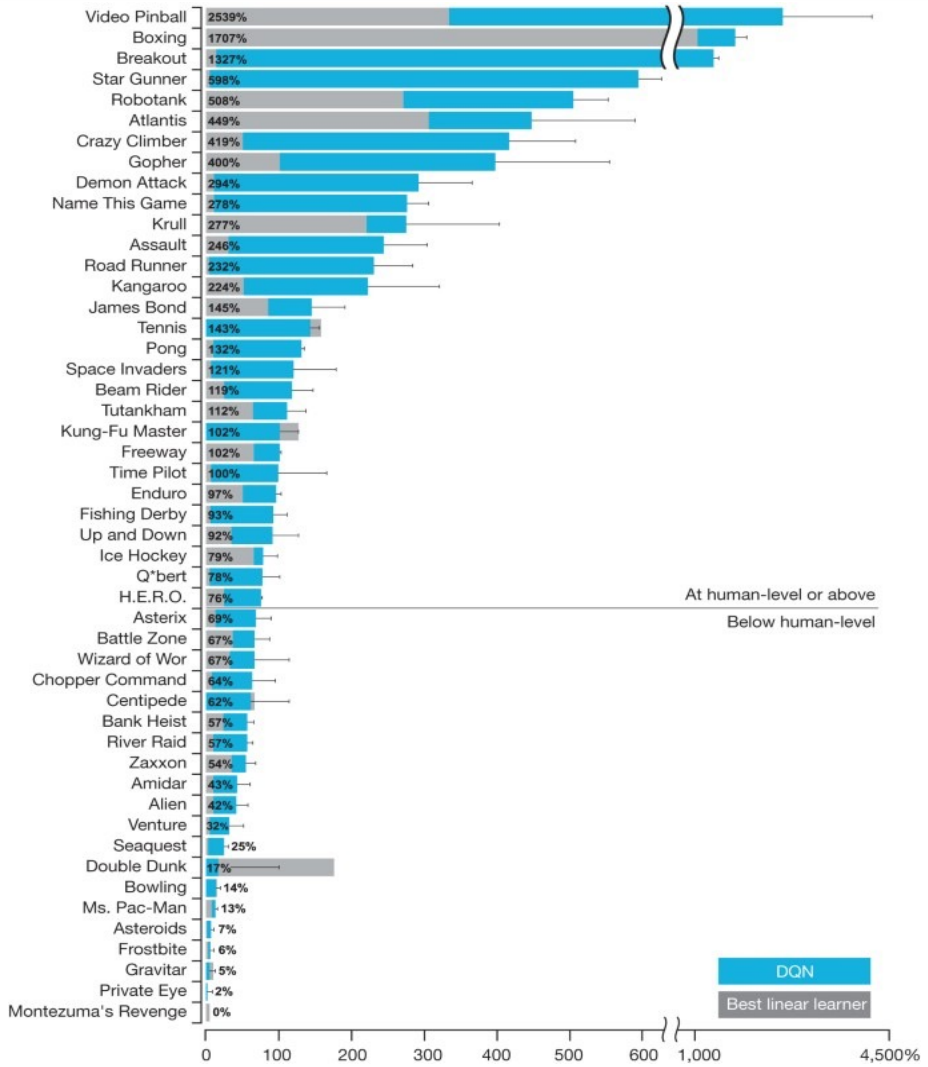
- Monitoring cognitive capability can be assumed by using command-answer system so called **“Text-Adventure-Games”**
- The system enables to **create a learning environments** about cognitive competence, which are closely-related to the meaning of ‘learning from scratch’
- The system creates a simulated world, like text-adventure games. each sentences are produced with a simple grammer. The advantages of following

•Difficulty/Complexity is controlled
•Training and Evaluation data are provided
•Evaluation through question answering is easy to overlook the process



http://www.thespermwhale.com/jaseweston/babi/abordes-ICLR.pdf

Why game?



Old games offer a great variety of controlled environments

https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf

Tasks

- Single supporting Facts

  - The task is to find out the single supporting fact in order to provide the answer.
  - Answer was previously given, and the size of fact cannot be bigger than two.
  - Typical question type is to find out **the location of Agent**

John travelled to the office.
Daniel travelled to the office.
Where is Daniel? office

- Two Supporting Facts

  - This is slightly harder task of single supporting facts.
  - Two supporting facts have to be chained in order to answer the question

John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? playground

- Three Supporting Facts

  - Finding three supporting detail is mandatory for model to derive answer.

