Adicionando robustez semântica a sistemas de diálogo

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

May 10, 2018

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

Geração de diálogo

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly

Sistemas de diálogo

- Goal-driven systems:
  - serviços de suporte técnico
  - marketing
  - sistemas de reserva
  - sistemas de informação
- Non-goal-driven systems
  - Conversas livres (sem fim específico)

Sistemas de diálogo baseados em redes neurais

Modelos de linguagem baseados em redes neurais

Nos chamamos de modelo de linguagem uma distribuição de probabildiade sobre uma sequencia de tokens em uma lingua natural.

$$P(x_1, x_2, x_3, x_4) = p$$

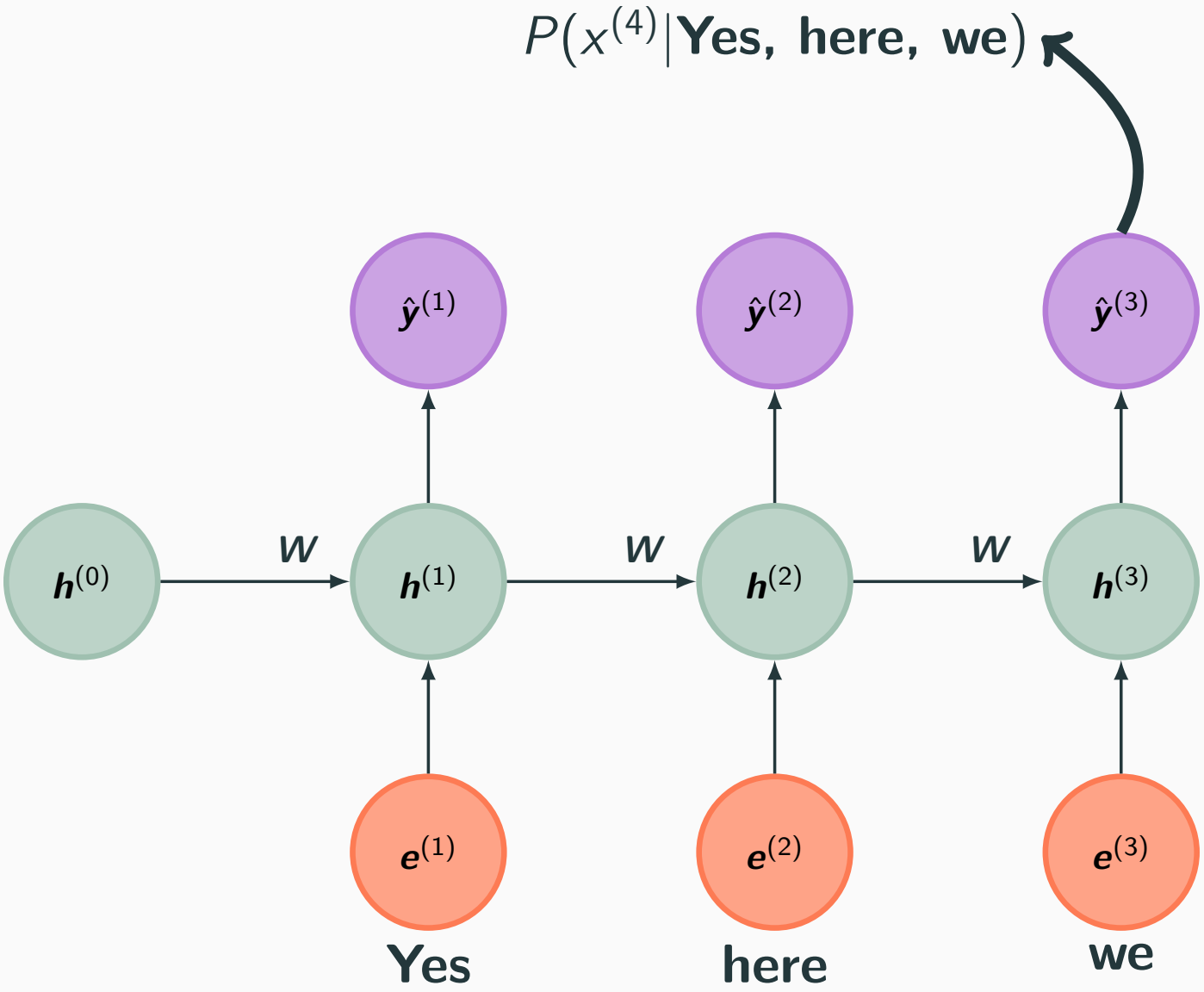
Em vez de usar uma abordagem que seja específica para o domínio da linguagem natural, podemos usar um modelo para predição de dados sequencias: **uma rede recorrente (RNN)**.

Nossa tarefa de aprendizado é estimar a distribuição de probabilidade

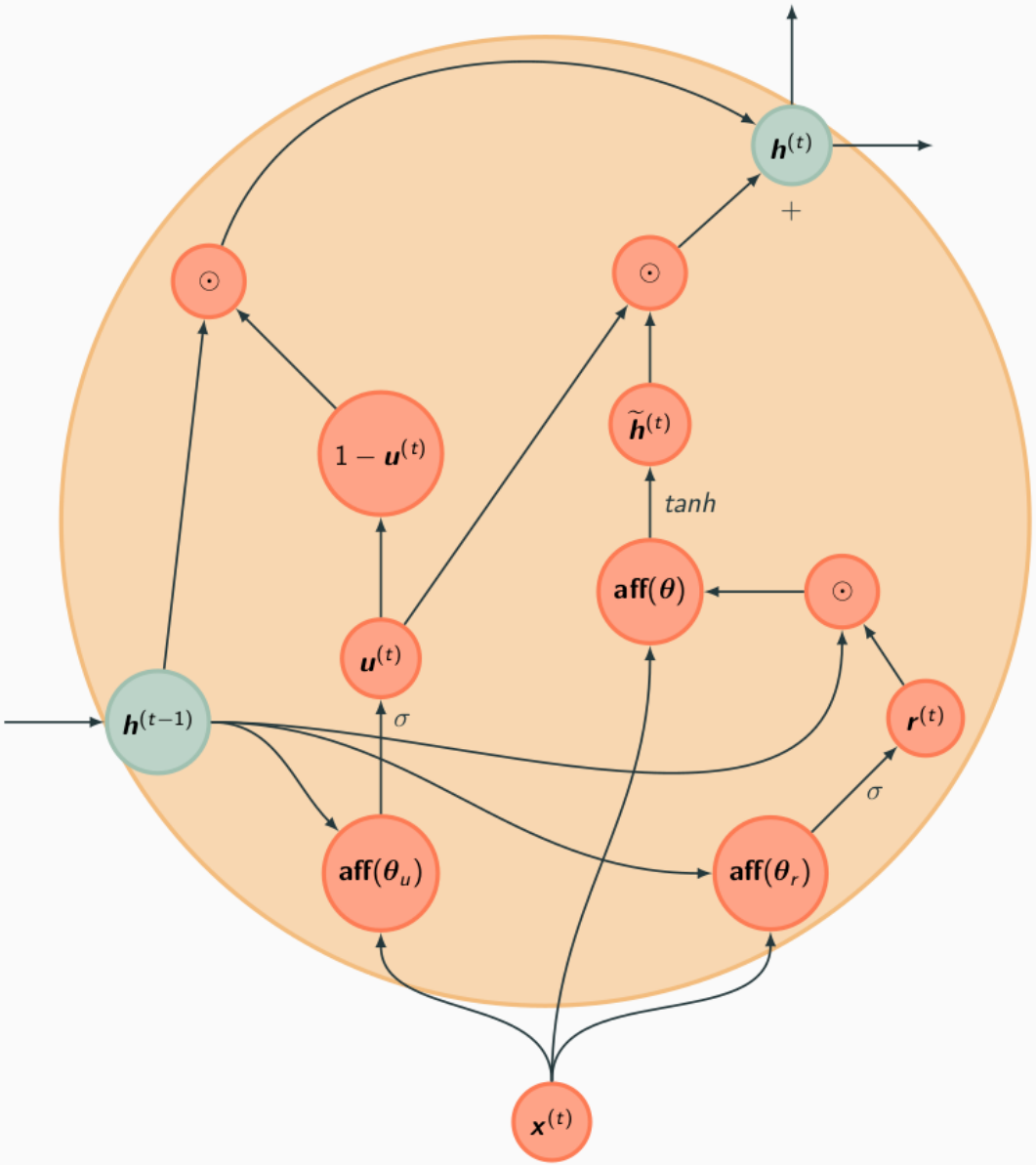
$$P(x_n = \text{palavra}_{j^*} | x_1, \ldots, x_{n-1})$$

