Adding semantic robustness to dialog agents

Felipe Salvatore
https://felipessalvatore.github.io/

July 6, 2018

IME-USP: Instituto de Matemática e Estatística - Universidade de São Paulo

Research problem

- Create a set of tasks that incorporate logic reasoning to boost performance of the current dialog agents.
- Perform a stress test in the existing *neural network based end-to-end dialog systems*.
- Integrate linguistic reasoning with visual references to create a new set of visual question answering (VQA) tasks.
- Define new models to achieve better results in the tasks proposed above.

Background

Neural network based language model

We call language model a probability distribution over sequences of tokens in a natural language.

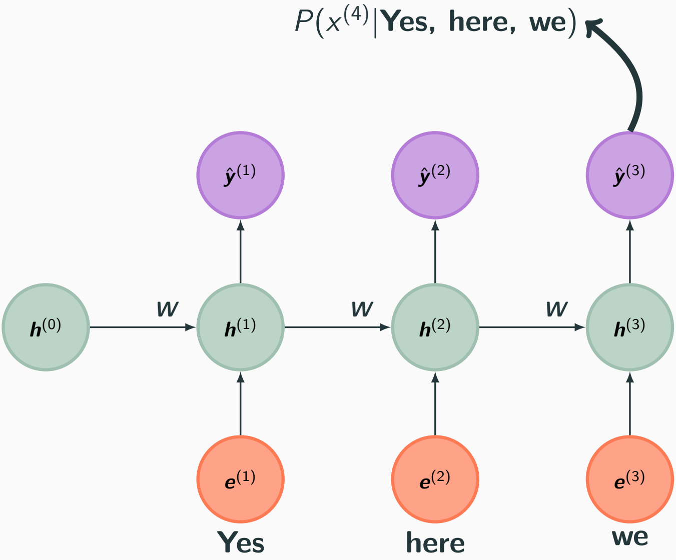
$$P(x_1, x_2, x_3, x_4) = p \tag{1}$$

Since [7], we use a Recurrent Neural Network (RNN) to estimate the probability distribution

$$P(x_n = \text{word}_{j^*} | x_1, \ldots, x_{n-1}) \tag{2}$$

for any $(n-1)$-sequence of words $x_1, \ldots, x_{n-1}$.

Neural network based language model



GRU: Gated Recurrent Units

$$\widetilde{\boldsymbol{h}}^{(t)} = tahn(\boldsymbol{W}(\boldsymbol{h}^{(t-1)} \odot \boldsymbol{r}^{(t)}) + \boldsymbol{U}\boldsymbol{x}^{(t)} + \boldsymbol{b}) \tag{3}$$

$$\boldsymbol{r}^{(t)} = \sigma(\boldsymbol{W}_r\boldsymbol{h}^{(t-1)} + \boldsymbol{U}_r\boldsymbol{x}^{(t)} + \boldsymbol{b}_r) \tag{4}$$

$$\boldsymbol{u}^{(t)} = \sigma(\boldsymbol{W}_u\boldsymbol{h}^{(t-1)} + \boldsymbol{U}_u\boldsymbol{x}^{(t)} + \boldsymbol{b}_u) \tag{5}$$

$$\boldsymbol{h}^{(t)} = \boldsymbol{u}^{(t)} \odot \widetilde{\boldsymbol{h}}^{(t)} + (1 - \boldsymbol{u}^{(t)}) \odot \boldsymbol{h}^{(t-1)} \tag{6}$$

LSTM: Long Short Term Memory

$$\boldsymbol{f}^{(t)} = \sigma(\boldsymbol{W}_f \boldsymbol{h}^{(t-1)} + \boldsymbol{U}_f \boldsymbol{x}^{(t)} + \boldsymbol{b}_f) \tag{7}$$

$$\boldsymbol{i}^{(t)} = \sigma(\boldsymbol{W}_i \boldsymbol{h}^{(t-1)} + \boldsymbol{U}_i \boldsymbol{x}^{(t)} + \boldsymbol{b}_i) \tag{8}$$