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? office

Tasks

- Two Argument Relations: Subject vs Object

  - Solving problem about differentiating and recognize subject and object

  - Linguistically have problem among this.
  - Seems like this is a counter-example of bag-or words
  - Not a semantic approach of directional analysis

  → Relational expression are not 100% S-O structure. Able to catch Idiom about direction?

The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? office
What is the bedroom north of? bathroom

- Yes/No Questions

  - Simplest case, examining whether the model can revise answer about True, False question.

John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? no
Is Daniel in the bathroom? yes

- Counting

  - Evaluating ability of **simple counting**: Number of certain property

Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? two

Tasks

- Positional Reasoning
  - Quite similar task compare to T4, but more focused on predicate itself (e.g. A be to the right of the B)

The triangle is to the right of the blue square.
The red square is on top of the blue square.
The red sphere is to the right of the blue square.
Is the red sphere to the right of the blue square? yes
Is the red square to the left of the triangle? yes

- Reasoning about Size
  - This task is inspired by Winograd schema(Levesque AAAI11)
  - T3 and T6 are prerequisites.
  - Task about reasoning about Relative Size

The trophy doesn't fit in the brown suitcase because it's too big.
What is too big?
~~the trophy~~
the suitcase

cf. Winograd Schema Challenge
•Alternative to the Turing Test
•Pair of sentences that differ only in one or two words and that contain a referential ambiguity
•The ambiguity can be resolved in opposite directions in two sentences
•The answer is obvious to the human reader, but cannot easily be founded using selectional restrictions or statistical technique.

Tasks

- Path finding

  - This task gives location of each object, and required to **find the path** between locations: give directions sequences.
  - Have to find way that is the most effective, since the task requires **search**

The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? west, north
How do you go from office to bathroom? north, west

Simulation

- Simulated world is composed of entities of various types(locations, object, persons, etc)
- Simple automated grammer(Verb synonym)
- 150 words, 4 actors, 6 locations, and 3 objects per task
    - Entities have internal states: location, mental state of actors, properties

1. Location

   : Have to correspond with nearby surroundings

2. Actors

   : Pre-specified rules based on common sense control their behaviour
   : Actions enforce coherence in the simulation

3. Random Valid Actions

   : Each task has limitation of actions

Simulation control panel

- Symbols
  - Can the system switch to the other language?
  - And other simpler symbolic systems?
  - (added) Any better symbolic system that has function of symbol system? (e.g sudo code)

    ⇒ Used other language, shuffled English letters

- Memory
  - How far should one remember?
  - Is an external source of knowledge necessary?(Memory Network)

    ⇒ Added irrelevant facts, which assess maximum of memory, but exact number of data is not shown.
    ⇒ Able to use external resources(common sense) in order to solve the task

Results

- Linguistics
  - How is reasoning altered by ambiguities?
  - (added) How can model distinguish differences between ambiguities? (Semantic, Syntactic, Lexial, Anaphoric,Ellipsis)
  - Phase Embedding

    ⇒ T20 examines the most simple(sometimes cannot get credit from linguists) type of anaphoric task, but still cannot cover whole concept of coreference, which need to adapt a real coreference dataset.

    ⇒ T13 seems to cover the coreference when the pronoun can refer to multiple actors, but this task checks same linguistic ability compared with T20. Seems to analyze more about how NLP model cope with linguistic concept.

    ⇒ T14 mentions that the task examine understanding of time expression, but the concept ‘time’ should be more precise. The total concept should be an ‘aspect’ and special time period should check under time schema, especially boundedness, durational/punctual, telic/atelic,, and so on

Results

- Reasoning
  - How many facts should be chained together?
  - T8 is related to a QA task related to database search, which can be considered as a ability to make a query

e.g

- Intersection: Who is in the park carrying food

- Union: Who has milk or cookies

- Set difference: Who is in the park apart from Bill?

Dataset

- format

```
ID text
ID text
ID text
ID question[tab]answer[tab]supporting_fact ID.
...
```

ID : Number of the sentence
text: Storyline text
supporting_fact ID: ID that the answer clue be located

- example: Task 13: Compound Coreference

```
1 Sandra travelled to the kitchen.
2 Sandra travelled to the hallway.
3 Mary went to the bathroom.
4 Sandra moved to the garden.
5 Where is Sandra?	Garden	4
```

Sentence 1 to 4 are storyline texts, while sentence 5 is the question, answer, supporting_fact ID.