para qualquer $(n-1)$ sequencia de palavras $x_1, \ldots, x_{n-1}$.

O modelo de linguagem com RNN



GRU: Gated Recurrent Units



Exemplo: TrumpBot

`https://github.com/felipessalvatore/MyTwitterBot`

Felipe Salvatore
@Felipessalvador

Hillary can make america great again. @greta @MarkBurnettTV #DinheiroNãoCompra #SecretBallot #خسوف_القمر

Traduzir do inglês

15:10 - 7 de ago de 2017

Felipe Salvatore
@Felipessalvador

Obama is all beautiful. I agree with people attacking me. Amazing. @CLewandowski_ #SecretBallot @garyplayer @greta

Traduzir do inglês

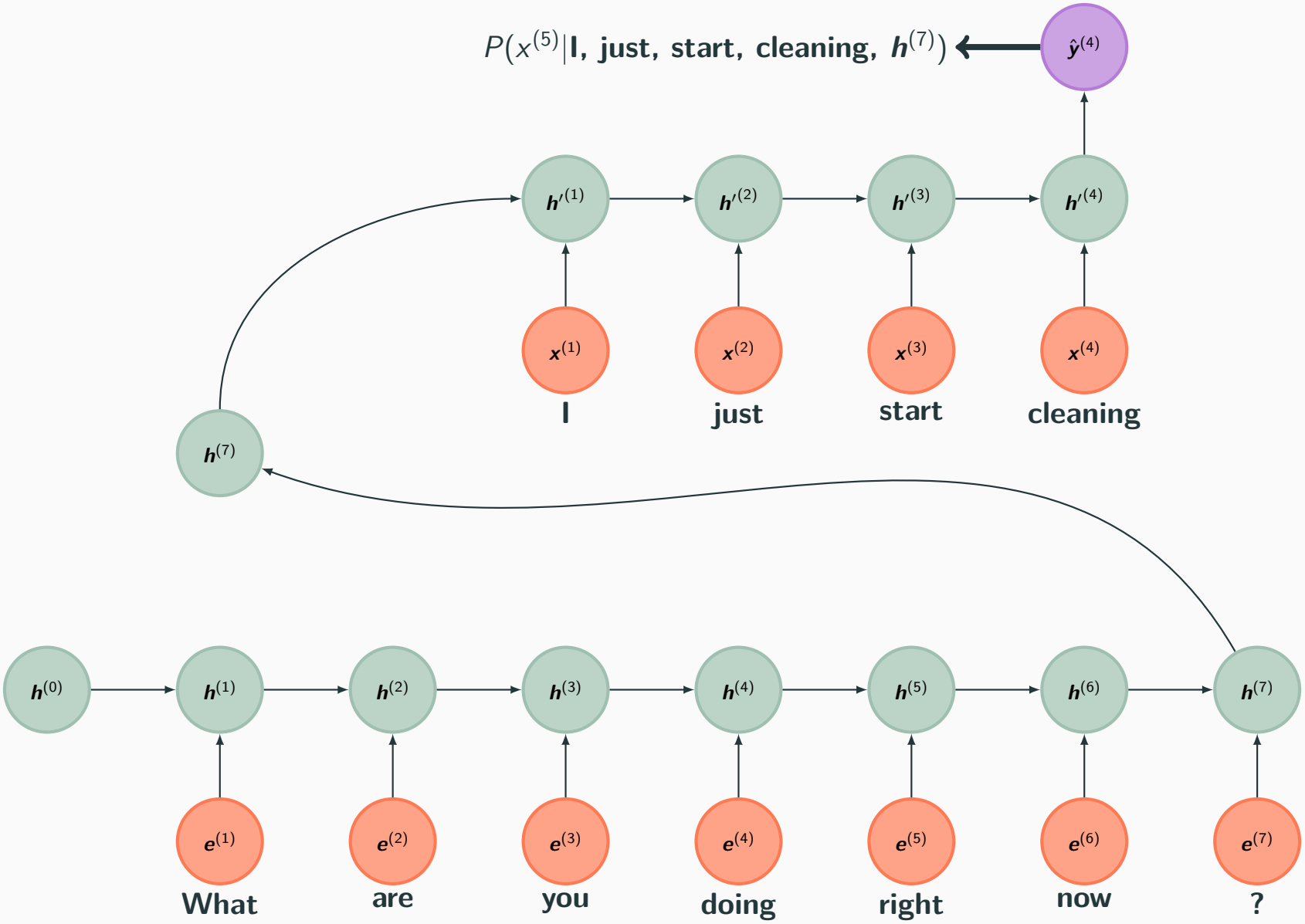
14:40 - 7 de ago de 2017

Exemplo: Funk Generator

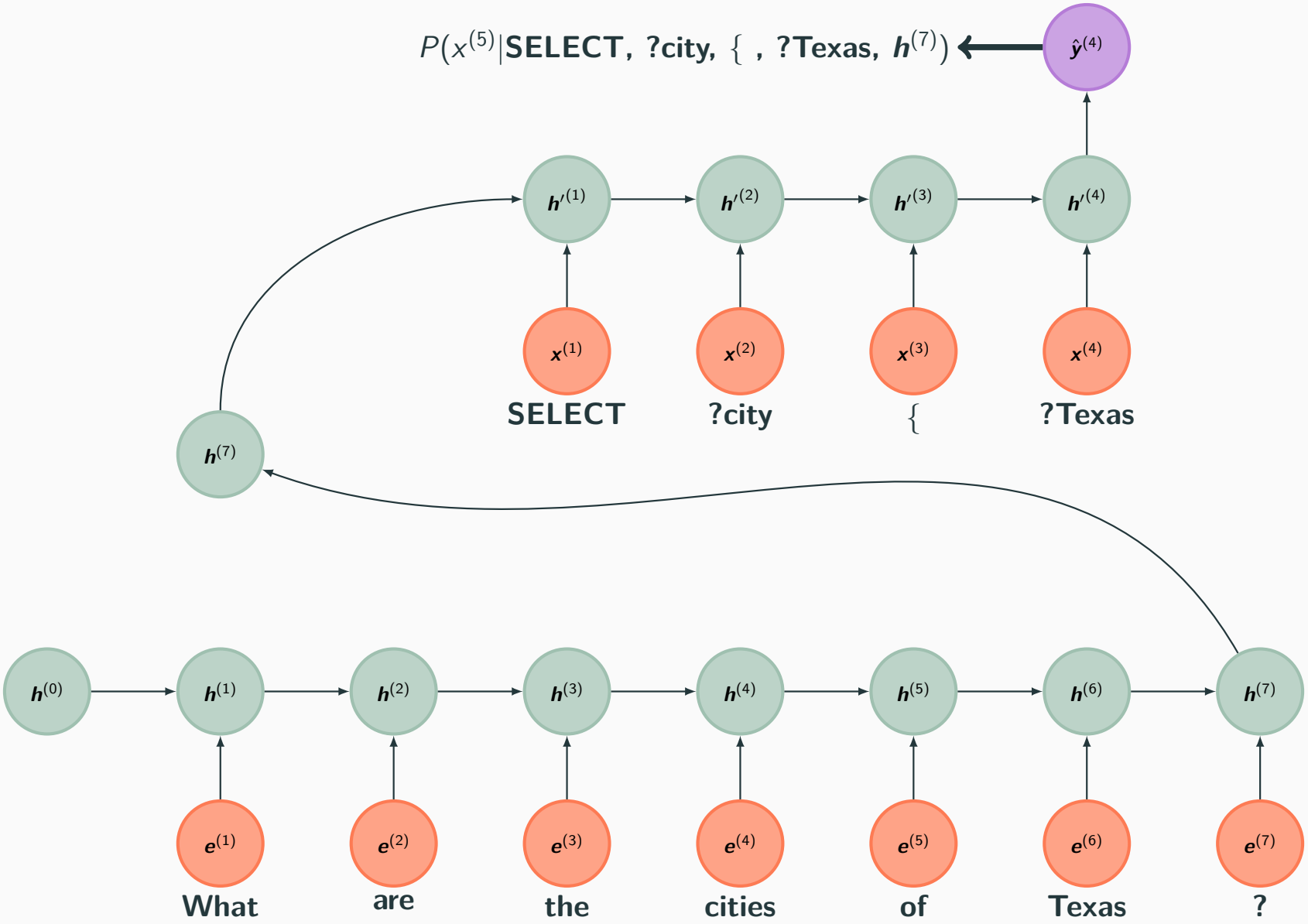
`https://github.com/lucasmoura/funk_generator`

É o dj que tá tocando e não sabe de nada
Eu já tô no clima e já tô no meu nome
Cordão de ouro no pescoço eu tô na moda
Com a camisa da
Louis
Vuitton
Pulo da morena que elas gosta
E se eu te pego no baile
De captiva de citroen ou de hayabusa
Tu viu a 1100 cilindradas
Se eu tô no litoral de cordão de ouro
De cordão de ouro no pescoço

Seq2seq: diálogo [8]



Seq2seq: tradução [5]



Exemplo de diálogo [8]

Human: *what is moral ?*
Machine: *what empowered humanity , what intellectual the essence is .*
Human: *what is immoral ?*
Machine: *the fact that you have a child .*
Human: *what is morality ?*
Machine: *what is altruism ?*
Human: *what is the definition of altruism ?*
Machine: *if you don 't believe in god , then you don 't know .*
Human: *ok ... so what is the definition of morality ?*
Machine: *well , the truth is , you 're not a believer in god almighty .*
Human: *tell me the definition of morality , i am quite upset now !*
Machine: *i 'm not ashamed of being a philosopher !*

Métricas