$$\boldsymbol{o}^{(t)} = \sigma(\boldsymbol{W}_o \boldsymbol{h}^{(t-1)} + \boldsymbol{U}_o \boldsymbol{x}^{(t)} + \boldsymbol{b}_o) \tag{9}$$

$$\tilde{\boldsymbol{c}}^{(t)} = tahn(\boldsymbol{W}\boldsymbol{h}^{(t-1)} + \boldsymbol{U}\boldsymbol{x}^{(t)} + \boldsymbol{b}) \tag{10}$$

$$\boldsymbol{c}^{(t)} = \boldsymbol{f}^{(t)} \odot \boldsymbol{c}^{(t-1)} + \boldsymbol{i}^{(t)} \odot \tilde{\boldsymbol{c}}^{(t)} \tag{11}$$

$$\boldsymbol{h}^{(t)} = \boldsymbol{o}^{(t)} \odot tanh(\boldsymbol{c}^{(t)}) \tag{12}$$

Sequence-to-sequence

- $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, source sentence
- $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(m)}$, target sentence
- f_{enc} (the *encoder*), a RNN
- f_{dec} (the *encoder*), a language model

$$\boldsymbol{s} = f_{enc}(\boldsymbol{x}^{(n)}, \boldsymbol{h}^{(n-1)}) \tag{13}$$

$$\tilde{\boldsymbol{h}}^{(t)} = f_{dec}(\boldsymbol{y}^{(t)}, \tilde{\boldsymbol{h}}^{(t-1)}) \tag{14}$$

$$p(y_t | y_1, \ldots, y_{t-1}, x_1, \ldots, x_n) = softmax(\boldsymbol{W}_s \tilde{\boldsymbol{h}}^{(t)} + \boldsymbol{b}_s) \tag{15}$$

Attention

$$\boldsymbol{a}_{ts} = \frac{exp(score(\tilde{\boldsymbol{h}}^{(t)}, \boldsymbol{h}^{(s)}))}{\sum_j exp(score(\tilde{\boldsymbol{h}}^{(t)}, \boldsymbol{h}^{(j)}))} \tag{16}$$

$$score(\tilde{\boldsymbol{h}}^{(t)}, \boldsymbol{h}^{(s)}) = \begin{cases} \tilde{\boldsymbol{h}}^{(t)}.^{\top} \boldsymbol{h}^{(s)} \\ \tilde{\boldsymbol{h}}^{(t)}.^{\top} \boldsymbol{W}_a \boldsymbol{h}^{(s)} \\ \boldsymbol{v}_a^{\top} tahn(\boldsymbol{W}_a[\tilde{\boldsymbol{h}}^{(t)}; \boldsymbol{h}^{(s)}]) \end{cases} \tag{17}$$

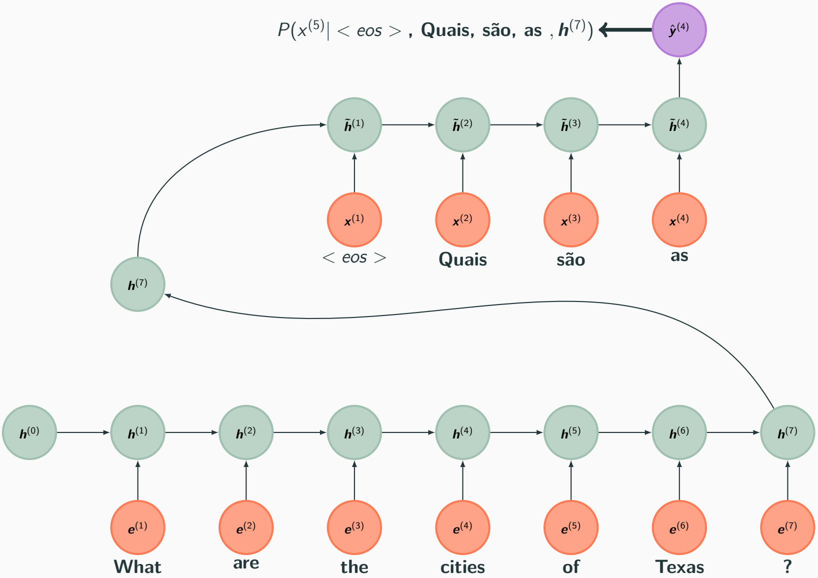
$$\boldsymbol{c}^{(t)} = \sum_s \boldsymbol{a}_{ts} \boldsymbol{h}^{(s)} \tag{18}$$

$$\tilde{\boldsymbol{h}}_{out}^{(t)} = tahn(\boldsymbol{W}_c[\boldsymbol{c}^{(t)}; \boldsymbol{h}^{(t)}]) \tag{19}$$

