Machine Reading, Models and Applications

Julien Perez
Machine Learning and Optimization group

4th July, 2018

NAVER LABS

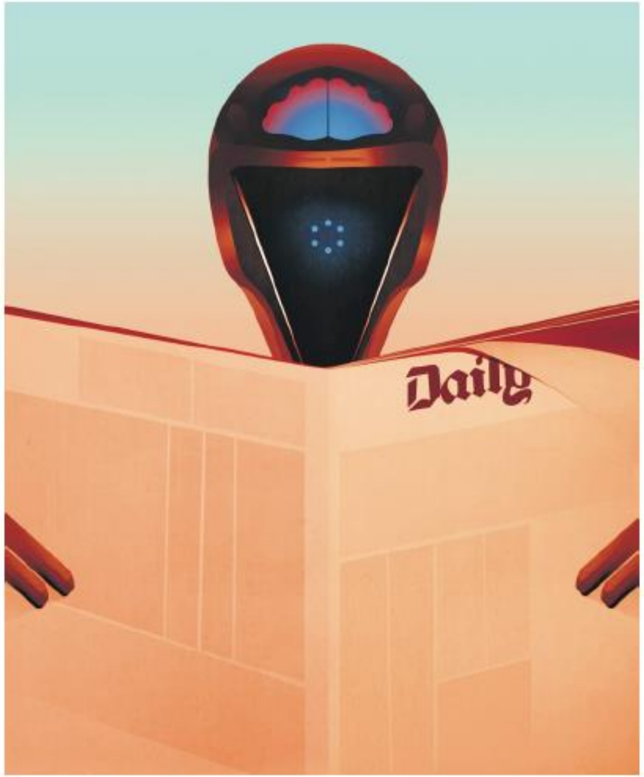
Content

1. Machine reading tasks

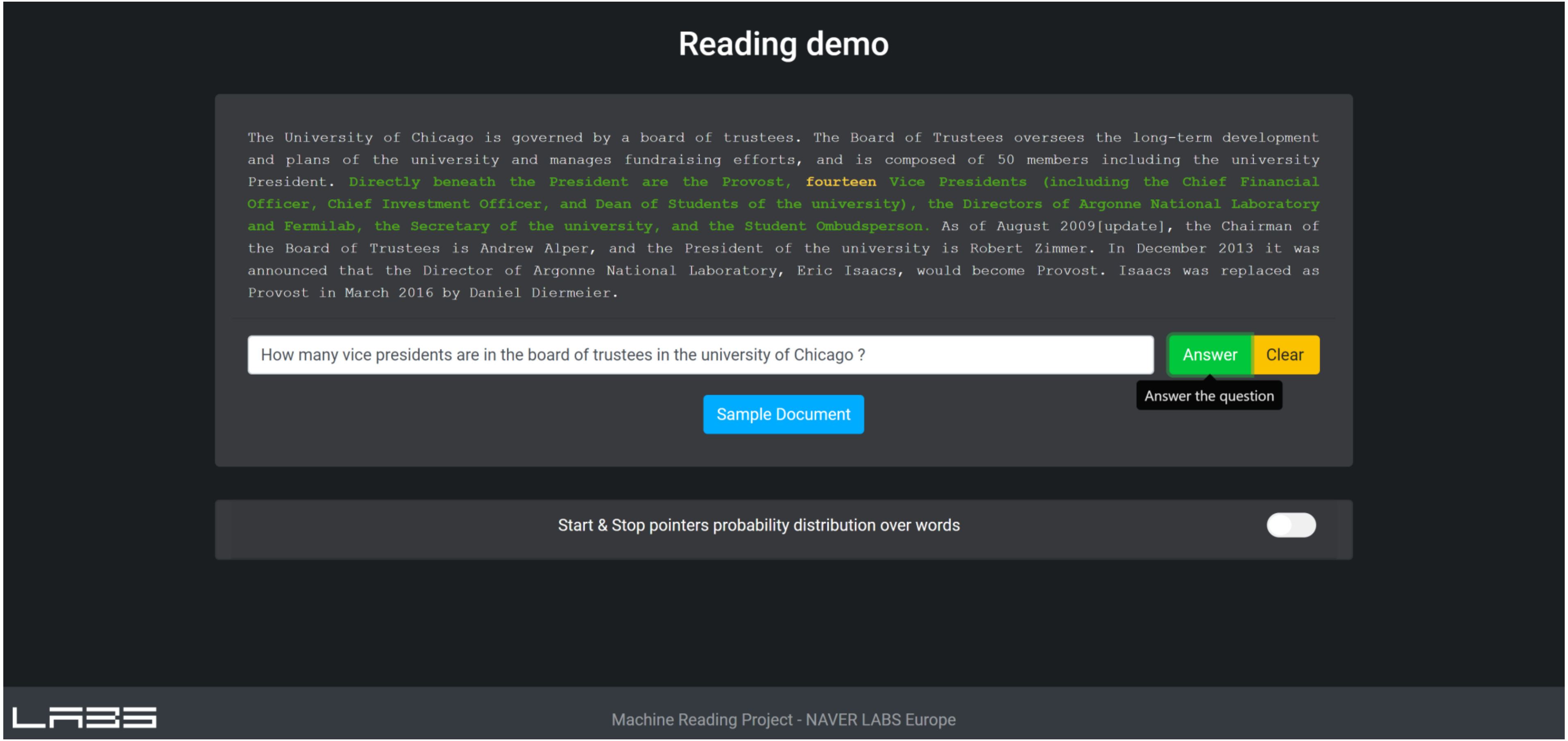
2. Models of reading

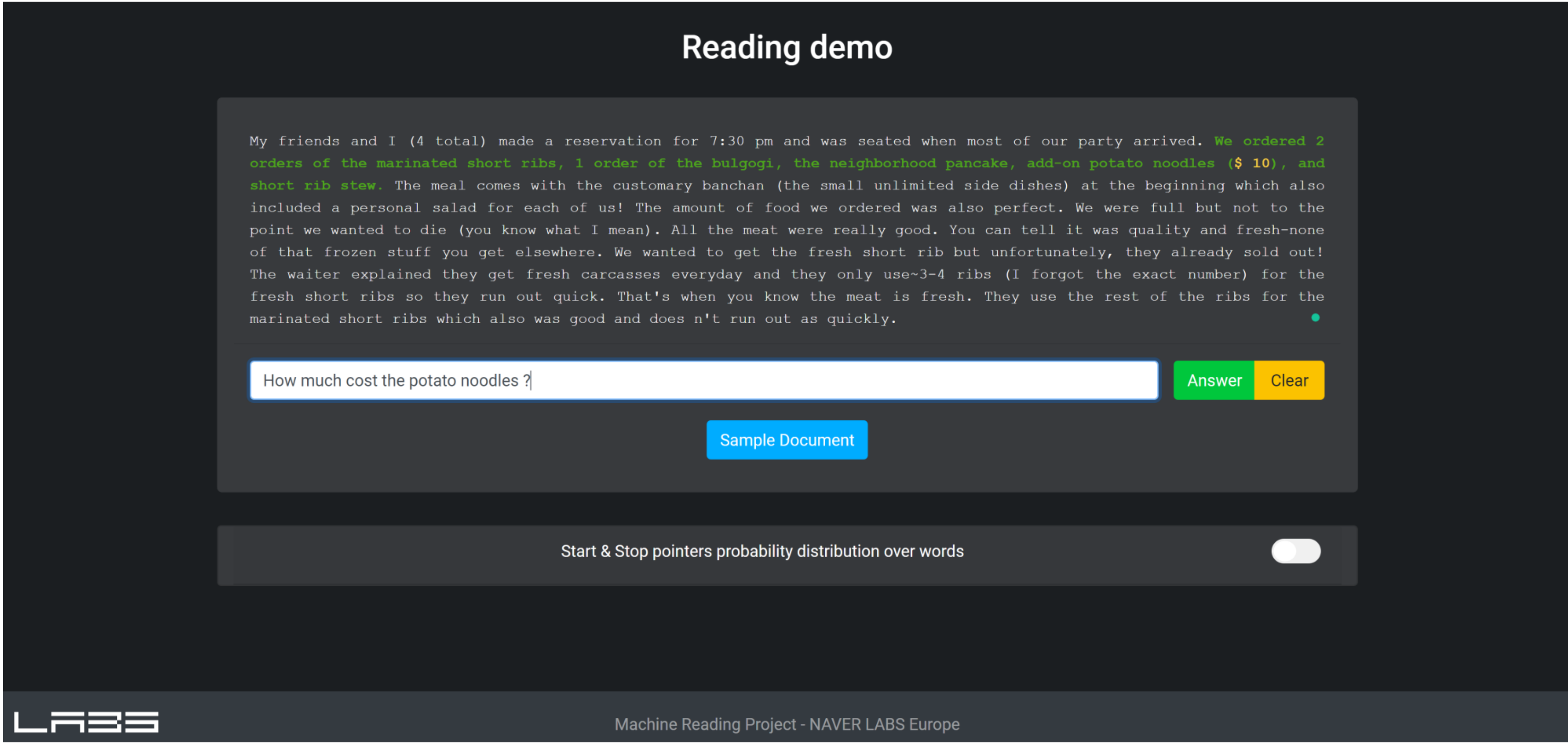
3. Applications

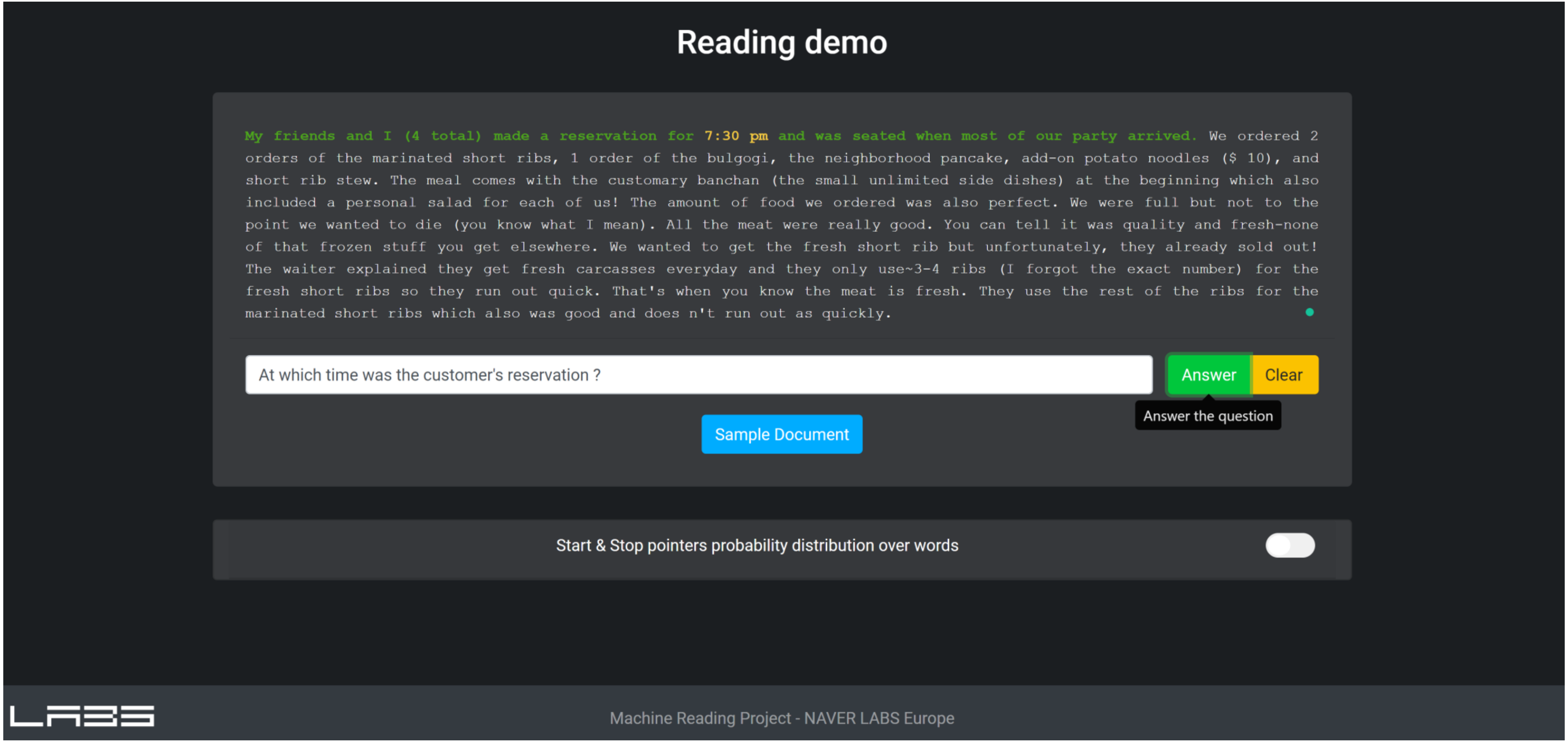
4. Open Questions

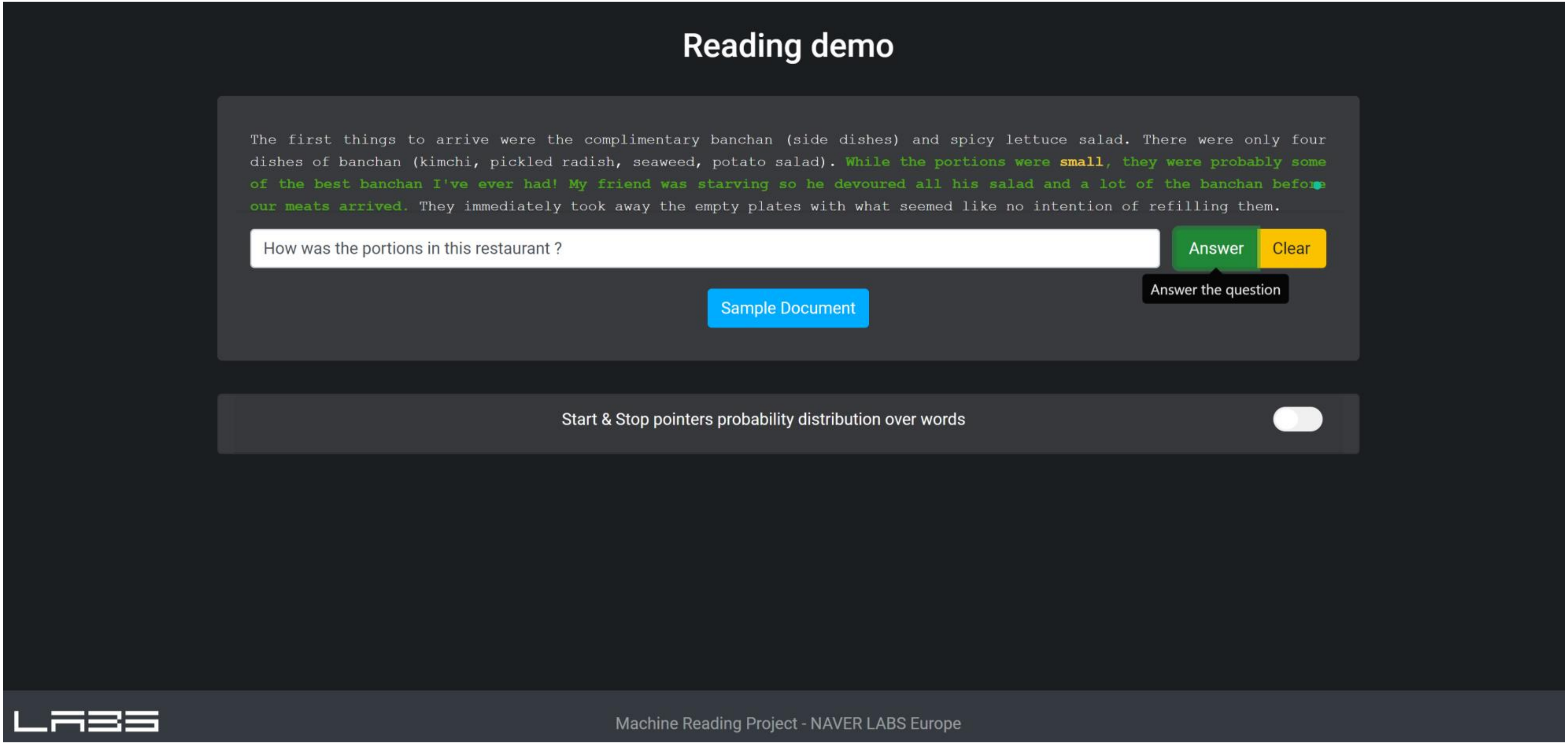


Courtesy of Phil Blunsom


Reading demo
The University of Chicago is governed by a board of trustees. The Board of Trustees oversees the long-term development and plans of the university and manages fundraising efforts, and is composed of 50 members including the university President. Directly beneath the President are the Provost, fourteen Vice Presidents (including the Chief Financial Officer, Chief Investment Officer, and Dean of Students of the university), the Directors of Argonne National Laboratory and Fermilab, the Secretary of the university, and the Student Ombudsperson. As of August 2009[update], the Chairman of the Board of Trustees is Andrew Alper, and the President of the university is Robert Zimmer. In December 2013 it was announced that the Director of Argonne National Laboratory, Eric Isaacs, would become Provost. Isaacs was replaced as Provost in March 2016 by Daniel Diermeier.
How many vice presidents are in the board of trustees in the university of Chicago ?
Answer
Clear
Answer the question
Sample Document
Start & Stop pointers probability distribution over words
Machine Reading Project - NAVER LABS Europe


Reading demo
My friends and I (4 total) made a reservation for 7:30 pm and was seated when most of our party arrived. We ordered 2 orders of the marinated short ribs, 1 order of the bulgogi, the neighborhood pancake, add-on potato noodles ($ 10), and short rib stew. The meal comes with the customary banchan (the small unlimited side dishes) at the beginning which also included a personal salad for each of us! The amount of food we ordered was also perfect. We were full but not to the point we wanted to die (you know what I mean). All the meat were really good. You can tell it was quality and fresh-none of that frozen stuff you get elsewhere. We wanted to get the fresh short rib but unfortunately, they already sold out! The waiter explained they get fresh carcasses everyday and they only use~3-4 ribs (I forgot the exact number) for the fresh short ribs so they run out quick. That's when you know the meat is fresh. They use the rest of the ribs for the marinated short ribs which also was good and does n't run out as quickly.