Avaliação humana [4]

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)

Avaliação automática: BLEU (bilingual evaluation understudy) [7]

Essa métrica compara n-gramas (até 4) da resposta candidata com os n-gramas da refência da tradução e conta o numero de acertos. Essa métrica também penaliza traduções muito curtas:

$$BLUE(r, \hat{r}) = min\left(1, \frac{len(\hat{r})}{len(r)}\right)\left(\prod_{n=1}^{4} precision_n(r, \hat{r})\right)^{\frac{1}{4}}$$

em que $precision_n(r, \hat{r})$ é o número de overlap de n gramas de r e $\hat{r}$ dividido pelo número de todos os n-gramas de $\hat{r}$.

- $BLUE(r, \hat{r}) \in [0, 1]$

Avaliação automática: problemas

Context of Conversation	**Context of Conversation**
A: dearest! question. how many thousands of people can panaad occupy? B: @user panaad has <number> k seat capacity while rizal has <number> k thats why they choose rizal i think .	A: never felt more sad than i am now B: @user aww why ? A: @user @user its a long story ! sure you wanna know it ? bahaha and thanks for caring btw <heart>
Ground Truth Response	**Ground Truth Response**
A: now i know about the siting capacity . thanks for the info @user great evening.	A: @user i don 't mind to hear it i 've got all day and youre welcome <number>
Proposed Response	**Proposed Response**
A: @user makes sense. thanks!	A: @user i know , i 'm just so happy for you !

"In particular, we show that these metrics (BLEU, METEOR, ROUGE) have only a small positive correlation on the chitchat oriented Twitter dataset, and no correlation at all on the technical Ubuntu Dialogue Corpus." [3]

De diálogos abertos para pequenas tarefas

bAbI [9]

Criar uma série de pequenas tarefas para testar diferentes capacidades de um sistema de diálogo.