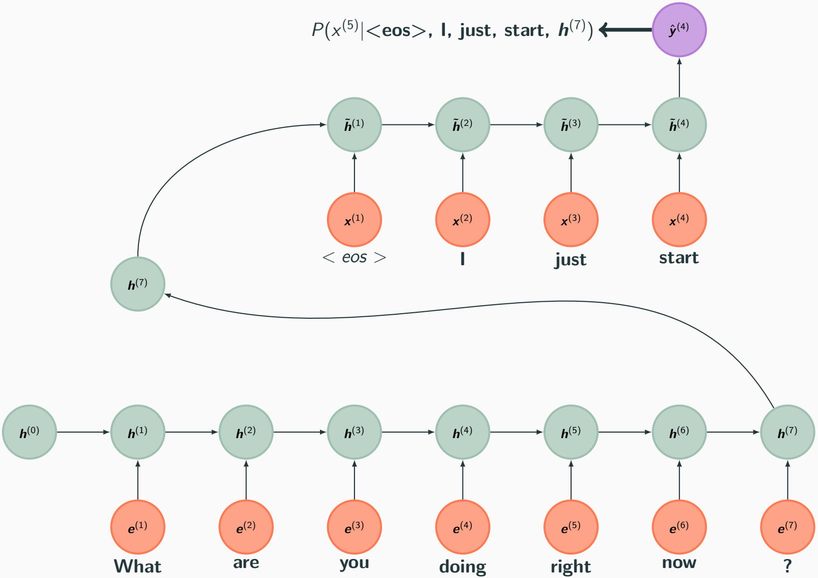
$$p(y_t | y_1, \ldots, y_{t-1}, x_1, \ldots, x_n) = softmax(\boldsymbol{W}_s \tilde{\boldsymbol{h}}_{out}^{(t)} + \boldsymbol{b}_s) \tag{20}$$

Neural network based dialog systems

Seq2seq applied to translation



Seq2seq applied to dialog [8]



MemNN

- $U_1, \ldots, U_n$ context
- q question
- a answer

We have $k = 1, \ldots, K$ memory layers:

- $\{\boldsymbol{m}^{(k)}{}_i\}$, memory vectors
- $\boldsymbol{u}^{(k)}$, input vector
- $\boldsymbol{p}^{(k)}$, match between $\boldsymbol{u}^{(k)}$ and each $\boldsymbol{m}_i^{(k)}$
- $\{\boldsymbol{c}^{(k)}{}_i\}$, another representation of the context $U_1, ..., U_n$
- $\boldsymbol{o}^{(k)}$, output.
- $\hat{\boldsymbol{a}} = softmax(\boldsymbol{W}(\boldsymbol{o}^K))$, candidate answer

How to evaluate dialogs?

Human evaluation [5]

In the first trial, we asked the following questions to the users, for each response:

1. How appropriate is the response overall? (overall, scale of 1-5)
2. How on-topic is the response? (topicality, scale of 1-5)
3. How specific is the response to some context? (specificity, scale of 1-5)
4. How much background information is required to understand the context? (background, scale of 1-5)

1. Adequacy: the meaning equivalence between the generated and control sentence.
2. Fluency: the syntactic correctness of the generated sequence.
3. Readability: efficacy of the generated sentence in a particular context.