How much cost the potato noodles ?
Answer
Clear
Sample Document
Start & Stop pointers probability distribution over words
Machine Reading Project - NAVER LABS Europe


Reading demo
My friends and I (4 total) made a reservation for 7:30 pm and was seated when most of our party arrived. We ordered 2 orders of the marinated short ribs, 1 order of the bulgogi, the neighborhood pancake, add-on potato noodles ($ 10), and short rib stew. The meal comes with the customary banchan (the small unlimited side dishes) at the beginning which also included a personal salad for each of us! The amount of food we ordered was also perfect. We were full but not to the point we wanted to die (you know what I mean). All the meat were really good. You can tell it was quality and fresh-none of that frozen stuff you get elsewhere. We wanted to get the fresh short rib but unfortunately, they already sold out! The waiter explained they get fresh carcasses everyday and they only use~3-4 ribs (I forgot the exact number) for the fresh short ribs so they run out quick. That's when you know the meat is fresh. They use the rest of the ribs for the marinated short ribs which also was good and does n't run out as quickly.
At which time was the customer's reservation ?
Answer
Clear
Answer the question
Sample Document
Start & Stop pointers probability distribution over words
Machine Reading Project - NAVER LABS Europe


Reading demo
The first things to arrive were the complimentary banchan (side dishes) and spicy lettuce salad. There were only four dishes of banchan (kimchi, pickled radish, seaweed, potato salad). While the portions were small, they were probably some of the best banchan I've ever had! My friend was starving so he devoured all his salad and a lot of the banchan before our meats arrived. They immediately took away the empty plates with what seemed like no intention of refilling them.
How was the portions in this restaurant ?
Answer
Clear
Answer the question
Sample Document
Start & Stop pointers probability distribution over words
Machine Reading Project - NAVER LABS Europe

Multi Documents Answering

In Blade Runner, which company built the replicants ?

Wikipedia

Predict

Clear

Matched documents

Blade Runner (franchise)	657.27
Blade Runner	657.27
Blade Runner 2049	609.13
Replicant	609.13
Blade Runner (1997 video game)	543.61