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

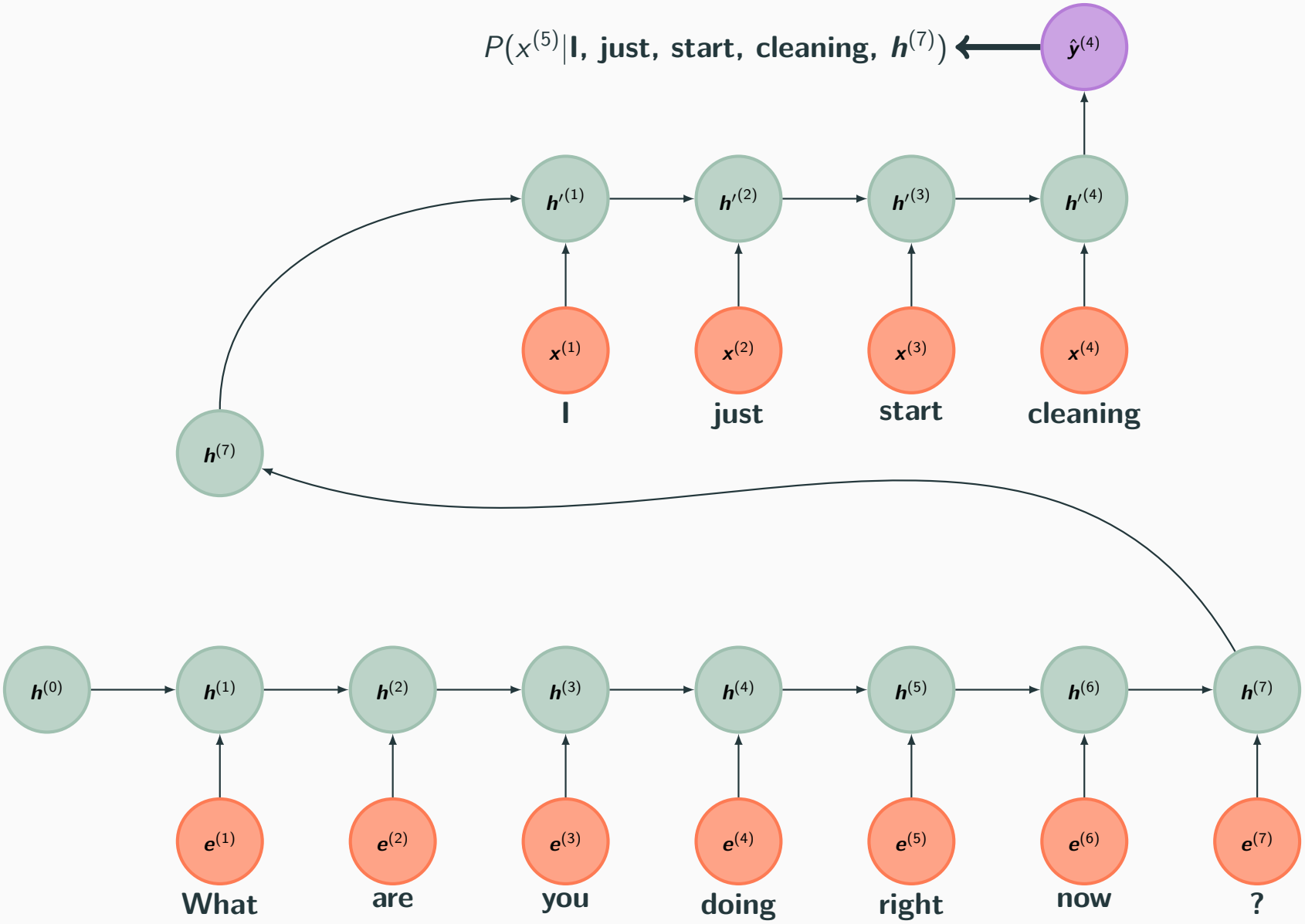
Task 6: Yes/No Questions

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

Seq2seq

- $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ sentença de entrada (source)
- $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ sentença de saida (target)
- $\boldsymbol{s} = f_{enc}(\boldsymbol{x}_n, \boldsymbol{h}_{n-1})$ representação até a entrada $\boldsymbol{x}_n$
  - $\boldsymbol{h}_0$ iniciado aleatoriamemnte
  - $\boldsymbol{h}_n = \boldsymbol{s}$
- $\boldsymbol{h}'_j = f_{dec}(\boldsymbol{y}_j, \boldsymbol{h}'_{j-1})$ representação saida no instante j
  - $\boldsymbol{h}'_0 = \boldsymbol{s}$
- $p(y_j|y_1, \ldots, y_{j-1}, x_1, \ldots, x_n) = softmax(\boldsymbol{W}_s \boldsymbol{h}'_j + \boldsymbol{b}_s)$
- $\hat{y}_j = \arg\max p(y_j|y_1, \ldots, y_{j-1}, x_1, \ldots, x_n)$

Seq2seq


$P(x^{(5)}|\text{I, just, start, cleaning, } h^{(7)})$
$\hat{y}^{(4)}$
$h'^{(1)}$
$h'^{(2)}$
$h'^{(3)}$
$h'^{(4)}$
$x^{(1)}$
$x^{(2)}$
$x^{(3)}$
$x^{(4)}$
I
just
start
cleaning
$h^{(7)}$
$h^{(0)}$
$h^{(1)}$
$h^{(2)}$
$h^{(3)}$
$h^{(4)}$
$h^{(5)}$
$h^{(6)}$
$h^{(7)}$
$e^{(1)}$
$e^{(2)}$
$e^{(3)}$
$e^{(4)}$
$e^{(5)}$
$e^{(6)}$
$e^{(7)}$
What
are
you
doing
right
now
?

Modelo de atenção

No modelo de atenção vamos contruir um vetor de contexto $\boldsymbol{c}_t$ para selecionar as informações na sentença de entrada.
Uma vez costruido $\boldsymbol{c}_t$, definimos um estado de atenção

$$\tilde{\boldsymbol{h}}_t = tahn(\boldsymbol{W}_c[\boldsymbol{c}_t; \boldsymbol{h}_t])$$

e geramos a predição

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

Como definimos c_t?

Construimos uma matrix de alinhamento a tal que

- a_{ts} é a probabilidade da representação de entrada $\boldsymbol{h}_s$ ser relevante para a saída $\hat{y}_t$.

- $a_{ts} = \frac{exp(score(\tilde{\boldsymbol{h}}_t, \boldsymbol{h}_s))}{\sum_j exp(score(\tilde{\boldsymbol{h}}_t, \boldsymbol{h}_j))}$

- $score(\tilde{\boldsymbol{h}}_t, \boldsymbol{h}_s) = \begin{cases} \tilde{\boldsymbol{h}}_t^\top \boldsymbol{h}_s \\ \tilde{\boldsymbol{h}}_t^\top \boldsymbol{W}_a \boldsymbol{h}_s \\ \boldsymbol{v}_a^\top tahn(\boldsymbol{W}_a[\tilde{\boldsymbol{h}}_t; \boldsymbol{h}_s]) \end{cases}$

- $\boldsymbol{c}_t$ é uma soma ponderada de todas as representações de entrada:

$$\boldsymbol{c}_t = \sum_{s=1}^{n} a_{ts} \boldsymbol{h}_s$$

Modelo de memória

- $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ sentenças de contexto
- $\boldsymbol{q}$ pergunta
- $\boldsymbol{a}$ resposta
- $\{\boldsymbol{s}_i\} \rightarrow^{\boldsymbol{A}} \{\boldsymbol{m}_i\}$ (vetores de memôria)
- $\boldsymbol{q} \rightarrow^{\boldsymbol{B}} \boldsymbol{u}$ (estado interno)
- $\{p_i\} = \{softmax(\boldsymbol{u}^T \boldsymbol{m}_i)\}$ ("match" entre $\boldsymbol{m}_i$ e $\boldsymbol{u}$)
- $\{\boldsymbol{s}_i\} \rightarrow^{\boldsymbol{C}} \{\boldsymbol{c}_i\}$
- $\boldsymbol{o} = \sum_i p_i \boldsymbol{c}_i$
- $\hat{\boldsymbol{a}} = softmax(\boldsymbol{W}(\boldsymbol{o} + \boldsymbol{u}))$

Podemos ter k camadas e memória (hops)