BLEU (bilingual evaluation understudy)

$$P_n = \frac{\text{number of } n\text{-grams in both } \hat{y} \text{ and } y}{\text{number of } n\text{-grams appearing in } \hat{y}} \tag{21}$$

$$BP = \begin{cases} 1 & \text{if } len(\hat{y}) > len(y) \\ \exp\left(1 - \frac{len(y)}{len(\hat{y})}\right) & \text{otherwise} \end{cases} \tag{22}$$

$$BLEU = BP \; \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log P_n\right) \tag{23}$$

METEOR (Metric for Evaluation of Translation with Explicit ORdering)

$$P = \frac{\text{number of unigrams in both } \hat{y} \text{ and } y}{\text{number of unigrams appearing in } \hat{y}} \tag{24}$$

$$R = \frac{\text{number of unigrams in both } \hat{y} \text{ and } y}{\text{number of unigrams appearing in } y} \tag{25}$$

$$F_{mean} = \frac{10PR}{R + 9P} \tag{26}$$

$$METEOR = F_{mean}(1 - penalty) \tag{27}$$

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

$$P_{lcs} = \frac{lcs(\hat{y}, y)}{len(\hat{y})} \tag{28}$$

$$R_{lcs} = \frac{lcs(\hat{y}, y)}{len(y)} \tag{29}$$

$$ROUGE_L = \frac{(1 + \beta^2) P_{lcs} R_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{30}$$

where β is usually set to favour recal ($\beta = 1.2$).

Problems [4]

metric	Spearman	*p*-value	Pearson	*p*-value
BLEU	0.34	< 0.01	0.14	0.17
METEOR	0.19	0.06	0.19	0.05
ROUGE	0.12	0.22	0.1	0.34

Table 1: Correlation between automatic metrics and human judgments based on dialog generated on Twitter

metric	Spearman	*p*-value	Pearson	*p*-value
BLEU	0.12	0.23	0.11	0.26
METEOR	0.06	0.53	0.14	0.16
ROUGE	0.05	0.59	0.06	0.53

Table 2: Correlation between automatic metrics and human judgments based on dialog generated on Ubuntu

Creating simplified tasks as tests

bAbI [9]

One solution is to create a set of QA synthetic tasks to test different capabilities of a dialog agent.

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
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? A:office

Task 4: Two Argument Relations
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? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations
Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions
John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? A:no
Is Daniel in the bathroom? A:yes

ParlAI

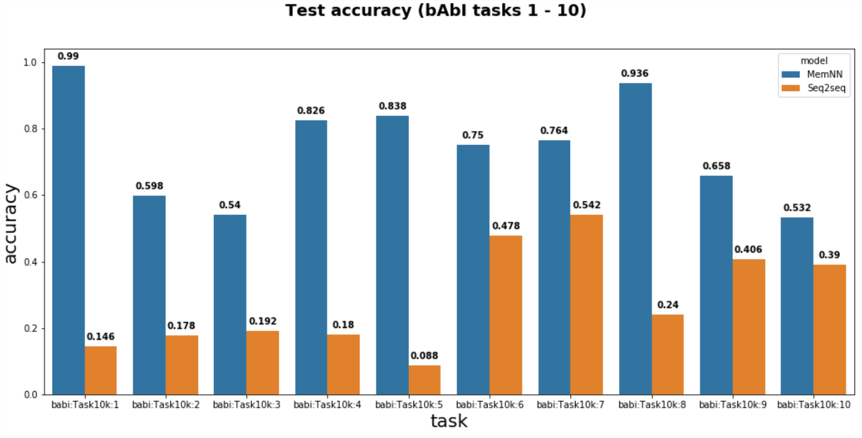
https://github.com/facebookresearch/ParlAI

"ParlAI (pronounced 'par-lay') is a framework for dialog AI research, implemented in Python.

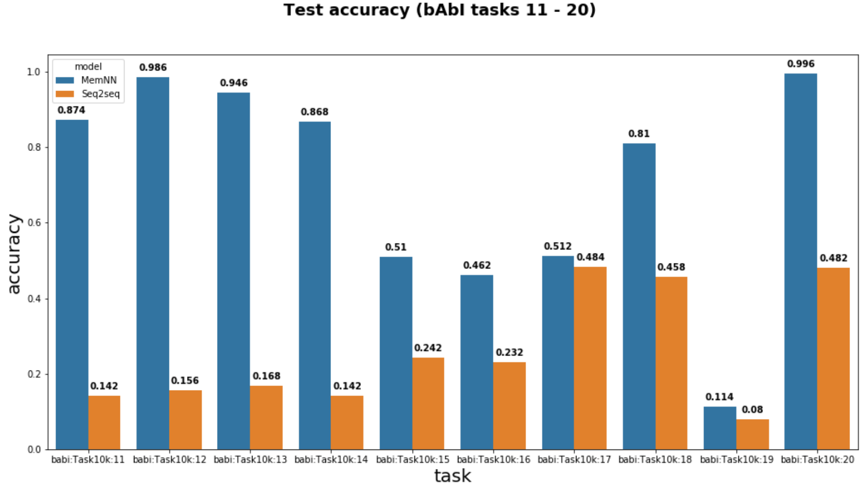
Its goal is to provide researchers:

- a unified framework for sharing, training and testing dialog models
- many popular datasets available all in one place, with the ability to multi-task over them
- seamless integration of Amazon Mechanical Turk for data collection and human evaluation"

Sanity check experiments



Sanity check experiments



Entailment-QA

bAbI: task 15

Basic Deduction

Task 15: Basic Deduction
Sheep are afraid of wolves.
Cats are afraid of dogs.
Mice are afraid of cats.
Gertrude is a sheep.
What is Gertrude afraid of? A:wolves

$$
\begin{gathered}
P^1 \text{ are afraid of } Q^1 \\
P^2 \text{ are afraid of } Q^2 \\
P^3 \text{ are afraid of } Q^3 \\
P^4 \text{ are afraid of } Q^4 \\
c^1 \text{ is a } P^1 \\
c^2 \text{ is a } P^2 \\
c^3 \text{ is a } P^3 \\
c^4 \text{ is a } P^4 \\
\text{What is } c^j \text{ afraid of? A: } Q^j
\end{gathered}
$$

bAbI: task 16

Basic Induction

Task 16: Basic Induction
Lily is a swan.
Lily is white.
Bernhard is green.
Greg is a swan.
What color is Greg? A:white

$$c^1 \text{ is a } P^1$$
$$c^1 \text{ is } C^1$$
$$c^2 \text{ is a } P^2$$
$$c^2 \text{ is } C^2$$
$$c^3 \text{ is a } P^3$$
$$c^3 \text{ is } C^3$$
$$c^4 \text{ is a } P^4$$
$$c^4 \text{ is } C^4$$
$$c \text{ is a } P^j$$
$$\text{What color is } c\text{? A: } C^j$$

Entailment-QA

1. **Boolean Connectives**
2. **First-Order Quantifiers**
3. **Synonymy**
4. **Antinomy**
5. **Hypernymy**
6. **Active/Passive voice**

Entailment-QA: task 1

- Entailment (s_1 implies s_2)
  - $\underbrace{P^1 a^1 \wedge \cdots \wedge P^n a^n}_{s_1}, \underbrace{P^j a^j}_{s_2}$
  - $\underbrace{P^j a^j}_{s_1}, \underbrace{P^1 a^1 \vee \cdots \vee P^n a^n}_{s_2}$
  - $\underbrace{Pa}_{s_1}, \underbrace{\neg\neg Pa}_{s_2}$

- Not entailment (s_1 does not imply s_2)
  - $\underbrace{P^j a^j}_{s_1}, \underbrace{P^1 a^1 \wedge \cdots \wedge P^n a^n}_{s_2}$
  - $\underbrace{P^1 a^1 \vee \cdots \vee P^n a^n}_{s_1}, \underbrace{P^j a^j}_{s_2}$
  - $\underbrace{Pa}_{s_1}, \underbrace{\neg Pa}_{s_2}$

Entailment-QA: task 1

Ashley is fit

Ashley is not fit

The first sentence implies the second sentence? A: no

Avery is nice and Avery is obedient

Avery is nice

The first sentence implies the second sentence? A: yes

Elbert is handsome or Elbert is long

Elbert is handsome

The first sentence implies the second sentence? A: no

Entailment-QA: task 2

- Entailment
  - $\forall x Px$, Pa
  - $Pa, \exists x Px$
- Contradiction
  - $\forall x Px, \neg Pa$
  - $\forall x Px, \exists x \neg Px$
- Neutral
  - Pa, Qa
  - $\forall x Px, \neg Qa$