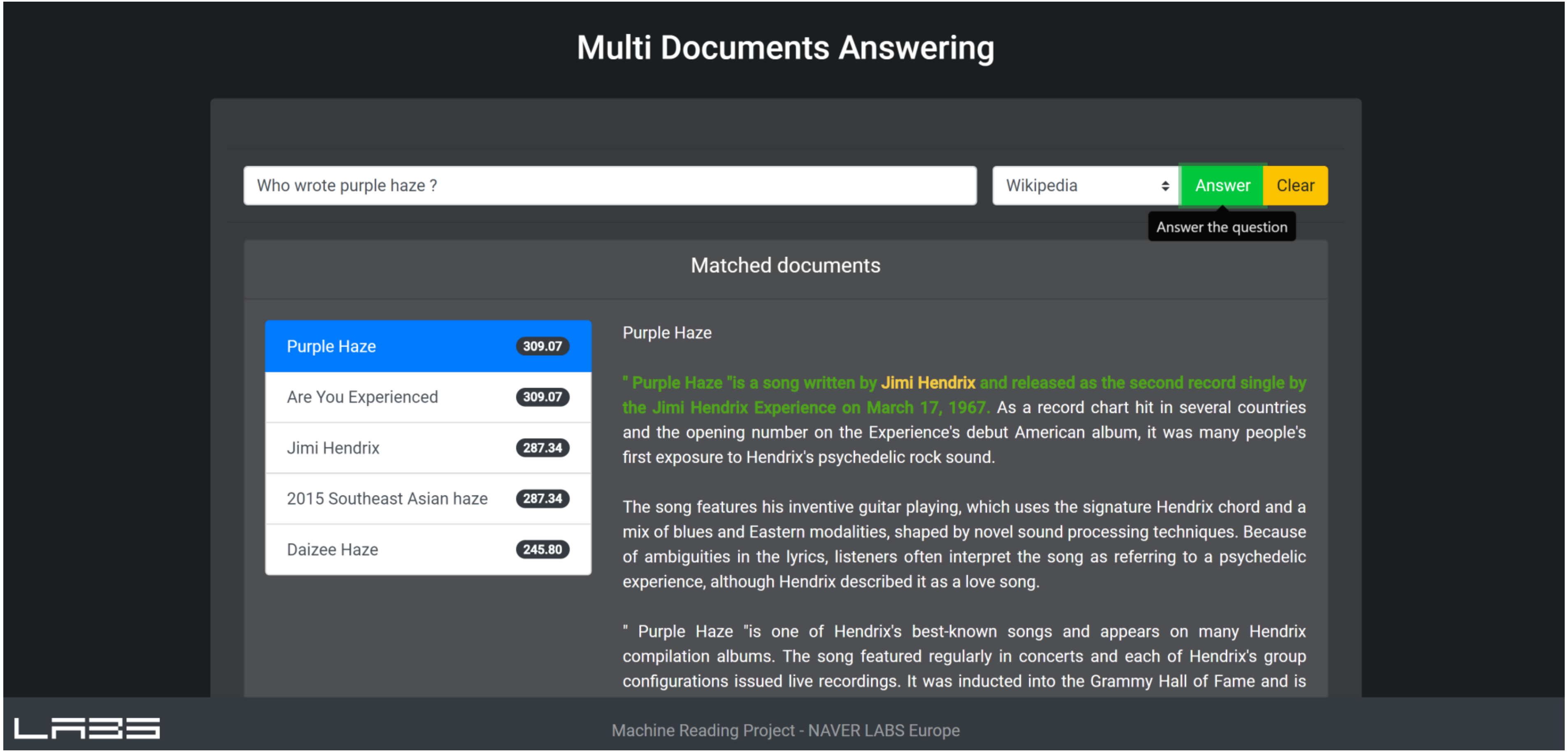
Blade Runner

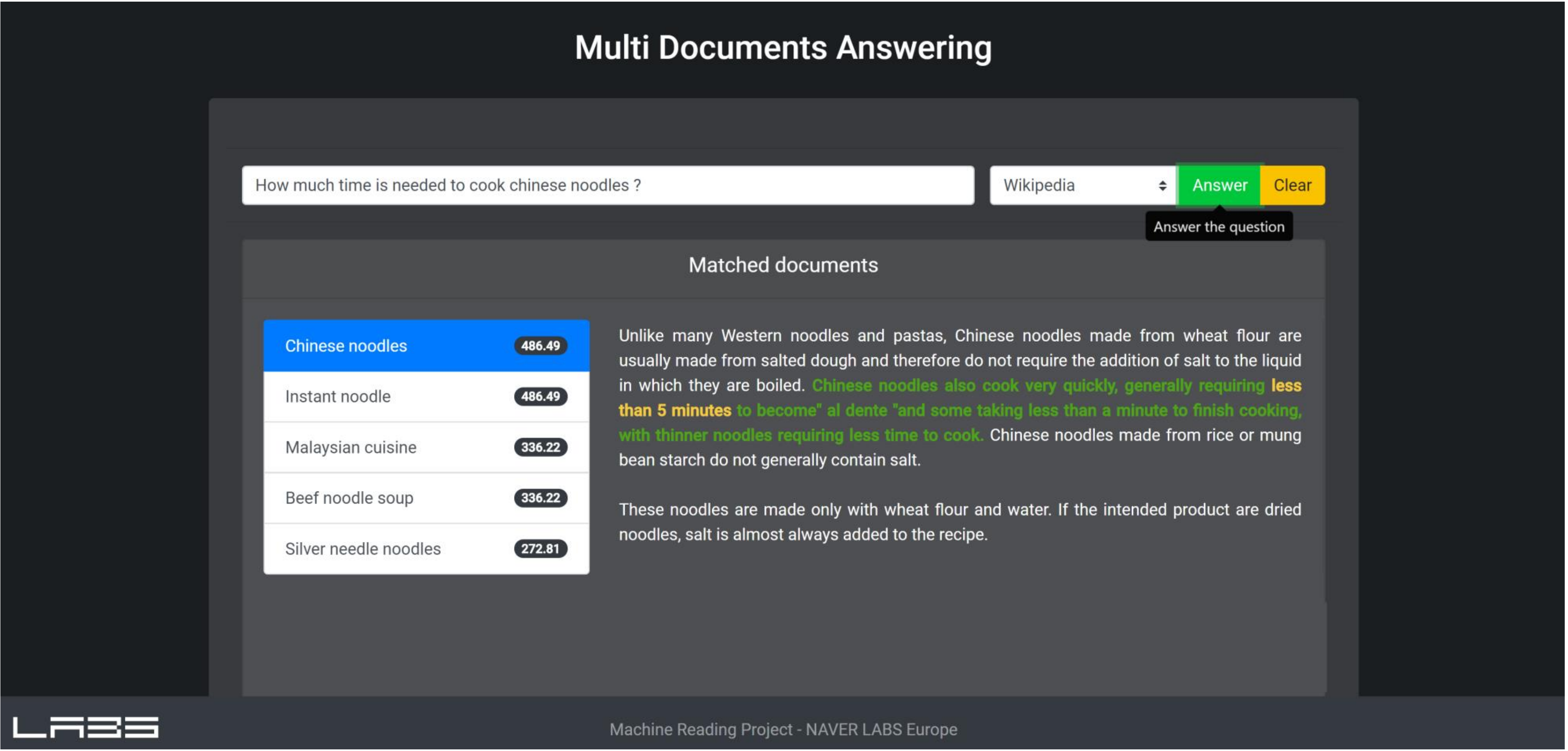
Blade Runner is a 1982 American neo-noir science fiction film directed by Ridley Scott, written by Hampton Fancher and David Peoples, and starring Harrison Ford, Rutger Hauer, Sean Young, and Edward James Olmos. It is a loose adaptation of Philip K. Dick's novel "Do Androids Dream of Electric Sheep?" (1968). The film is set in a dystopian future Los Angeles of 2019, in which synthetic humans known as replicants are bioengineered by the powerful **Tyrell Corporation** to work on off-world colonies. When a fugitive group of replicants led by Roy Batty (Hauer) escapes back to Earth, burnt-out cop Rick Deckard (Ford) reluctantly agrees to hunt them down.

" Blade Runner "initially underperformed in North American theaters and polarized critics; some praised its thematic complexity and visuals, while others were displeased with its unconventional pacing and plot. It later became an acclaimed cult film regarded as one of the all-time best science fiction movies. Hailed for its production design depicting a" retrofitted "future," Blade Runner "is a leading example of neo-noir cinema. The soundtrack, composed by Vangelis, was nominated in 1983 for a BAFTA and a Golden Globe as best original score.

The film has influenced many science fiction films, video games, anime, and television series. It brought the work of Philip K. Dick to the attention of Hollywood, and several later big-budget films were based on his work. In the year after its release," Blade Runner "won the Hugo Award for Best Dramatic Presentation, and in 1993 it was selected for preservation in

Machine Reading Project - NAVER LABS Europe


Multi Documents Answering
Who wrote purple haze ?
Wikipedia
Answer
Clear
Answer the question
Matched documents
Purple Haze 309.07
Are You Experienced 309.07
Jimi Hendrix 287.34
2015 Southeast Asian haze 287.34
Daizee Haze 245.80
Purple Haze
" Purple Haze "is a song written by Jimi Hendrix and released as the second record single by the Jimi Hendrix Experience on March 17, 1967. As a record chart hit in several countries and the opening number on the Experience's debut American album, it was many people's first exposure to Hendrix's psychedelic rock sound.
The song features his inventive guitar playing, which uses the signature Hendrix chord and a mix of blues and Eastern modalities, shaped by novel sound processing techniques. Because of ambiguities in the lyrics, listeners often interpret the song as referring to a psychedelic experience, although Hendrix described it as a love song.
" Purple Haze "is one of Hendrix's best-known songs and appears on many Hendrix compilation albums. The song featured regularly in concerts and each of Hendrix's group configurations issued live recordings. It was inducted into the Grammy Hall of Fame and is
Machine Reading Project - NAVER LABS Europe


Multi Documents Answering
How much time is needed to cook chinese noodles ?
Wikipedia
Answer
Clear
Answer the question
Matched documents
Chinese noodles 486.49
Instant noodle 486.49
Malaysian cuisine 336.22
Beef noodle soup 336.22
Silver needle noodles 272.81
Unlike many Western noodles and pastas, Chinese noodles made from wheat flour are usually made from salted dough and therefore do not require the addition of salt to the liquid in which they are boiled. Chinese noodles also cook very quickly, generally requiring less than 5 minutes to become" al dente "and some taking less than a minute to finish cooking, with thinner noodles requiring less time to cook. Chinese noodles made from rice or mung bean starch do not generally contain salt.
These noodles are made only with wheat flour and water. If the intended product are dried noodles, salt is almost always added to the recipe.
Machine Reading Project - NAVER LABS Europe

Multi Documents Answering

When CNRS was founded ?

Wikipedia | Answer | Clear

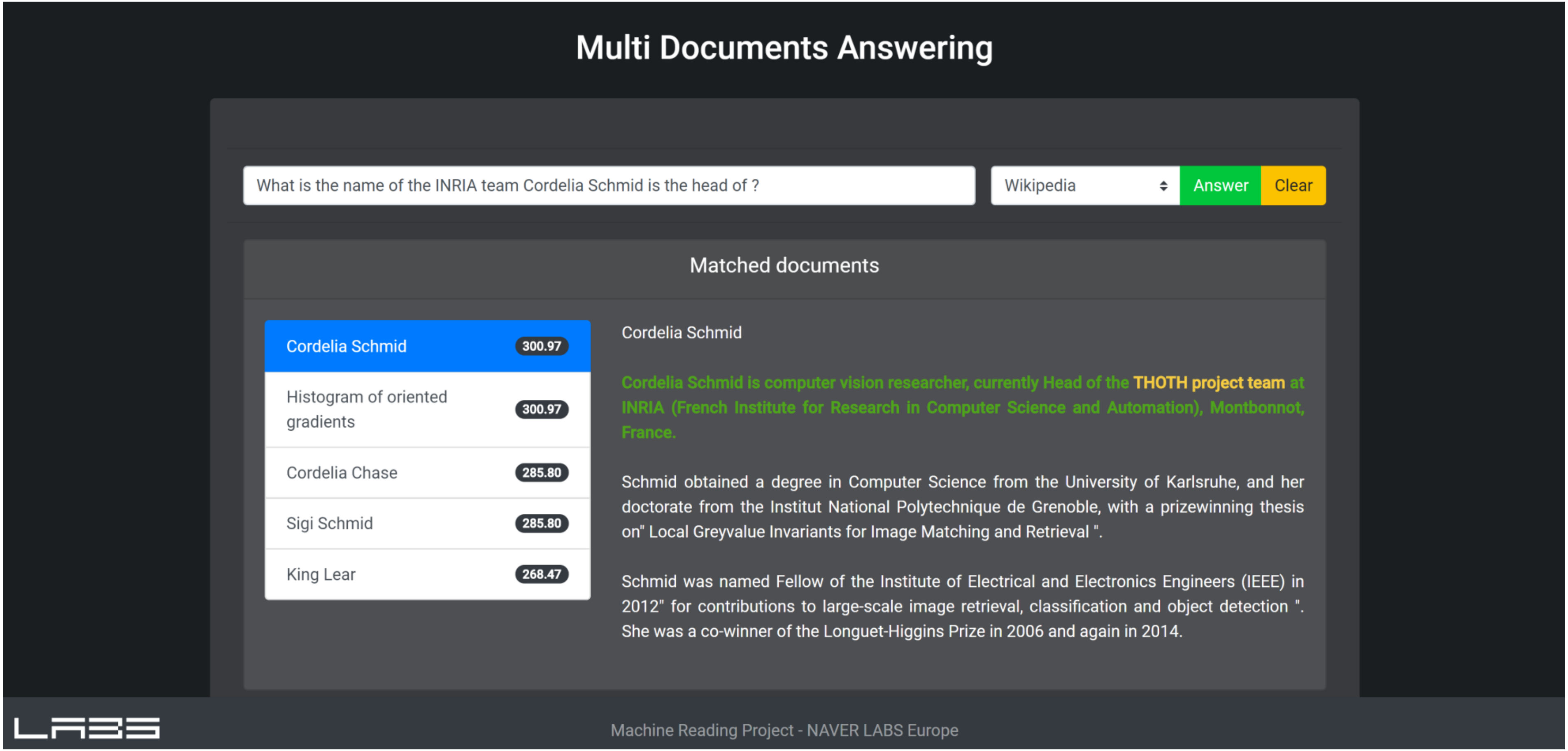
Matched documents

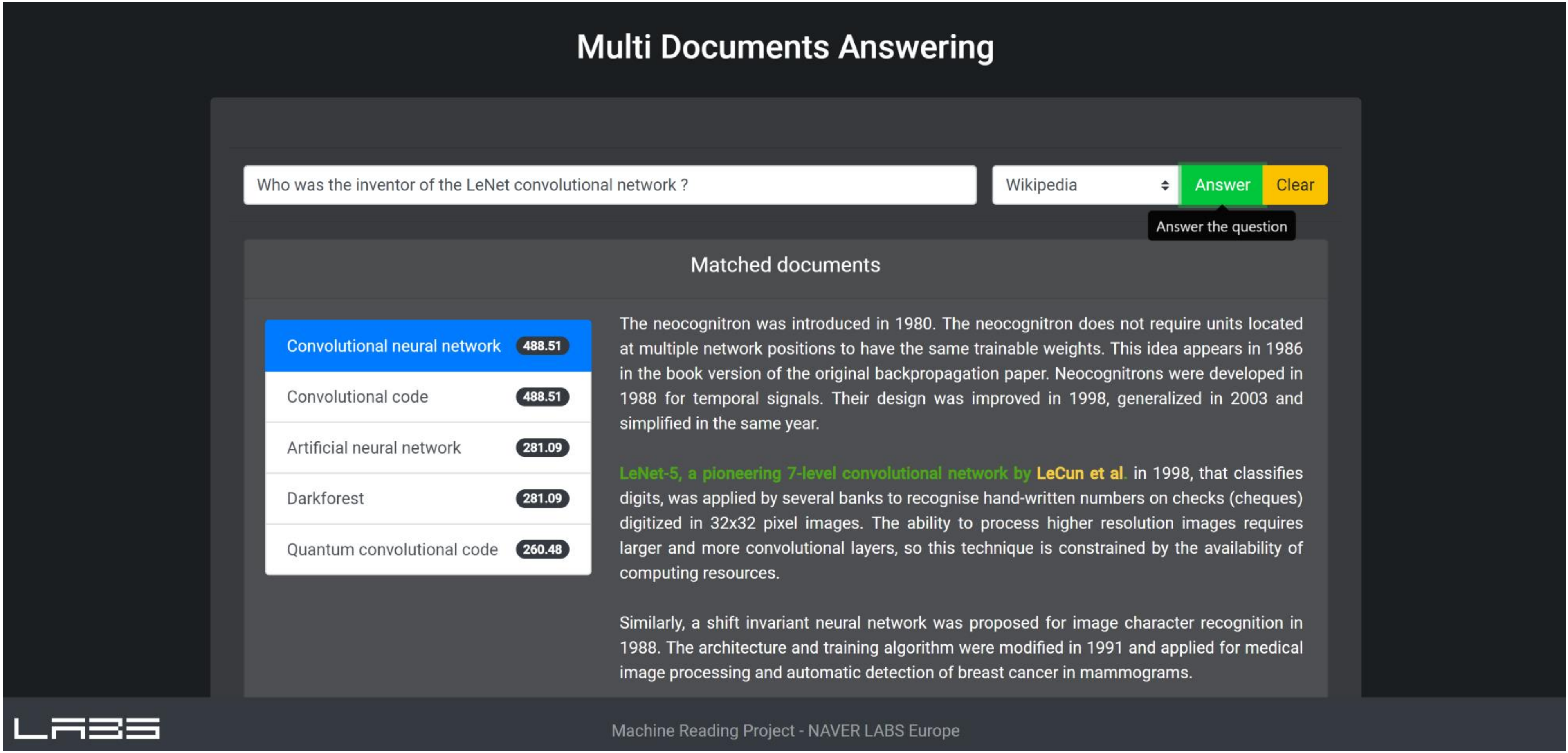
Institut Charles Sadron	152.27
Centre national de la recherche scientifique	152.27
Human and Social Sciences Library Paris Descartes-CNRS	140.11
Christian Cambillau	140.11
Rhodia (company)	124.45

Centre national de la recherche scientifique

All permanent support employees are recruited through annual nationwide competitive campaigns. Following a 1983 reform, the candidates selected have the status of civil servants and are part of the public service.