- $\boldsymbol{u}^k = \boldsymbol{u}^{k-1} + \boldsymbol{o}^{k-1}$
- $\{\boldsymbol{s}^k{}_i\} \rightarrow^{vectA^k} \{\boldsymbol{m}^k{}_i\}$
- $\{\boldsymbol{s}^k{}_i\} \rightarrow^{vectC^k} \{\boldsymbol{c}^k{}_i\}$
- $\{p^k{}_i\} = \{softmax(\boldsymbol{u}^k{}^\top \boldsymbol{m}_i^k)\}$
- $\boldsymbol{o}^k = \sum_i p^k{}_i \boldsymbol{c}^k{}_i$
- $\hat{\boldsymbol{a}} = softmax(\boldsymbol{W}(\boldsymbol{o}^k + \boldsymbol{u}^k))$

ParlAI

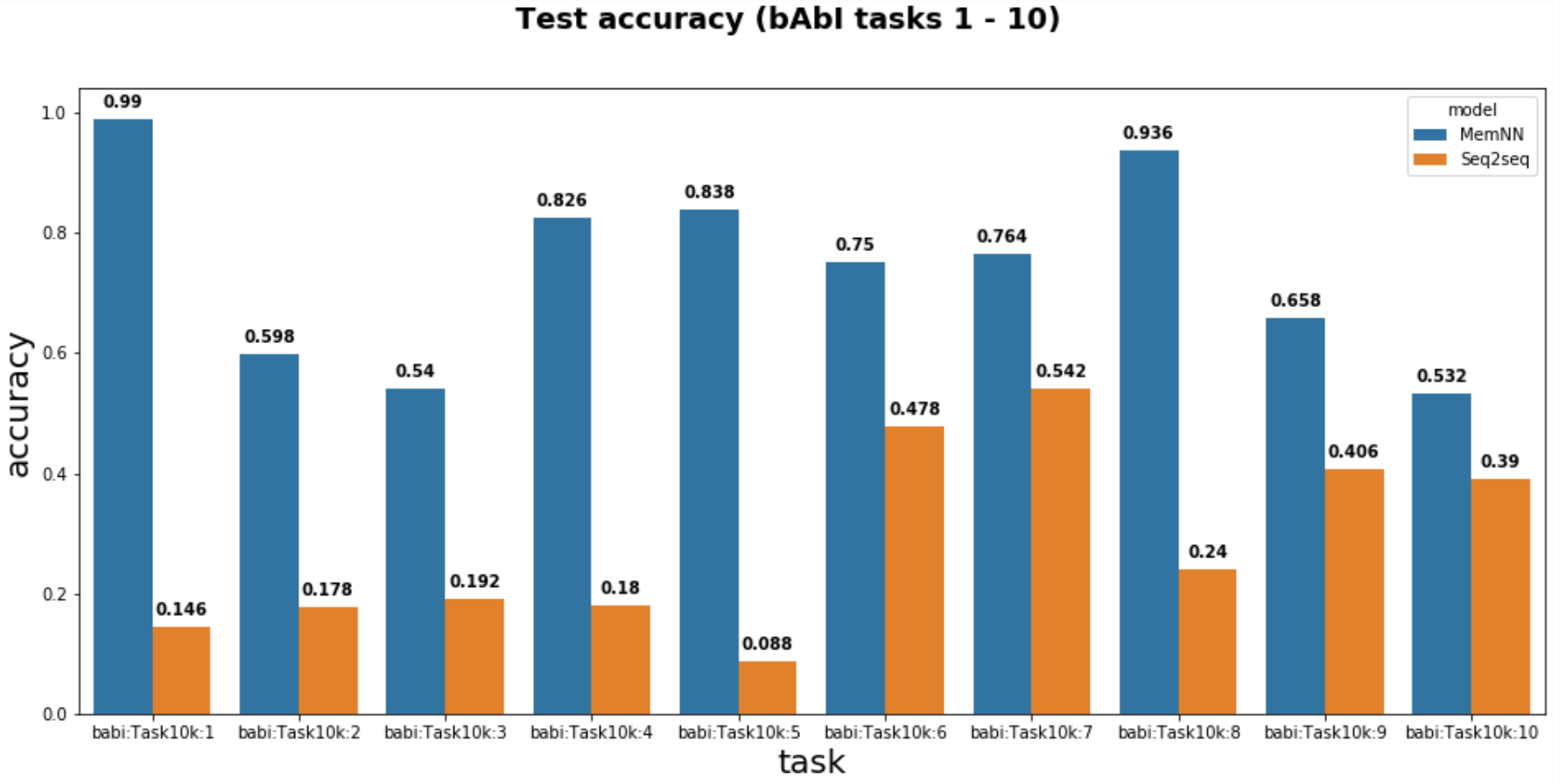
`https://github.com/facebookresearch/ParlAI`

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

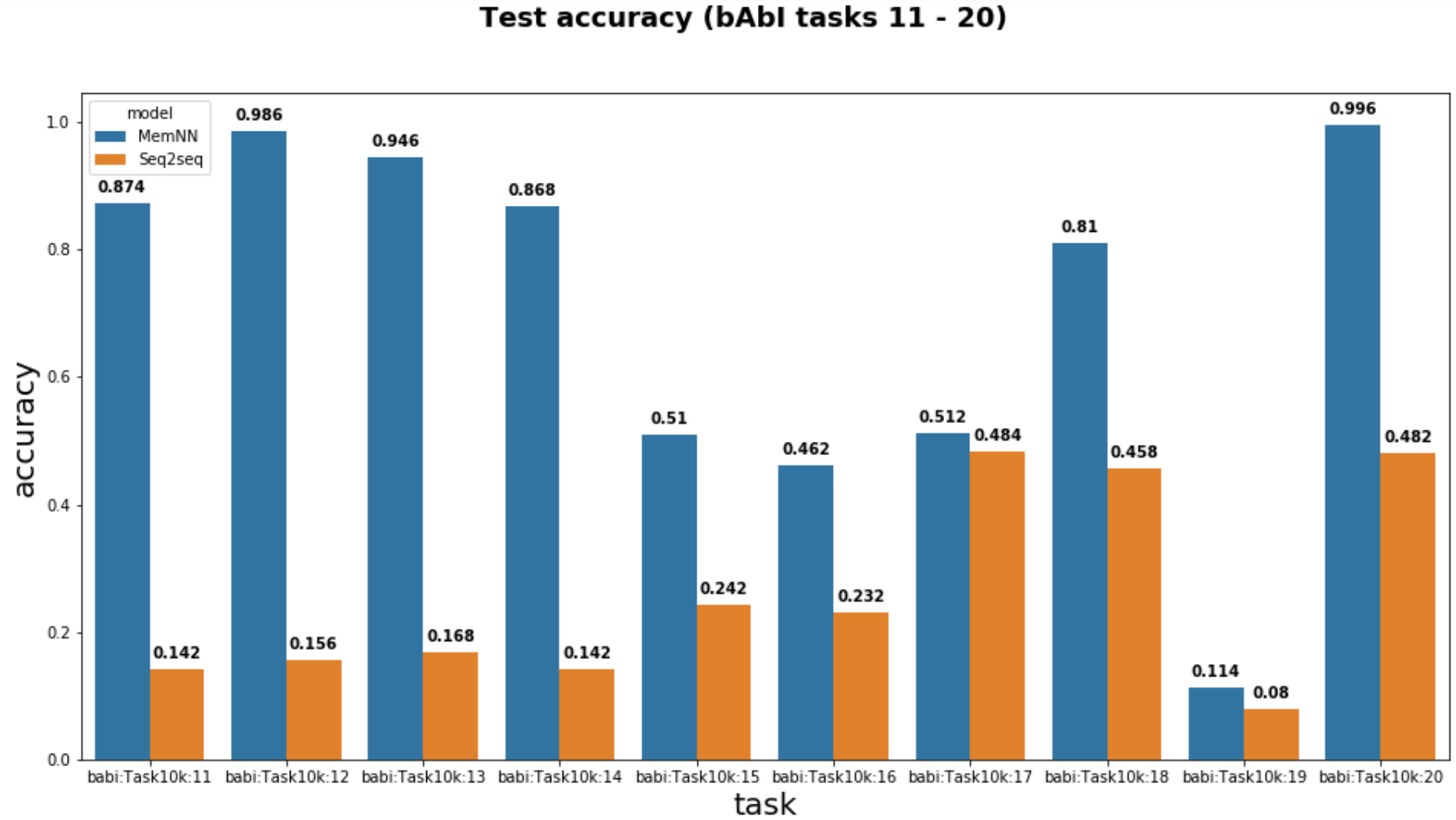
Its goal is to provide researchers:

- 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"

Experimentos



Experimentos



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

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

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 *is a* P^1
c^1 *is* C^1
c^2 *is a* P^2
c^2 *is* C^2
c^3 *is a* P^3
c^3 *is* C^3
c^4 *is a* P^4
c^4 *is* C^4
c *is a* P^j
What color is c? *A:* C^j

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.

Quora question pairs [1]

Who creates bitcoins?
Who invented Bitcoin?
Are the above questions duplicate? A: no

How aeroplanes fly?
How do airplanes fly?
Are the above questions duplicate? A: yes

What actually is brexit?
What is brexit?
Are the above questions duplicate? A: yes

hat is my ethnicity?
What does ethnicity mean?
Are the above questions duplicate? A: no

DialogGym

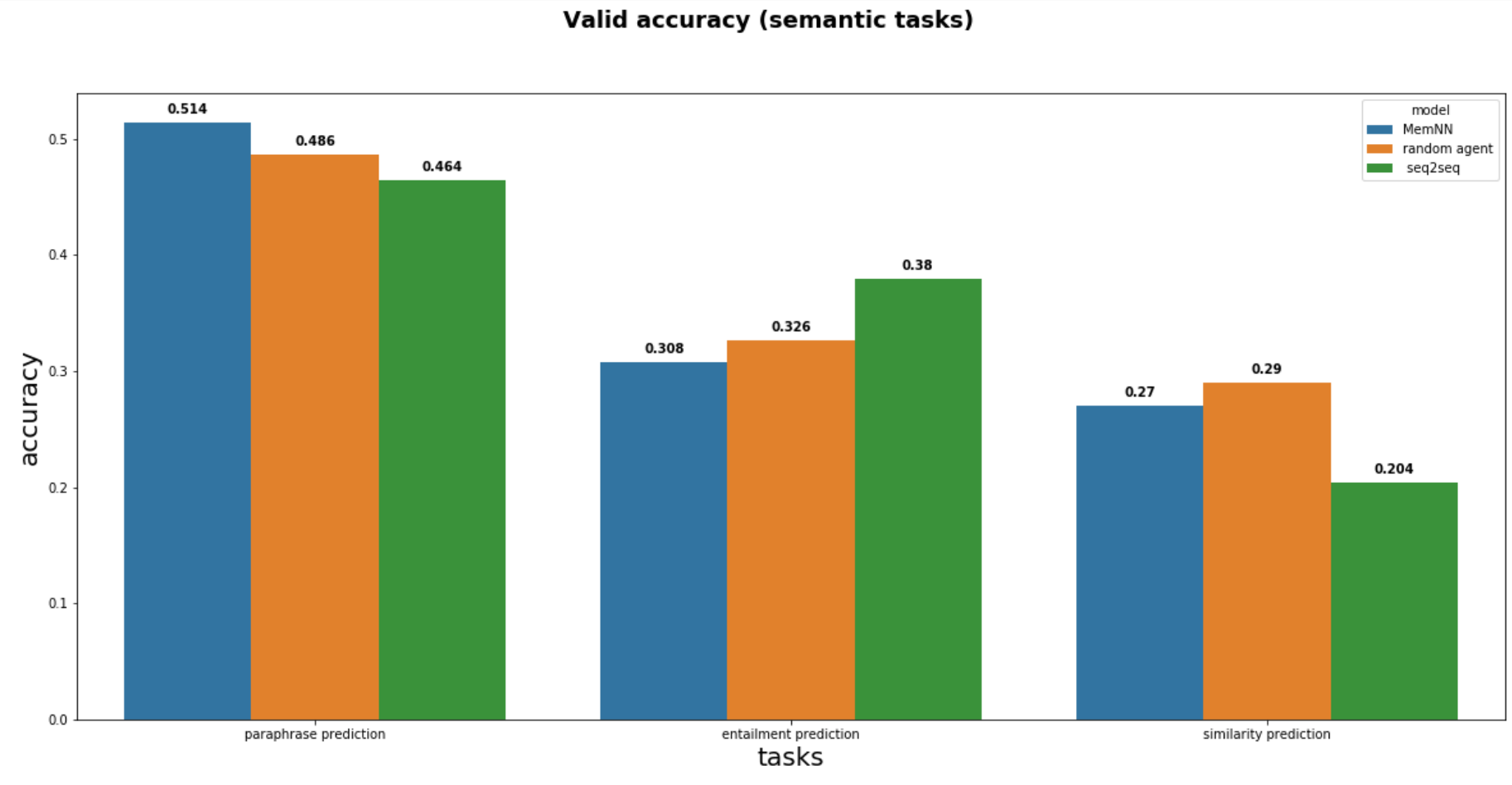
DialogGYM

https://github.com/felipessalvatore/DialogGym

Um novo conjunto de tarefas

- **Task 1: entailment prediction** Given two sentences p and q the agent is asked to detect a basic entailment relation between them, i.e., the agent should respond if p implies q, if p contradicts q or if p is neutral to q. For example, the sentences "*A man is thinking*" and "*There is no man thinking*" is given to the agent, he needs to detect the quantifier to spot the contradiction between these two informations.
- **Task 2: similarity prediction** The agent is questioned to indicate how related are the meaning of two sentences, e.g., "*A man is reading the email. Someone is reading the email. Are the sentences above related?*". There are only 4 possible answers: "not related", "somewhat related", "related", "strongly related".
- **Task 3: paraphrase prediction** The agent is asked (a yes/no question) to identify if two given questions express the same meaning using different words, e.g., "*Who was Pele? Who is Pele? Are the above questions duplicate?*".