Entailment-QA: task 2

Every person is lively

Belden is lively

What is the semantic relation? A: entailment

Every person is short

There is one person that is not short

What is the semantic relation? A: contradiction

Every person is beautiful

Abilene is not blue

What is the semantic relation? A: neutral

Entailment-QA: task proxy

SICK (Sentences Involving Compositional Knowledge) [6]

Relatedness score	Example
1.6	A: "*A man is jumping into an empty pool*" B: "*There is no biker jumping in the air*"
2.9	A: "*Two children are lying in the snow and are making snow angels*" B: "*Two angels are making snow on the lying children*"
3.6	A: "*The young boys are playing outdoors and the man is smiling nearby*" B: "*There is no boy playing outdoors and there is no man smiling*"
4.9	A: "*A person in a black jacket is doing tricks on a motorbike*" B: "*A man in a black jacket is doing tricks on a motorbike*"

Table 1: Examples of sentence pairs with their gold relatedness scores (on a 5-point rating scale).

Entailment label	Example
ENTAILMENT	A: "*Two teams are competing in a football match*" B: "*Two groups of people are playing football*"
CONTRADICTION	A: "*The brown horse is near a red barrel at the rodeo*" B: "*The brown horse is far from a red barrel at the rodeo*"
NEUTRAL	A: "*A man in a black jacket is doing tricks on a motorbike*" B: "*A person is riding the bicycle on one wheel*"

Table 2: Examples of sentence pairs with their gold entailment labels.

Entailment-QA: task proxy

There is no dog leaping in the air

A dog is leaping high in the air and another is watching

What is the semantic relation? A: contradiction

A man is exercising

A baby is laughing

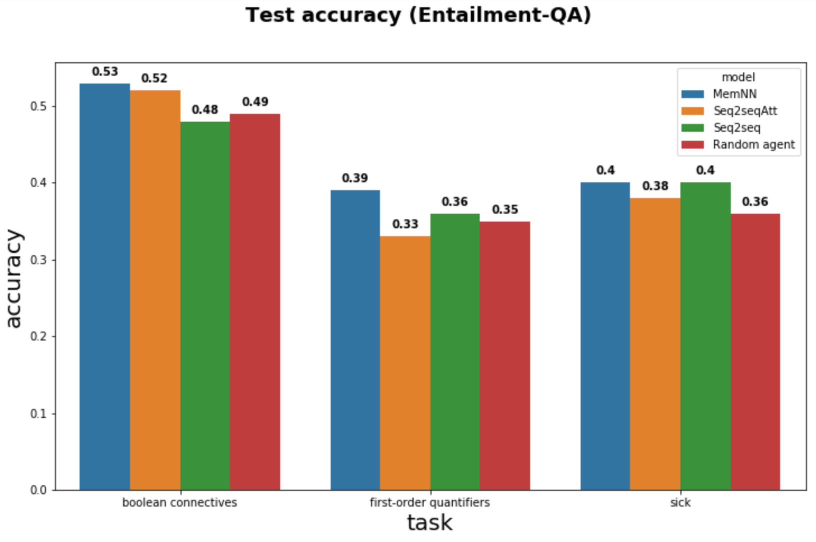
What is the semantic relation? A: neutral