The CNRS was created on **19 October 1939** by decree of President Albert Lebrun. Since 1954, the centre has annually awarded gold, silver, and bronze medals to French scientists and junior researchers. In 1966, the organisation underwent structural changes, which resulted in the creation of two specialised institutes: the National Astronomy and Geophysics Institute in 1967 (which became the National Institute of Sciences of the Universe in 1985) and the Institut national de physique nucléaire et de physique des particules (IN2P3; English: National Institute of Nuclear and Particle Physics) in 1971.


Multi Documents Answering
What is the name of the INRIA team Cordelia Schmid is the head of ?
Wikipedia
Answer
Clear
Matched documents
Cordelia Schmid 300.97
Histogram of oriented gradients 300.97
Cordelia Chase 285.80
Sigi Schmid 285.80
King Lear 268.47
Cordelia Schmid
Cordelia Schmid is computer vision researcher, currently Head of the THOTH project team at INRIA (French Institute for Research in Computer Science and Automation), Montbonnot, France.
Schmid obtained a degree in Computer Science from the University of Karlsruhe, and her doctorate from the Institut National Polytechnique de Grenoble, with a prizewinning thesis on" Local Greyvalue Invariants for Image Matching and Retrieval ".
Schmid was named Fellow of the Institute of Electrical and Electronics Engineers (IEEE) in 2012" for contributions to large-scale image retrieval, classification and object detection ". She was a co-winner of the Longuet-Higgins Prize in 2006 and again in 2014.
Machine Reading Project - NAVER LABS Europe


Multi Documents Answering
Who was the inventor of the LeNet convolutional network ?
Wikipedia
Answer
Clear
Answer the question
Matched documents
Convolutional neural network 488.51
Convolutional code 488.51
Artificial neural network 281.09
Darkforest 281.09
Quantum convolutional code 260.48
The neocognitron was introduced in 1980. The neocognitron does not require units located at multiple network positions to have the same trainable weights. This idea appears in 1986 in the book version of the original backpropagation paper. Neocognitrons were developed in 1988 for temporal signals. Their design was improved in 1998, generalized in 2003 and simplified in the same year.
LeNet-5, a pioneering 7-level convolutional network by LeCun et al. in 1998, that classifies digits, was applied by several banks to recognise hand-written numbers on checks (cheques) digitized in 32x32 pixel images. The ability to process higher resolution images requires larger and more convolutional layers, so this technique is constrained by the availability of computing resources.
Similarly, a shift invariant neural network was proposed for image character recognition in 1988. The architecture and training algorithm were modified in 1991 and applied for medical image processing and automatic detection of breast cancer in mammograms.
Machine Reading Project - NAVER LABS Europe

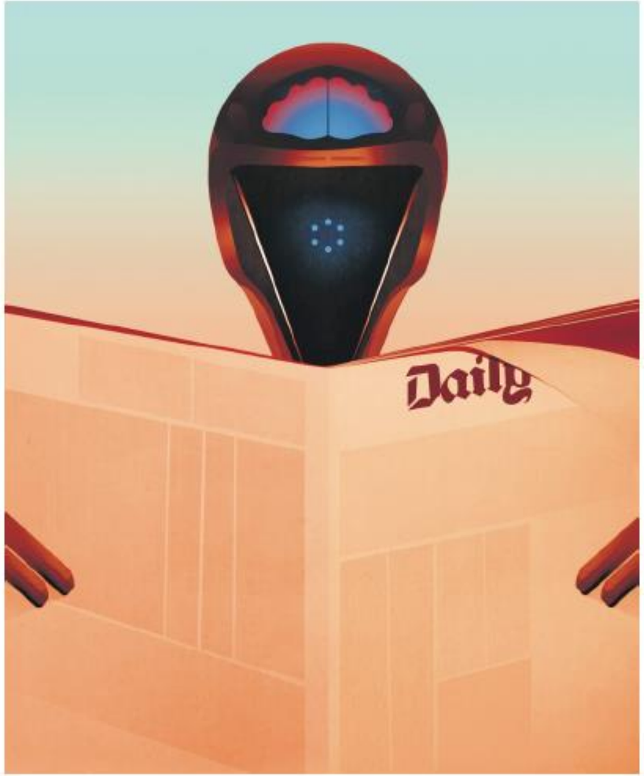
Content

1. Machine reading tasks
    - Definition
    - State of the art approaches
    - Dataset taxonomy

2. Models of reading

3. Applications

4. Open Questions



Courtesy of Phil Blunsom

Machine Reading
motivations

Human knowledge is (**mainly**) stored in natural language

Natural Language is an **efficient** support of knowledge transcription

Language is efficient because of its **contextuallity** that leads to **ambiguity**

Languages assume **apriori knowledge** of the world

The Library of Trinity College Dublin

Definition

"A machine comprehends **a passage of text** if, for any **question** regarding that text, it can be **answered** correctly by a majority of native speakers.

The machine needs to provide a string which human readers would agree both
1. Answers that question
2. Does not contain information irrelevant to that question." *(Burges, 2013)*

Applications

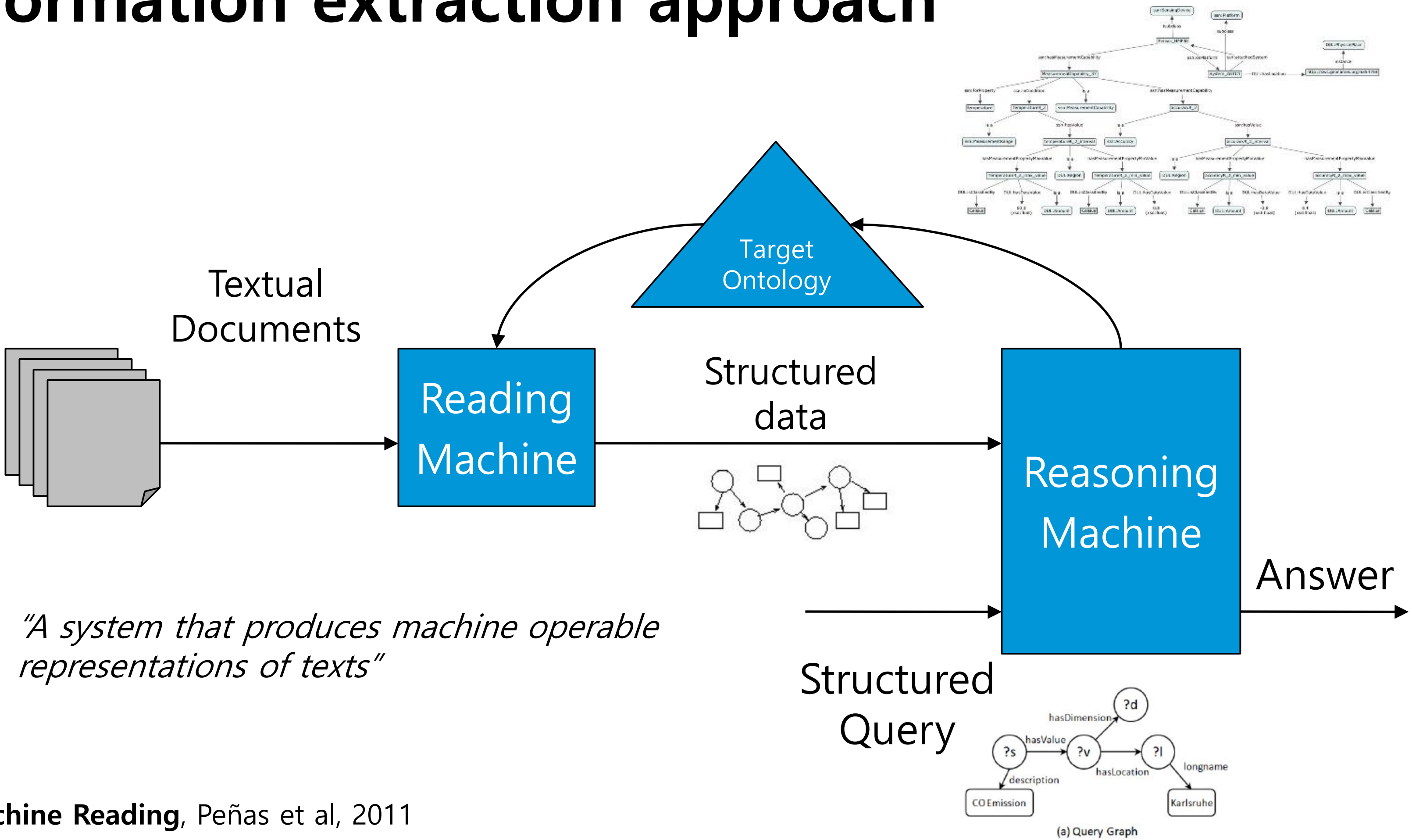
- Collection of documents as KB
- Social media mining
- Dialog understanding
- Fact checking – Fake news detection

Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

Information extraction approach



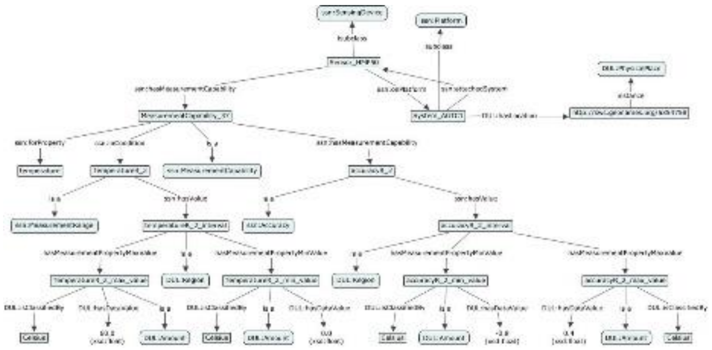
"A system that produces machine operable representations of texts"

[1] **Machine Reading**, Peñas et al, 2011

Information extraction approach

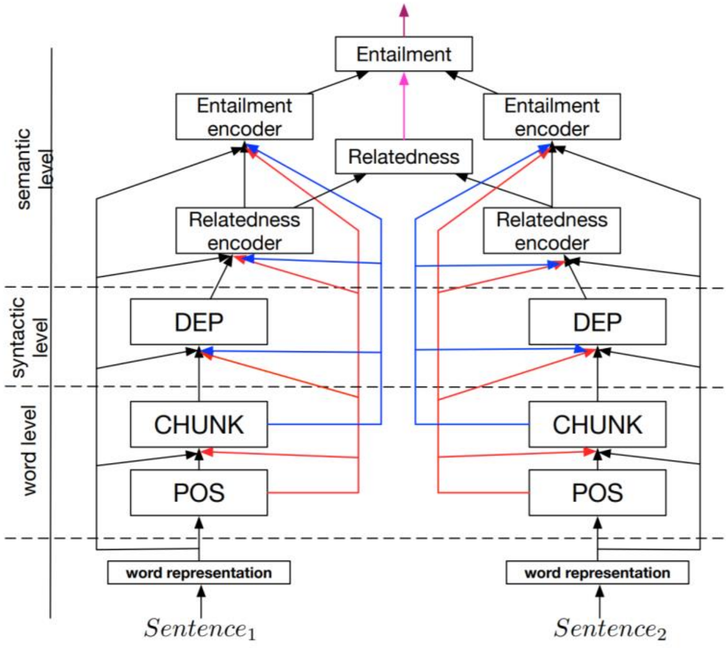
" A system that produces machine operable representations of texts "

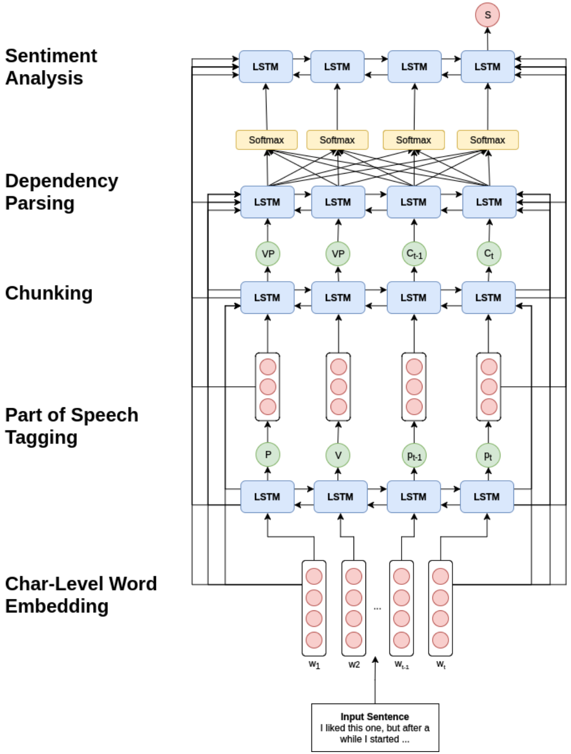
… but we have 3 problems here



1. *Fixed/Predefined ontologies*
2. *Fixed/Predefined lexical domain*
3. ***Data duplication by structuration***

Classic Deep NLP approach





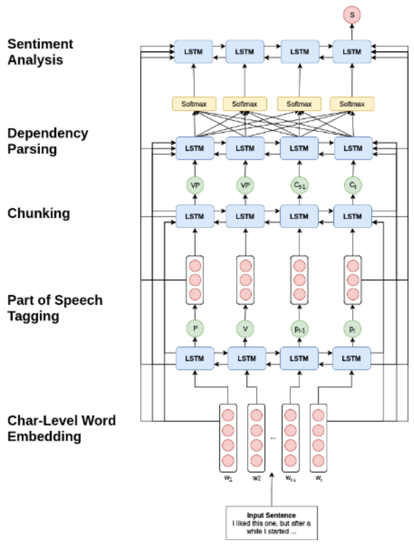
[2] **A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks**, Socher et al, 2017

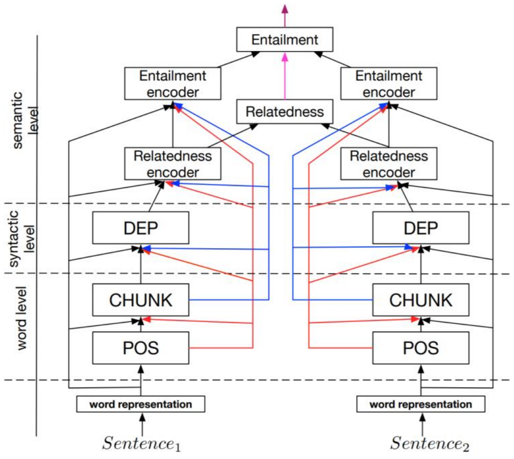
Classic Deep NLP approach

"Machine reading, yet another (Deep) NLP task ? "

… but we have 3 problems here

1. Is (Language dependant) syntax a requirement to semantics ?
2. Additional (unnecessary) requirement
   - Annotations
   - Priors
3. Not end-to end machine comprehension





Machine Reading

End-to-end formulation of natural language comprehension

Document

James was always getting in trouble. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. ***Then he walked to the fast food restaurant*** *and ordered 15 bags of fries. He didn't pay, and instead headed home.*

Question: Where did James go after he went to the grocery store?