The correct answer of question is located between two [tab],\t .
The researcher can infer the clue of the question, by peeking supporting_fact_ID.

The storyline text be usually made up with 15~20 sentences, often inserted with questions.

Noticable fact is that the question answer clue does not locate right before the question, but the whole storyline.

Question answering machine has to consider whole text in order to make a right answer.

https://www.kaggle.com/datasets/roblexnana/the-babi-tasks-for-nlp-qa-system

Dataset

- format

```
ID text
ID text
ID text
ID question[tab]answer[tab]supporting_fact ID.
...
```

ID : Number of the sentence
text: Storyline text
supporting_fact ID: ID that the answer clue be located

- Code

https://github.com/BSPL-KU/bspl-ku.github.io/blob/d6f7ae43aa8d8ca7060f0126b0adaec75e7feb34/babi_preprocessing.ipynb
https://github.com/BSPL-KU/bspl-ku.github.io/blob/d6f7ae43aa8d8ca7060f0126b0adaec75e7feb34/babi_preprocessing.ipynb

- example: Task 13: Compound Coreference

```
1 Sandra travelled to the kitchen.
2 Sandra travelled to the hallway.
3 Mary went to the bathroom.
4 Sandra moved to the garden.
5 Where is Sandra?    Garden    4
```

Sentence 1 to 4 are storyline texts, while sentence 5 is the question, answer, supporting_fact ID.

The correct answer of question is located between two [tab],\t .
The researcher can infer the clue of the question, by peeking supporting_fact_ID.

The storyline text be usually made up with 15~20 sentences, often inserted with questions.

Noticable fact is that the question answer clue does not locate right before the question, but the whole storyline.

Question answering machine has to consider whole text in order to make a right answer.

https://www.kaggle.com/datasets/roblexnana/the-babi-tasks-for-nlp-qa-system

Experiment

- Comparing methods via task solving
  - N-gram classifier (weak)
  - LSTMs (weak)
  - MenNN (strong)
  - extensions of MN(AM,NG,NONLINEAR) (strong)
  - structured SVM (external resources)

•Weakly supervised models → given question answer pairs at training time
•Strongly supervised models → give a set of supporting facts
•Methods in the last external resources track can use labeled data from other sources

Methods

- N-gram classfier (Richardson et al. (2013)
  - SLM, if the sentence is k1,k2,k3 and the word to predict is w,MenNN (strong)
  - Bag-of-N-grams for all sentences in the story that share at least one word with the question
  - Learn a linear classifier to predict the answer

$$P(w|k1,k2,k3) = \frac{count(k1,k2,k3,w)}{count(k1,k2,k3)}$$

- MenNN
  - Controller NN performs inference over stored previous statements
  - I: (input feature map) – convert input sentence x to an internal feature representation I(x).*
  - G: (generalization) – update the current memory state m given the new input: mi = G(mi,I(x),m), ∀i.
  - O: (output feature map) – compute output o given the new input and the memory: o =O(I (x), m).
  - R: (response) – finally, decode output features o to give the final textual response to the user: r = R(o).

Methods

- MenNN
  - finding the first supporting fact with match score with the question
  - find the second supporting fact with both question and first fact
  - Matching function consists of → Mapping the bag-of words for the question and fact into embedding by summing word embedding

$$o_1 = O_1(x, \mathbf{m}) = \arg\max_{i=1,\ldots,N} s_O(x, \mathbf{m}_i)$$

$$o_2 = O_2(q, \mathbf{m}) = \arg\max_{i=1,\ldots,N} s_O([x, \mathbf{m}_{o_1}], \mathbf{m}_i)$$

$$s(x, y) = \Phi_x(x)^\top U^\top U \Phi_y(y).$$

Methods

- MenNN
  - U = n * D matrix where D is the number of features and n is the embedding dimension
  - PHI function map the original text to D-dimensional feature space
  - Response function consider RNNs, setup limit response to be a single word by ranking them.
  - Response function consider RNNs, setup limit response to be a single word by ranking them.

$$s(x, y) = \Phi_x(x)^\top U^\top U \Phi_y(y).$$

$$r = R(q, w) = \mathbf{argmax}_{w \in W} \; s_R([x, \mathbf{m}_{o_1}, \mathbf{m}_{o_2}], w)$$

Methods

- Adaptive memories
  - Infer more than 2, model predict inference or “stop” class(scoring a special fact m0)

```
i=1
oi =O(x,m)
while oi ≠ mØ do

i←i+1

oi = O([x,mo1,...,moi−1],m)
end while
```