Some dogs are playing in a river

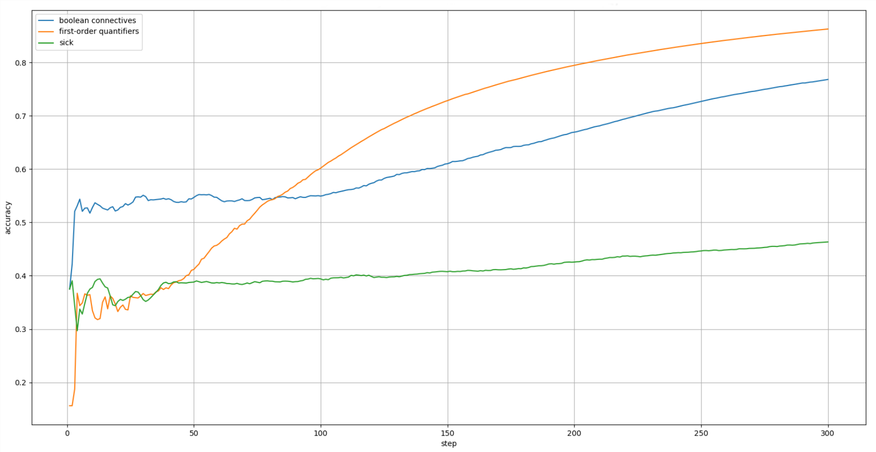
Some dogs are playing in a stream

What is the semantic relation? A: entailment

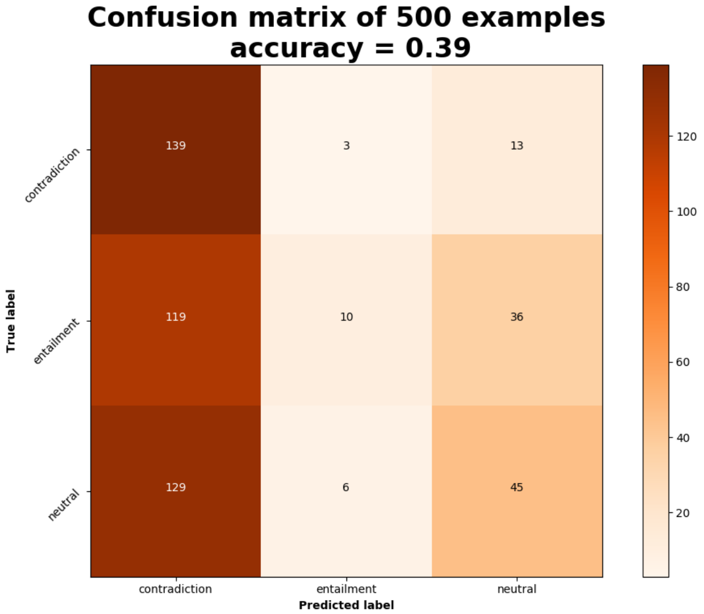
Preliminary Results



Preliminary Results



Preliminary Results



Future Steps

- Try to overcome the reported overfitting problem.
- Finish the Entailment-QA corpus.
- Explore new models not mentioned here, like Dynamic Memory Networks [3] and Memory Attention and Composition (MAC) cell [2].
- Create a visual version of the Entailment-QA to test logical inference with images.
- Check the reinforcement learning on dialog.
- Review the literature on the theory of comparing models [1].

Schedule

Activity	2016		2017		2018		2019		2020
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st
Courses									
Teaching Assist. (PAE)									
Bibliographic Review									
Software Implementation									
Qualification Writing									
Qualification Exam									
Finishing Entailment-QA task									
Visual Entailment-QA task									
Improve Training									
Adding new models									
Reinforcement Learning Methods									
Model Comparison Theory									
Thesis Writing									
Thesis Defense									

References I

A. Benavoli, G. Corani, J. Demsar, and M. Zaffalon.
Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis.
Journal of Machine Learning Research, 18:77:1–77:36, 2017.

D. A. Hudson and C. D. Manning.
Compositional attention networks for machine reasoning.
CoRR, abs/1803.03067, 2018.

A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher.
Ask me anything: Dynamic memory networks for natural language processing.
CoRR, abs/1506.07285, 2015.

References II

C. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau.
How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation.
CoRR, abs/1603.08023, 2016.

R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau.
Towards an automatic turing test: Learning to evaluate dialogue responses.
CoRR, abs/1708.07149, 2017.

References III

M. Marelli, L. Bentivogl, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli.
Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment.
SemEval 2014, 2014.

T. Mikolov, S. Kombrink, L. Burget, J. Cernocký, and S. Khudanpur.
Extensions of recurrent neural network language.
IEEE, pages 5528–5531, 2011.

O. Vinyals and Q. V. Le.
A neural conversational model.
CoRR, abs/1506.05869, 2015.

References IV

J. Weston, A. Bordes, S. Chopra, and T. Mikolov.
Towards ai-complete question answering: A set of prerequisite toy tasks.
CoRR, abs/1502.05698, 2015.