- his deck
- his freezer
- a fast food restaurant
- his home

[3] **Teaching Machines to Read and Comprehend**, Blunsom et al, 2015
[4] **Text as knowledge bases,** Manning et al, 2016

Machine Reading

as Multi-choice question task

MCTest

- 500 passages
- 2000 questions about simple stories

RACE

- 28,000 passages
- 100,000 questions from English comprehension tests

[5] **MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text,** Richardson et al, 2013
[6] **RACE: Large-scale ReAding Comprehension Dataset From Examinations,** Lai et al, 2017

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.
One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.
His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.
After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

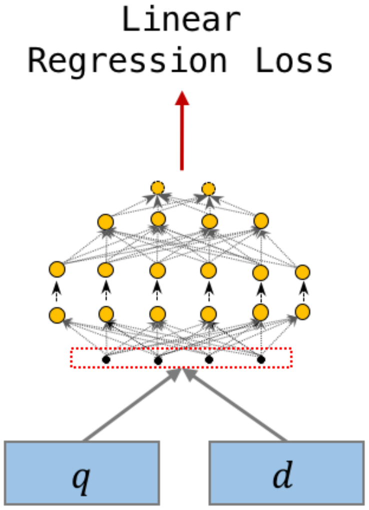
2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

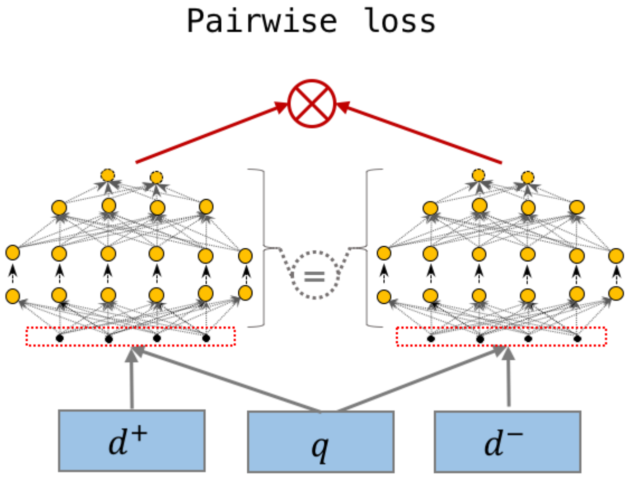
4) What did James do after he ordered the fries?
A) went to the grocery store
B) went home without paying
C) ate them
D) made up his mind to be a better turtle

Machine Reading

as Multi-choice question task



$$\mathcal{L}(b;\theta)=\frac{1}{|b|}\sum_{i=1}^{|b|}(\mathcal{S}(\{q,d\}_i;\theta)-s_{\{q,d\}_i})^2$$



$$\mathcal{L}(b;\theta)=\frac{1}{|b|}\sum_{i=1}^{|b|}\max\left\{0,\varepsilon-\ s_{\{q,d_1\}_i}-s_{\{q,d_2\}_i}\right\}$$

[5] **MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text,** Richardson et al, 2013

[6] **RACE: Large-scale ReAding Comprehension Dataset From Examinations,** Lai et al, 2017

Machine Reading

as Cloze style queries

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

Figure 1: **A Named Entity question from the CBT** (right), created from a book passage (left, in blue). In this case, the candidate answers C are both entities and common nouns, since fewer than ten named entities are found in the context.

Machine Reading
as Cloze style queries

	CNN			Daily Mail			CBT CN			CBT NE		
	train	valid	test	train	valid	test	train	valid	test	train	valid	test
# queries	380,298	3,924	3,198	879,450	64,835	53,182	120,769	2,000	2,500	108,719	2,000	2,500
Max # options	527	187	396	371	232	245	10	10	10	10	10	10
Avg # options	26.4	26.5	24.5	26.5	25.5	26.0	10	10	10	10	10	10
Avg # tokens	762	763	716	813	774	780	470	448	461	433	412	424
Vocab. size	118,497			208,045			53,185			53,063		

Table 1: Statistics on the 4 data sets used to evaluate the model. CBT CN stands for CBT Common Nouns and CBT NE stands for CBT Named Entites. Statistics were taken from (Hermann et al., 2015) and the statistics provided with the CBT data set.

[8] **Teaching Machines to Read and Comprehend**, Blunsom et al, 2015

Machine Reading
as Span selection

SQuAD
- 500 passages
- 100,000 questions on Wikipedia text
- Human annotated

- **TriviaQA**
  - 95k questions
  - 650k evidence documents
  - distant supervision

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

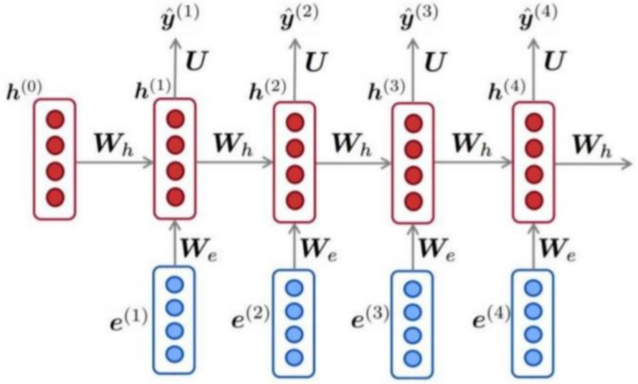
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

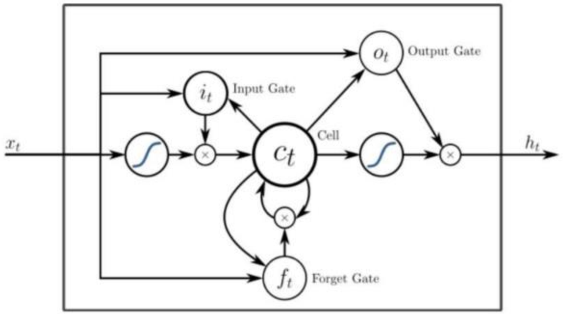
[9] **SQuAD: 100,000+ Questions for Machine Comprehension of Text,** Liang et al, 2016
[10] **TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension**, Zottlemoyer et al, 2017

Machine Reading
as Span selection

SQuAD
- 500 passages
- 100,000 questions on Wikipedia text
- Human annotated

- **TriviaQA**
  - 95k questions
  - 650k evidence documents
  - distant supervision





[9] **SQuAD: 100,000+ Questions for Machine Comprehension of Text,** Liang et al, 2016
[10] **TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension**, Zottlemoyer et al, 2017

Machine Reading

as Span selection

- 200k documents (~1M passages)
- 100k human generated questions
- Each query comes with approximately 10 passages


```
"passages": [
    {
        "url": "http://www.biography.com/people/ronald-reagan-9453198",

        "passage_text": "1984 Re-Election. In November 1984, Ronald Reagan was
            re-elected in a landslide, defeating Democratic challenger Walter
            Mondale. Reagan carried 49 of the 50 U.S. states in the election,
            and received 525 of 538 electoral votes—the largest number ever won
            by an American presidential candidate. "
    },
    {
        "url": "http://www.msnbc.com/the-last-word/watch/when-reagan-was-a
            -liberal-democrat-219696195576",

        "passage_text": "When Reagan was a liberal Democrat. In 1948, a very
            different sounding Ronald Reagan campaigned on the radio for
            Democrat Harry Truman. Listen to the old audio recording..."
    },

]

"query": "When was ronald reagan born ?",
"answer": "february 1911"
```


[11] **MS MARCO: A Human Generated MAchine Reading COmprehension Dataset**, Deng et al, 2016

Machine reading

Reasoning over knowledge extraction

- Textual data can specify reasoning capabilities

- **Goal**: build machines that can "understand " textual information, *i.e.* converting it into interpretable structured knowledge to be leveraged by humans and other machines alike.

- Optimized with categorical cross-entropy loss

$$CCE = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{J} y_j \cdot log(\hat{y}_j) + (1 - y_j) \cdot log(1 - \hat{y}_j)$$

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

Task 7: Counting
Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets
Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation
Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge
John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

[12] **Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks,** Weston and al

Machine Reading

Datasets

Before 2015:

- MCTest (Richardson et al, 2013): 2600 questions
- ProcessBank (Berant et al, 2014): 500 questions

More than 100k questions!

After 2015:

- **CNN/Daily Mail**
- Children Book Test
- WikiReading
- LAMBADA
- **SQuAD**
- Who did What
- Maluuba NewsQA
- MS MARCO
- NAVER DSTC6-T1

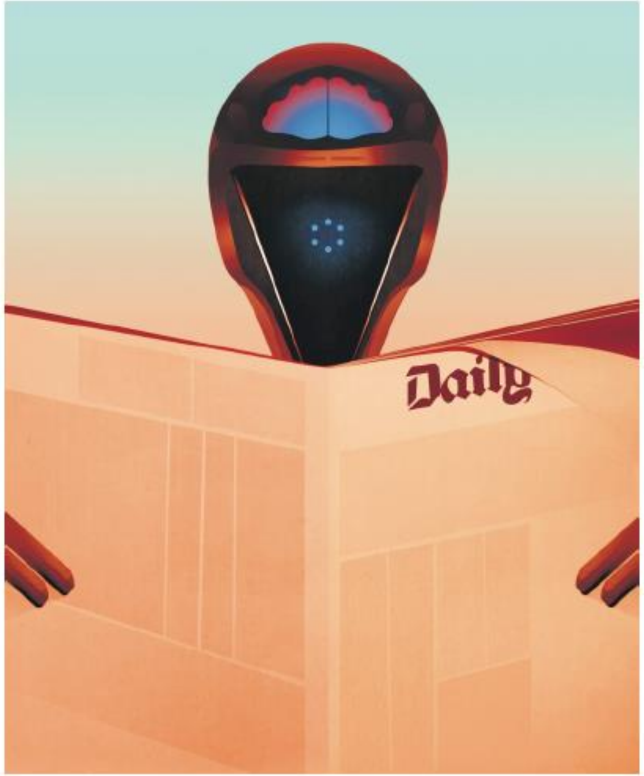
Content

1. Machine reading tasks

2. Models of reading
    1. Building blocks
    2. Retrieval models
    3. Reasoning models

3. Applications

4. Open Questions



Building blocks

Recurrent Neural Network

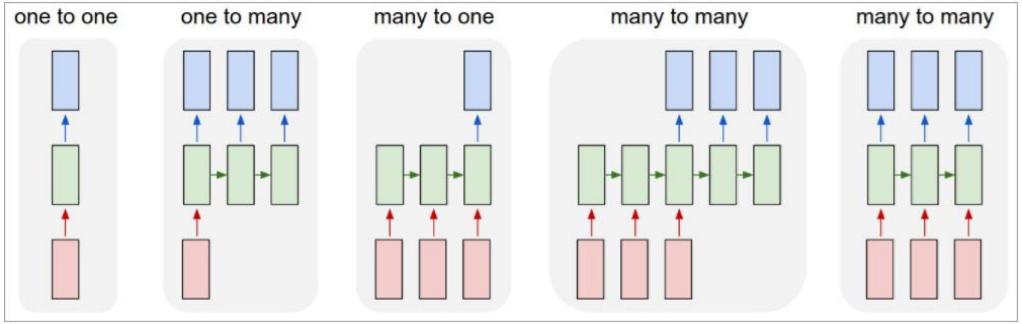
LSTM with a forget gate

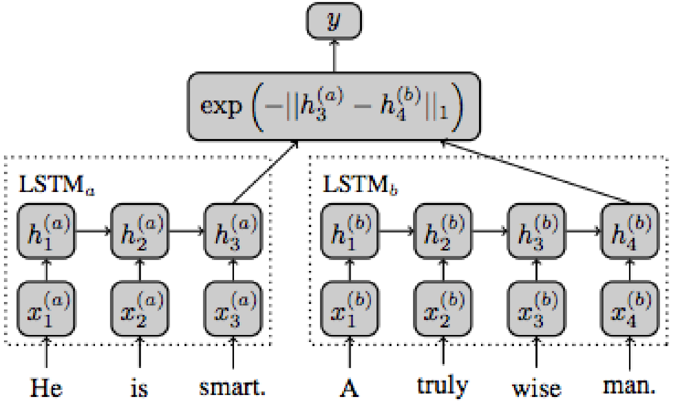
$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

where the initial values are $c_0 = 0$ and $h_0 = 0$
and the operator $\circ$ denotes the Hadamard product (entry-wise product).
The subscripts t refer to the time step.