- N-grams
  - Bag of 3-grams

Methods

- Nonlinearity
  - NN approach called multilinear map in order to solve problem of data size while using N-grams
  - 2 layer of NN with tanh nonlinearity
  - i is the position of the word in a sentence of length l, and for each position employ a n*n matrix P_p(i,l)
  - Nonlinear embedding, in order to assess the performance of nonlinear map that fail modelling
  - W is a n × n matrix.

$$s(q,d) = E(q) \cdot E(d); \quad E(x) = tanh(\sum_{i=1,\ldots,l} P_{p(i,l)} \Phi_x(x_i)^\top U)$$

$$p(i,l) = \lceil (iP_{sz})/l) \rceil$$

$$E(x) = tanh(W tanh(\Phi_x(x)^\top U)).$$

Dashboard

	Weakly Supervised		Uses External Resources	Strong Supervision (using supporting facts)						
TASK	N-gram Classifier	LSTM	Structured SVM *COREF+SRL features*	MemNN Weston et al. (2014)	MemNN *ADAPTIVE MEMORY*	MemNN *AM + N-GRAMS*	MemNN *AM + NONLINEAR*	MemNN *AM + NG + NL*	No. of ex. req. ≥ 95	MultiTask Training
1 - Single Supporting Fact	36	50	99	100	100	100	100	100	250 ex.	100
2 - Two Supporting Facts	2	20	74	100	100	100	100	100	500 ex.	100
3 - Three Supporting Facts	7	20	17	20	**100**	**99**	**100**	**100**	500 ex.	**98**
4 - Two Arg. Relations	50	61	98	71	69	**100**	73	**100**	500 ex.	80
5 - Three Arg. Relations	20	70	83	83	83	86	86	**98**	1000 ex.	**99**
6 - Yes/No Questions	49	48	99	47	52	53	**100**	**100**	500 ex.	**100**
7 - Counting	52	49	69	68	78	86	83	85	FAIL	86
8 - Lists/Sets	40	45	70	77	90	88	94	91	FAIL	93
9 - Simple Negation	62	64	100	65	71	63	**100**	**100**	500 ex.	**100**
10 - Indefinite Knowledge	45	44	99	59	57	54	**97**	**98**	1000 ex.	**98**
11 - Basic Coreference	29	72	100	100	100	100	100	100	250 ex.	100
12 - Conjunction	9	74	96	100	100	100	100	100	250 ex.	100
13 - Compound Coref.	26	94	99	100	100	100	100	100	250 ex.	100
14 - Time Reasoning	19	27	99	99	100	99	100	99	500 ex.	99
15 - Basic Deduction	20	21	96	74	73	**100**	77	**100**	100 ex.	**100**
16 - Basic Induction	43	23	24	27	**100**	**100**	**100**	**100**	100 ex.	94
17 - Positional Reasoning	46	51	61	54	46	49	57	65	FAIL	72
18 - Size Reasoning	52	52	62	57	50	74	54	**95**	1000 ex.	93
19 - Path Finding	0	8	49	0	9	3	15	36	FAIL	19
20 - Agent's Motivations	76	91	95	100	100	100	100	100	250 ex.	100
Mean Performance	34	49	79	75	79	83	87	93		92

- Each task 1000 questions for training and testing each, 95% above accuracy score is considered as "PASS"
- Highlighted numbers indicate tasks where extension of MenNN pass while origininal MemNN don't
- Green colour indicate amount of training data for each task to pass, FAIL means need more than 1000
- Multitask Training column gives accuracy of training all data at once

Result

- Standard MemNN outperforms N-gram and LSTM, but still fail at some tasks(because of insufficient modeling power)
- AM outperforms in T3, T16, T8, T19 → use AM in combination with subsequent model
- N-gram and SVM itself does not show any noticeable result(external resource does not stand out its effect)
- AM + NG + NL MemNN shows desirable to perform well on task, while using fewist number of examples, but still lack general search algorithms

Discussion

- Prerequisite set
  - If any learner fails on these tasks, they will fail on real-world tasks too.

  → Is it still necessary condition?

- Flexible framework
  - Simulation-based approach provide flexibility to detect and combine patterns in symbolic sequences(by removing lexical variability and so on)

- Testing learning methods
  - The tasks are not a substitue for real data but a complement which are designed as a test-bed for learning methods

Impact

- Has influence on development of MemN2N, Dynamic Memory Network, Neural Reasoner