Primeiros resultados



Podemos melhorar os resultados para as tarefas específicas

Features	Description	# of features
Negation	True if either sentence contains explicit negation; False otherwise	1
Word overlap	Ratio of overlapping word types to total word types in s_1 and s_2	1
Denotational constituent similarity	Positive normalized PMI of constituent nodes in the denotation graph	30
Distributional constituent similarity	Cosine similarity of vector representations of constituent phrases	30
Alignment	Ratio of number of aligned words to length of s_1 and s_2; max, min, average unaligned chunk length; number of unaligned chunks	23
Unaligned matching	Ratio of number of matched chunks to unaligned chunks; max, min, average matched chunk similarity; number of crossings in matching	31
Chunk alignment	Number of chunks; number of unaligned chunk labels; ratio of unaligned chunk labels to number of chunks; number of matched labels; ratio of matched to unmatched chunk labels	17
Synonym	Number of matched synonym pairs (w_1, w_2)	1
Hypernym	Number of matched hypernym pairs (w_1, w_2), number of matched hypernym pairs (w_2, w_1)	2
Antonym	Number of matched antonym pairs (w_1, w_2)	1

Por exemplo, em [2] os autores conseguiram 84.6% de accurácia no SICK.

Mas não queremos "tunar" um modelo para uma tarefa específica!

Olhando as perguntas geradas pelo SICK

The parrot is talking into the microphone
The parrot is speaking
What is the semantic relation? A: entailment

There is no man cutting tomatoes
A man is cutting tomatoes
What is the semantic relation? A: contradiction

Paper is being cut with scissors
Someone is cutting some paper with scissors
What is the semantic relation? A: entailment

An elder man is sitting on a bench and is angry
An elderly man is sitting on a bench
What is the semantic relation? A: entailment

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

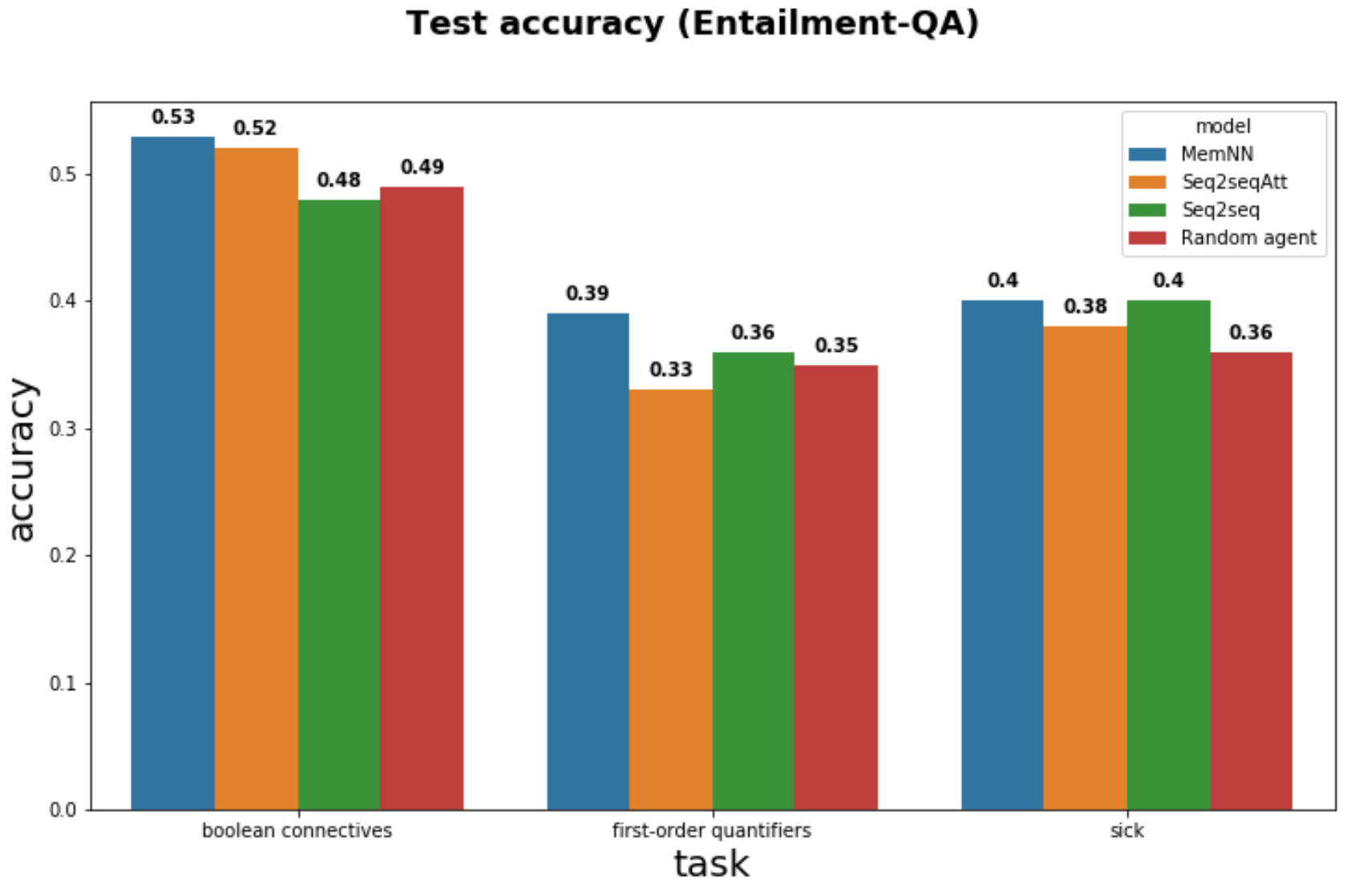
What is the semantic relation? A: contradiction