Variables

- $x_t \in \mathbb{R}^d$: input vector to the LSTM unit
- $f_t \in \mathbb{R}^h$: forget gate's activation vector
- $i_t \in \mathbb{R}^h$: input gate's activation vector
- $o_t \in \mathbb{R}^h$: output gate's activation vector
- $h_t \in \mathbb{R}^h$: output vector of the LSTM unit
- $c_t \in \mathbb{R}^h$: cell state vector
- $W \in \mathbb{R}^{h\times d}$, $U \in \mathbb{R}^{h\times h}$ and $b \in \mathbb{R}^h$: weight matrices and bias vector parameters





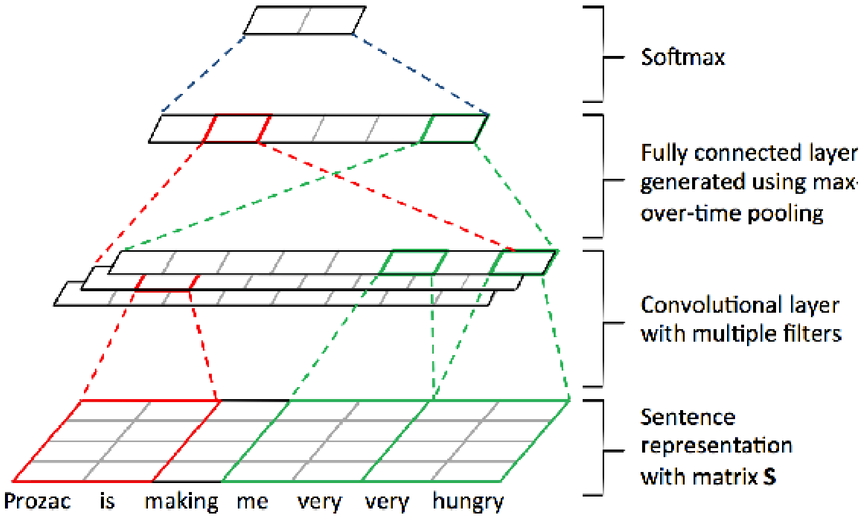
[13] **Long Short Term Memory**, Schmidhuber et al, 1997

Building blocks
Convolutional Network

Elements:

- Input sentence: $\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \ldots \oplus \mathbf{x}_n$
- Output local feature: $c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b)$
- Feature map: $\mathbf{c} = [c_1, c_2, \ldots, c_{n-h+1}]$
- Max-pooling layer
- Fully connected layer with softmax output for classification tasks

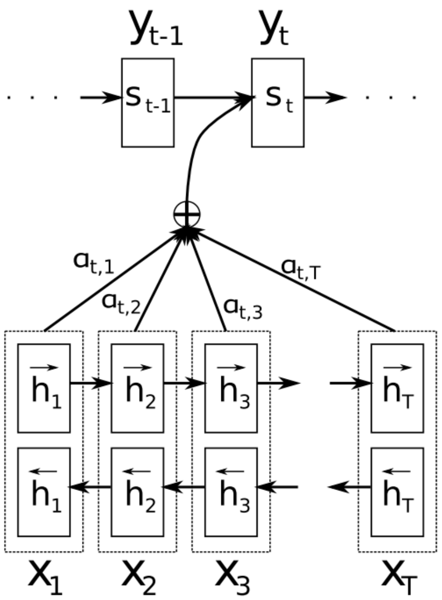
… Trivial to parallelize



[14] **Convolutional Neural Networks for Sentence Classification, Kim et al, 2017**

Building blocks

Attention mechanism



In Neural Machine Translation

- Encode each work in the input and output sentence into a vector
- Perform a linear combination of these vectors, weighted by « **attention score** »
- Use this combination as support to pick the next word

$$\alpha_{ts} = \frac{\exp\left(\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'=1}^{S} \exp\left(\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \qquad \text{[Attention weights]} \qquad (1)$$

$$\boldsymbol{c}_t = \sum_s \alpha_{ts} \bar{\boldsymbol{h}}_s \qquad \text{[Context vector]} \qquad (2)$$

$$\boldsymbol{a}_t = f(\boldsymbol{c}_t, \boldsymbol{h}_t) = \tanh(\boldsymbol{W}_c[\boldsymbol{c}_t; \boldsymbol{h}_t]) \qquad \text{[Attention vector]} \qquad (3)$$

[15] **Neural Machine Translation by Jointly Learning to Align and Translate**, Badhanau et al, 2015

Building blocks

Attention mechanism

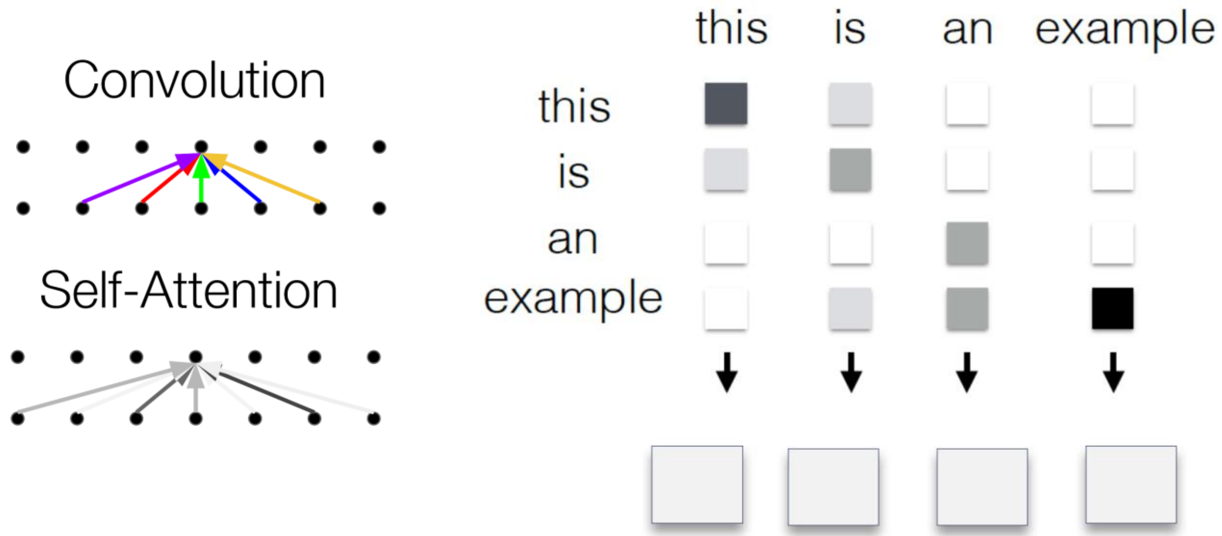
With q, a query and k, a key

			Reference
Multi-layer Perceptron	$a(q,k) = \tanh(\mathcal{W}_1[q,k])$	Flexible, often very good with large data	Bahdanau et al., 2015
Bilinear	$a(q,k) = q^T\mathcal{W}k$		Luong et al 2015
Dot Product	$a(q,k) = q^T k$	No parameters! But requires sizes to be the same	Luong et al. 2015
Scaled Dot Product	$a(q,k) = \frac{q^T k}{\sqrt{\lvert k \rvert}}$	Scale by size of the vector	Vaswani et al. 2017

Building blocks

Self-Attention mechanism

Each element in the sentence attends to other elements from the SAME sentence → context sensitive encodings!

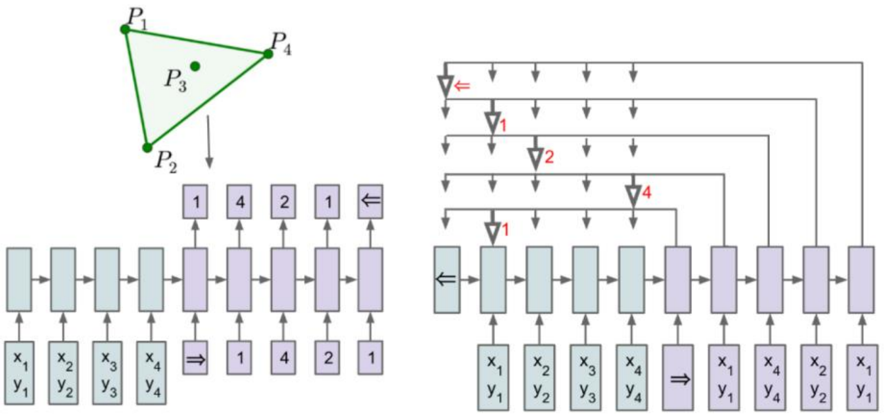


[16] **Attention Is All You Need**, Polosukhin et al, 2017

Building blocks

Pointer Networks

- Pointer networks are a variation of the seq-to-seq models.
- Instead of translating one sequence into another, the output is a sequence of pointers to the elements of the input series (i.e a permutation of the input sequence)



(a) Sequence-to-Sequence (b) Ptr-Net

[17] **Pointer Networks**, Vinyals et al, 2015

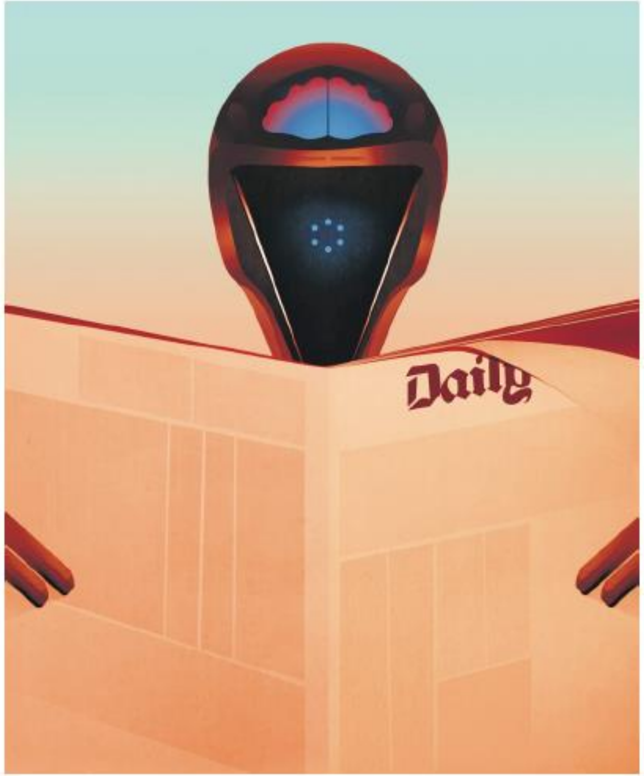
Content

1. Machine reading tasks

2. Models of reading
   1. Building blocks
   2. Retrieval models
   3. Reasoning models

3. Applications

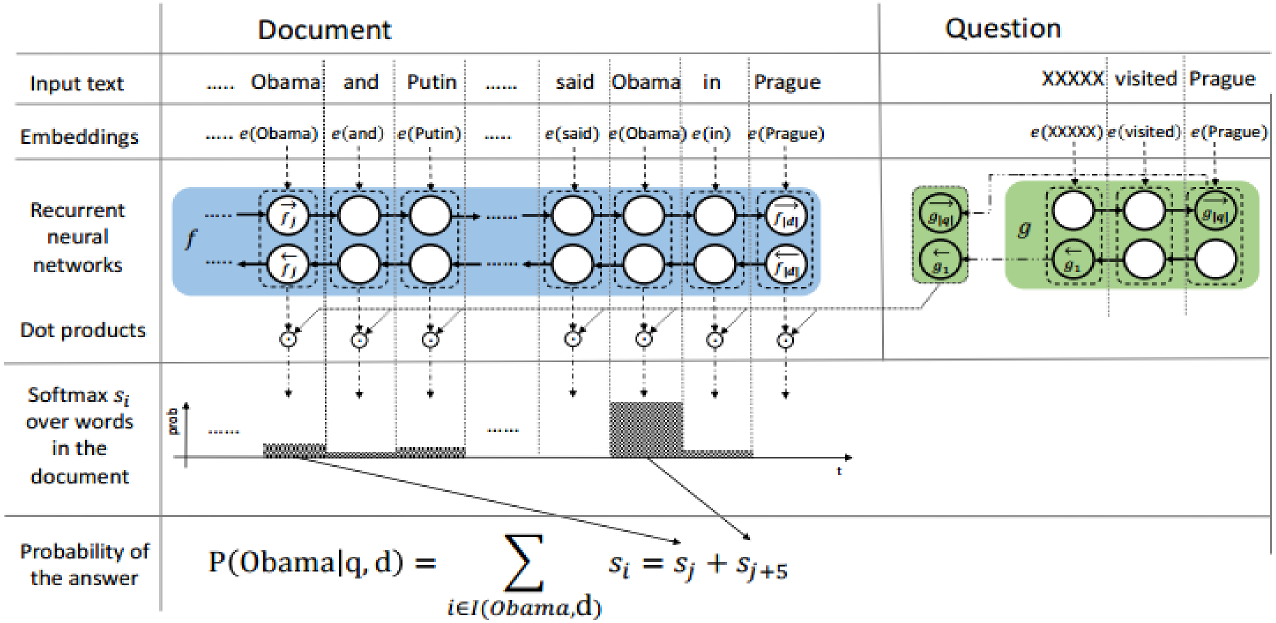
4. Open Questions



Courtesy of Phil Blunsom

Extractive models

Attention Sum Reader Network



$$s_i \propto \exp\left(f_i(\mathbf{d}) \cdot g(\mathbf{q})\right) \qquad (1)$$

$$P(w|\mathbf{q},\mathbf{d}) \propto \sum_{i \in I(w,\mathbf{d})} s_i \qquad (2)$$

where $I(w,\mathbf{d})$ is a set of positions where w appears in the document $\mathbf{d}$.

$$f_i(\mathbf{d}) = \overrightarrow{f_i}(\mathbf{d}) \,||\, \overleftarrow{f_i}(\mathbf{d}),$$

$$g(\mathbf{q}) = \overrightarrow{g_{|\mathbf{q}|}}(\mathbf{q}) \,||\, \overleftarrow{g_1}(\mathbf{q}).$$

Extractive models

Deep Long Short Term Memory readers

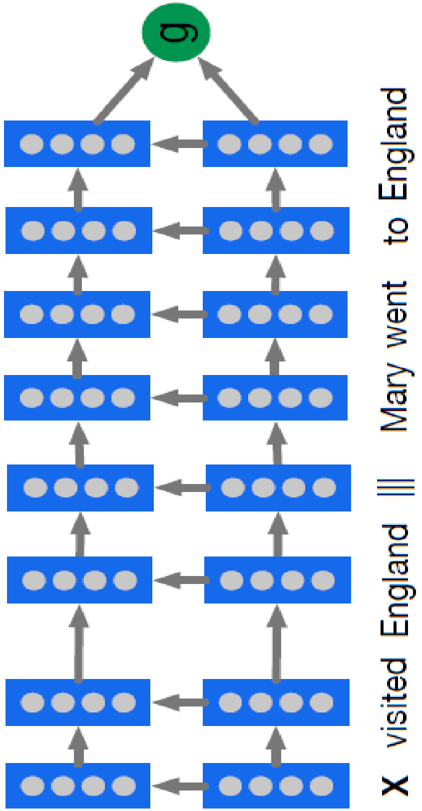
We employ a Deep LSTM cell with skip connections,

$$
\begin{aligned}
x'(t,k) &= x(t)||y'(t,k-1),\\
i(t,k) &= \sigma\left(W_{kxi}x'(t,k) + W_{khi}h(t-1,k) + W_{kci}c(t-1,k) + b_{ki}\right),\\
f(t,k) &= \sigma\left(W_{kxf}x(t) + W_{khf}h(t-1,k) + W_{kcf}c(t-1,k) + b_{kf}\right),\\
c(t,k) &= f(t,k)c(t-1,k) + i(t,k)\tanh\left(W_{kxc}x'(t,k) + W_{khc}h(t-1,k) + b_{kc}\right),\\
o(t,k) &= \sigma\left(W_{kxo}x'(t,k) + W_{kho}h(t-1,k) + W_{kco}c(t,k) + b_{ko}\right),\\
h(t,k) &= o(t,k)\tanh\left(c(t,k)\right),\\
y'(t,k) &= W_{ky}h(t,k) + b_{ky},\\
y(t) &= y'(t,1)||\ldots||y'(t,K),
\end{aligned}
$$

where $||$ indicates vector concatenation $h(t,k)$ is the hidden state for layer k at time t, and i, f, o are the input, forget, and output gates respectively.

$$g^{\mathrm{LSTM}}(d,q) = y(|d| + |q|)$$

with input $x(t)$ the concatenation of d and q separated by the delimiter $|||$.



[19] **Teaching Machines to Read and Comprehend**, Blunsom et al, 2015

Extractive models

Deep Long Short Term Memory readers

Denote the outputs of a bidirectional LSTM as $\overrightarrow{y}(t)$ and $\overleftarrow{y}(t)$. Form two encodings, one for the query and one for each token in the document,

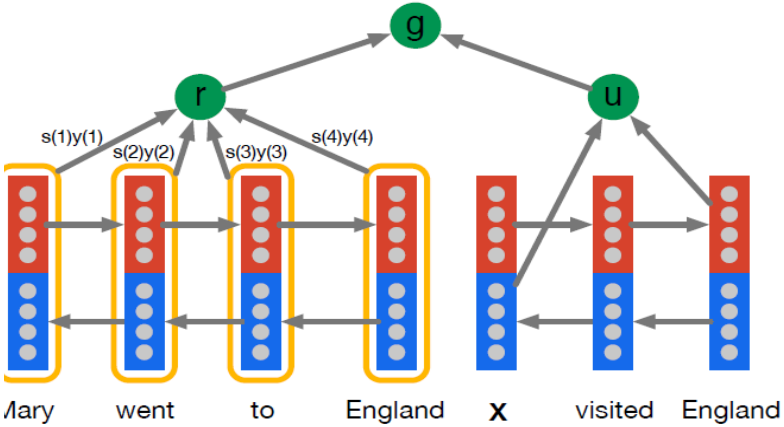
$$u = \overrightarrow{y_q}(|q|) \ || \ \overleftarrow{y_q}(1), \qquad y_d(t) = \overrightarrow{y_d}(t) \ || \ \overleftarrow{y_d}(t).$$

The representation r of the document d is formed by a weighted sum of the token vectors. The weights are interpreted as the model's attention,

$$m(t) = \tanh\left(W_{ym} y_d(t) + W_{um} u\right),$$
$$s(t) \propto \exp\left(\mathrm{w}_{ms}^{\intercal} m(t)\right),$$
$$r = y_d s.$$

Define the joint document and query embedding via a non-linear combination:

$$g^{AR}(d, q) = \tanh\left(W_{rg} r + W_{ug} u\right).$$



[19] **Teaching Machines to Read and Comprehend**, Blunsom et al, 2015

Extractive models

results

	valid	test	valid	test
Attentive Reader †	61.6	63.0	70.5	69.0
Impatient Reader †	61.8	63.8	69.0	68.0
MemNNs (single model) ‡	63.4	66.8	NA	NA
MemNNs (ensemble) ‡	66.2	69.4	NA	NA
Dynamic Entity Repres. (max-pool) ♯	71.2	70.7	NA	NA
Dynamic Entity Repres. (max-pool + byway)♯	70.8	72.0	NA	NA
Dynamic Entity Repres. + w2v ♯	71.3	72.9	NA	NA
Chen et al. (2016) (single model)	72.4	72.4	76.9	75.8
AS Reader (single model)	68.6	69.5	75.0	73.9
AS Reader (avg for top 20%)	68.4	69.9	74.5	73.5
AS Reader (avg ensemble)	73.9	**75.4**	78.1	77.1
AS Reader (greedy ensemble)	74.5	74.8	78.7	**77.7**

Table 2: Results of our AS Reader on the CNN and Daily Mail datasets. Results for models marked with † are taken from (Hermann et al., 2015), results of models marked with ‡ are taken from (Hill et al., 2015) and results marked with ♯ are taken from (Kobayashi et al., 2016). Performance of ‡ and ♯ models was evaluated only on CNN dataset.

	Named entity		Common noun	
	valid	test	valid	test
Humans (query) (*)	NA	52.0	NA	64.4
Humans (context+query) (*)	NA	***81.6***	NA	***81.6***
LSTMs (context+query) ‡	51.2	41.8	62.6	56.0
MemNNs (window memory + self-sup.) ‡	70.4	66.6	64.2	63.0
AS Reader (single model)	73.8	68.6	68.8	63.4
AS Reader (avg for top 20%)	73.3	68.4	67.7	63.2
AS Reader (avg ensemble)	74.5	70.6	71.1	**68.9**
AS Reader (greedy ensemble)	76.2	**71.0**	72.4	67.5

Table 3: Results of our AS Reader on the CBT datasets. Results marked with ‡ are taken from (Hill et al., 2015). (*)Human results were collected on 10% of the test set.

Extractive models

R-Net

- Extractive model
- Fully differentiable
- Based on 4 stacked layers
- Language independent

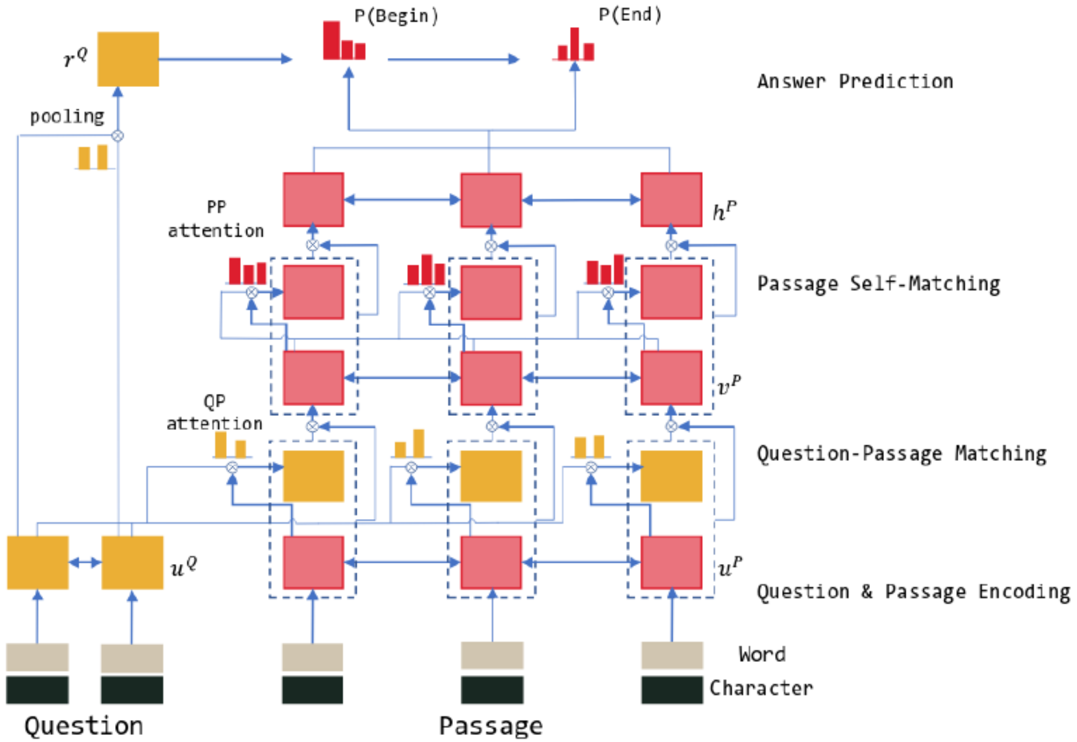
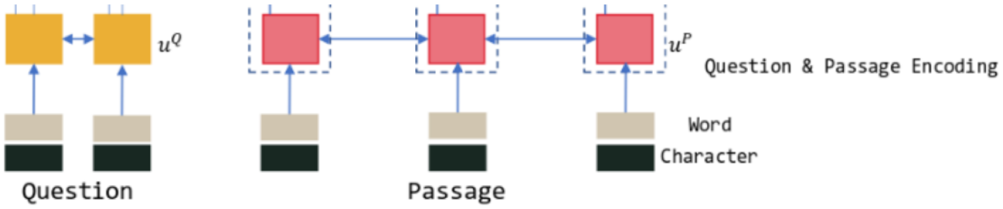


Figure 1: R-NET structure overview.

[20] **R-Net, technical report**, Microsoft Asia, 2017

Extractive models

R-Net – Question and Passage encoding



→ Let P = $\{w^P_1, \ldots, w^P_n\}$ be a document and Q = $\{w^Q_1, \ldots, w^Q_m\}$ a question regarding this passage.