Every person is beautiful

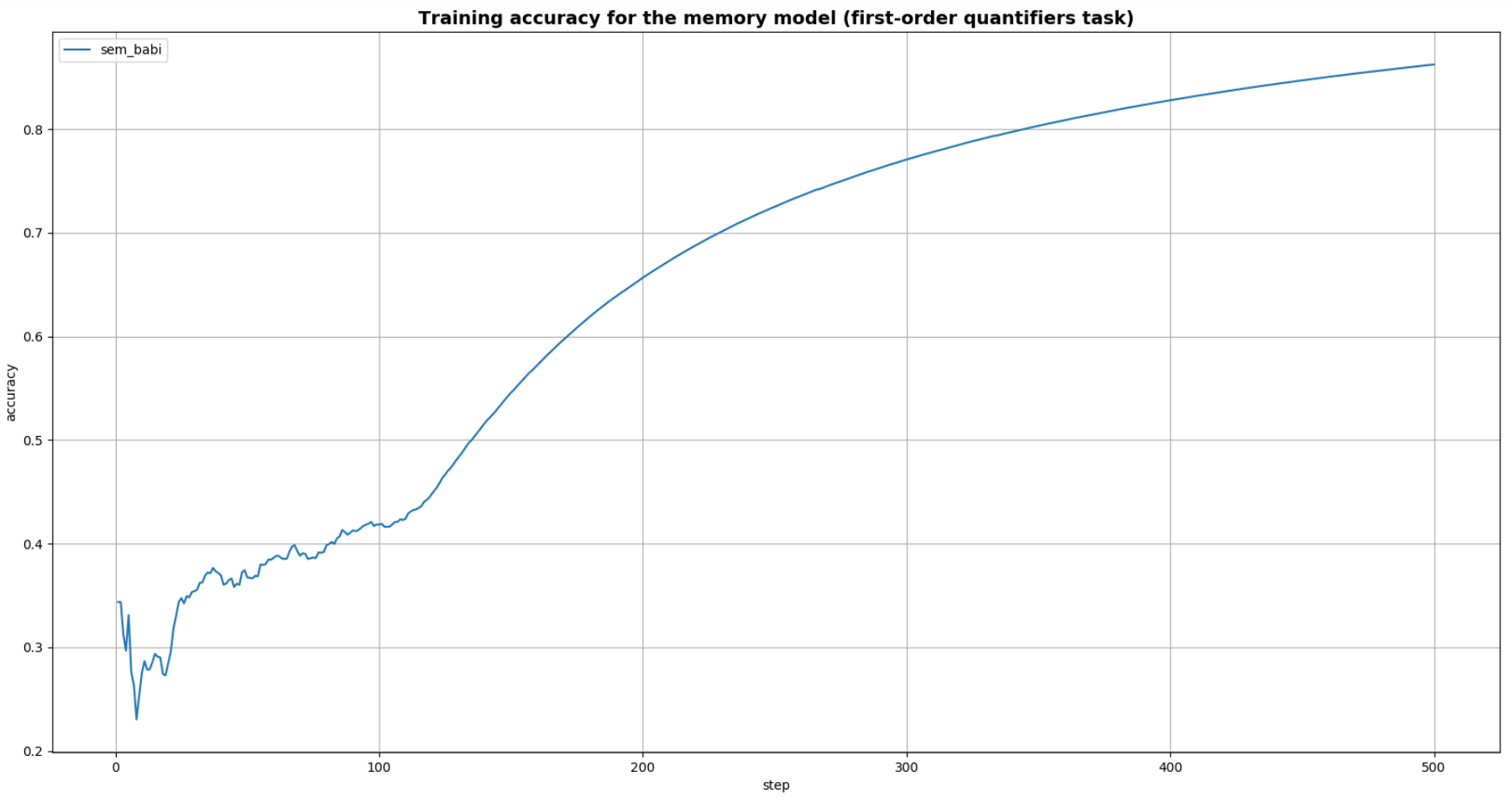
Abilene is not blue

What is the semantic relation? A: neutral

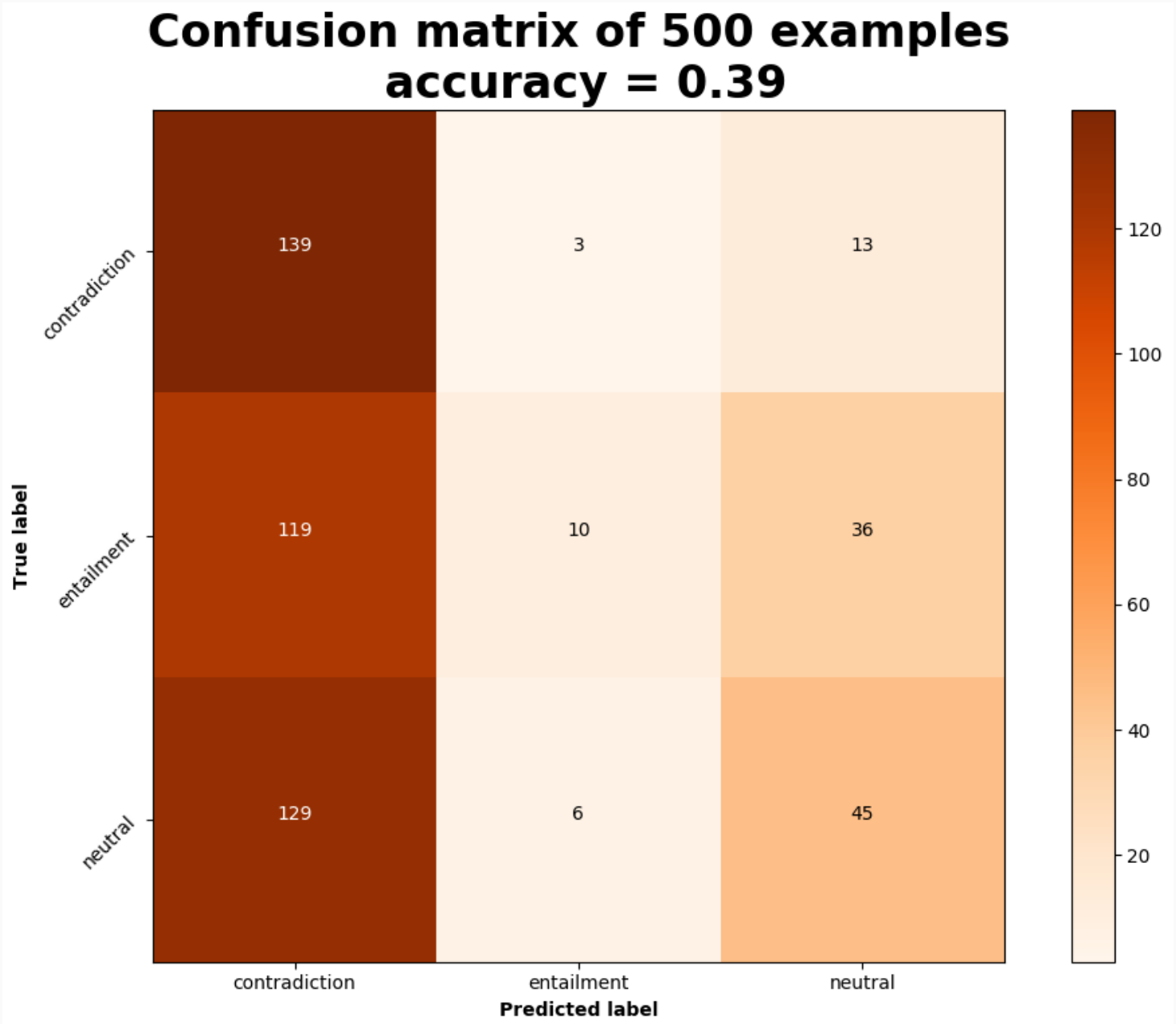
Resultados até agora



Resultados até agora



Resultados até agora



Próximos passos

- Terminar as tarefas
- Melhor o treinamento com os modelos atuais
- Explorar novos modelos

Referências I


Quora question pairs.
`https://www.kaggle.com/c/quora-question-pairs`.


A. Lai and J. Hockenmaier.
Illinois-lh: A denotational and distributional approach to semantics.
In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334. Association for Computational Linguistics, 2014.


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.

Referências II

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.

F. F. Luz and M. Finger.
Semantic parsing natural language into sparql: an lstm enconder- decoder neural net approach.
2017.

Referências 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.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu.
Bleu: A method for automatic evaluation of machine translation.
In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2001.

Referências IV

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

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.