→ First convert words to their word-level embeddings: $E_P = \{e^P_1, \ldots, e^P_n\}$ and $E_Q = \{e^Q_1, \ldots, e^Q_m\}$

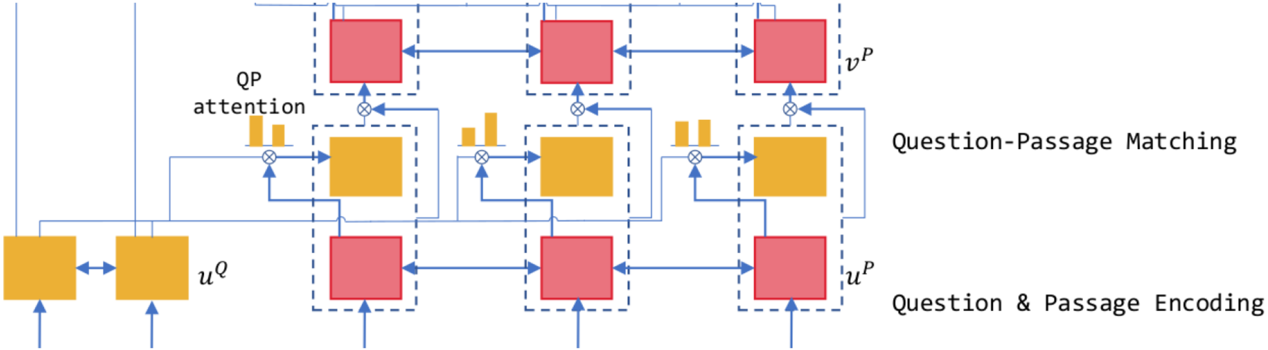
→ Generate character-level embeddings by taking the final states of a bidirectional RNN: $C_P = \{c^P_1, \ldots, c^P_n\}$ and $C_Q = \{c^Q_1, \ldots, c^Q_m\}$

→ Finnaly use a bidirectional RNN to produce $\mathbf{u^P}$ and $\mathbf{u^Q}$ the new representations of the passage and the question.

$$u_t^Q = \mathrm{BiRNN}_Q(u_{t-1}^Q, [e_t^Q, c_t^Q])$$
$$u_t^P = \mathrm{BiRNN}_P(u_{t-1}^P, [e_t^P, c_t^P])$$

Extractive models

R-Net - Question-Passage matching - Gated attention-based recurrent network



Objective: Incorporate question information into the passage representation

Solution: Attention-based RNN with an additional gate to determine the importance of information in the passage regarding a question

Extractive models

R-Net - Question-Passage matching - Gated attention-based recurrent network

From the question u^Q and a the document u^P, the model will compute a **question-aware representation** of the passage:

$$v^P_t = \mathrm{RNN}(\, v^P_{t-1}, [u^P_t, c_t]^* \,)$$

where $c_t = \mathrm{att}(\, u^Q, [u^P_t, v^P_{t-1}])$ is an attention-pooling vector of the whole question (u^Q)

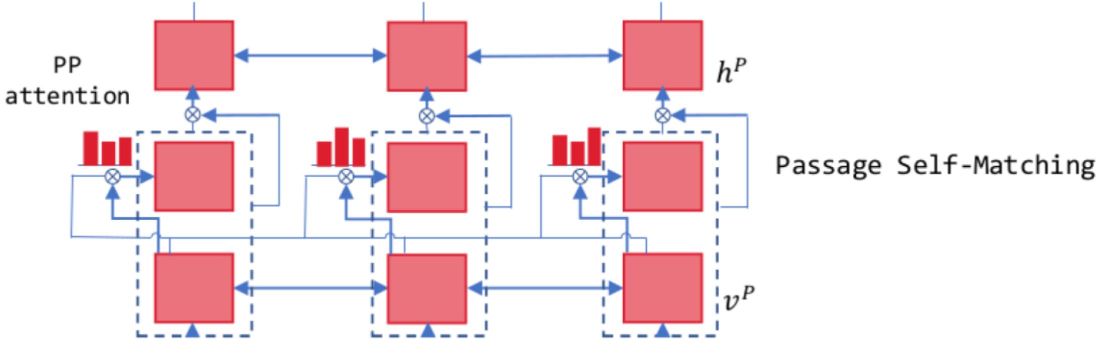
$$s_j^t = \mathrm{v}^{\mathrm{T}} \tanh(W_u^Q u_j^Q + W_u^P u_t^P + W_v^P v_{t-1}^P)$$
$$a_i^t = \exp(s_i^t) / \Sigma_{j=1}^m \exp(s_j^t)$$
$$c_t = \Sigma_{i=1}^m a_i^t u_i^Q$$

and $[u^P_t, c^P_t]^*$ a gated version of the input $[u^P_t, c^P_t]$

$$g_t = \mathrm{sigmoid}(W_g [u_t^P, c_t])$$
$$[u_t^P, c_t]^* = g_t \odot [u_t^P, c_t]$$

Extractive models

R-Net – Passage Self-Matching



Problem: Current representations v^P have a very limited knowledge of the context.

Solution: Match each token of the question-aware representation of the passage against the whole document

Extract evidence from the whole document according to the current passage word and question information

Extractive models

R-Net – Passage Self-Matching

From the question-aware representation of the passage (v^P), the model will compute a gated self-attention on it:

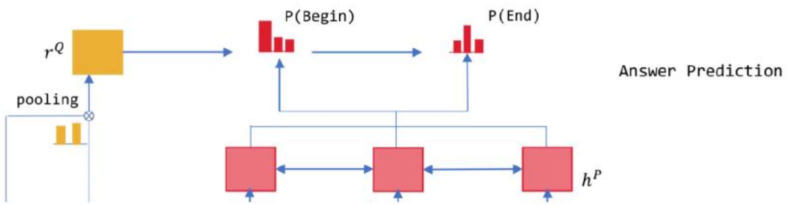
$$h^P_t = \text{BiRNN}(h^P_{t-1}, [v^P_t, c_t])$$

where $c_t = \text{att}(v^P, [u^P_t, v^P_{t-1}])$ is an attention-pooling vector of the whole passage (v^P)

$$s_j^t = \mathrm{v}^{\mathrm{T}}\tanh(W_v^P v_j^P + W_v^{\tilde{P}} v_t^P)$$
$$a_i^t = \exp(s_i^t)/\Sigma_{j=1}^n \exp(s_j^t)$$
$$c_t = \Sigma_{i=1}^n a_i^t v_i^P$$

Extractive models

R-Net – Output layer - Pointer Network



A **pointer network** will predict the start and end position of the answer.

The question vector is used as the initial state of the answer pointer network

Let (i,j) be the ground-truth of the start and end position of a question regarding a document.

Let yp^s_i and yp^e_j be the predicted probabilities of the word i to be the start of the answer and j the end of the answer.

Then the loss is defined as the sum of the predicted log probabilities of the gound-truth start and end position :

$$L = -\sum_N \log(yp_i^s) + \log(yp_j^e)$$

Extractive models

Performances on SQuAD and MsMARCO

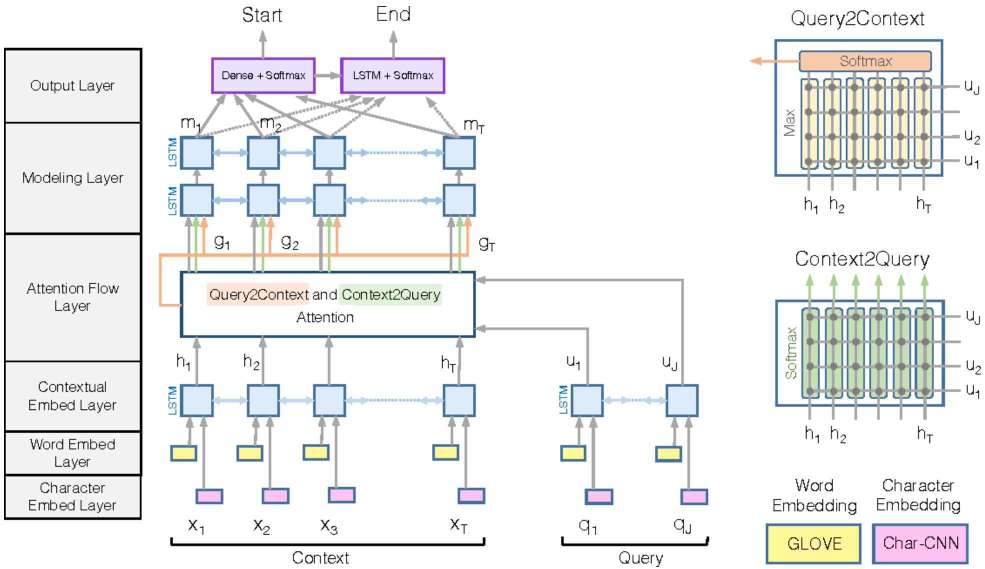
Results:

- State of the art model when the paper was published, in May 2017 on the SQuAD dataset
- Currently in the top 3
- State of the art on MS-MARCO

	Dev Set	Test Set
Single model	**EM / F1**	**EM / F1**
LR Baseline (Rajpurkar et al., 2016)	40.0 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.2	62.5 / 71.0
Attentive CNN context with LSTM (NLPR, CASIA)	- / -	63.3 / 73.5
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016b)	64.1 / 73.9	64.7 / 73.7
Dynamic Coattention Networks (Xiong et al., 2016)	65.4 / 75.6	66.2 / 75.9
Iterative Coattention Network (Fudan University)	- / -	67.5 / 76.8
FastQA (Weissenborn et al., 2017)	- / -	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
T-gating (Peking University)	- / -	68.1 / 77.6
RaSoR (Lee et al., 2016)	- / -	69.6 / 77.7
SEDT+BiDAF (Liu et al., 2017)	- / -	68.5 / 78.0
Multi-Perspective Matching (Wang et al., 2016)	- / -	70.4 / 78.8
FastQAExt (Weissenborn et al., 2017)	- / -	70.8 / 78.9
Mnemonic Reader (NUDT & Fudan University)	- / -	69.9 / 79.2
Document Reader (Chen et al., 2017)	- / -	70.7 / 79.4
ReasoNet (Shen et al., 2016)	- / -	70.6 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	- / -	70.6 / 79.5
jNet (Zhang et al., 2017)	- / -	70.6 / 79.8
Interactive AoA Reader (Joint Laboratory of HIT and iFLYTEK Research)	- / -	71.2 / 79.9
R-NET (Wang et al., 2017)	**71.1 / 79.5**	**71.3 / 79.7**
R-NET (March 2017)	**72.3 / 80.6**	**72.3 / 80.7**

Extractive models

Bidirectional Attention Flow for Machine Comprehension

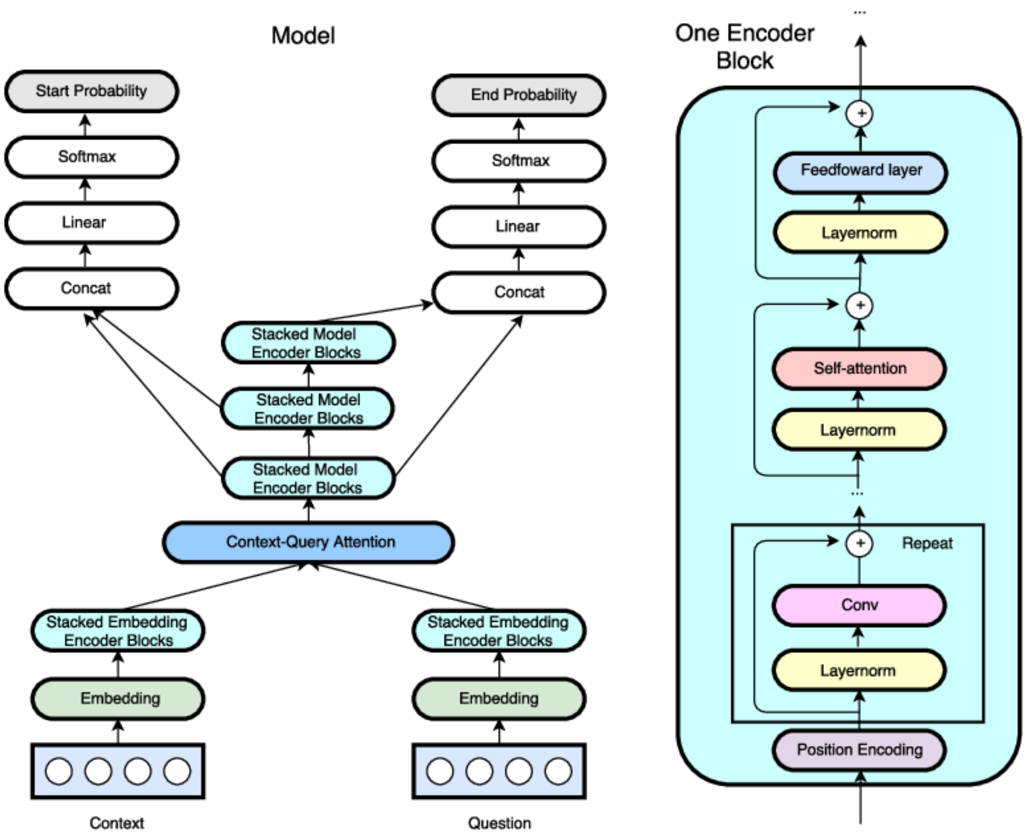


	CNN		DailyMail	
	val	test	val	test
Attentive Reader (Hermann et al., 2015)	61.6	63.0	70.5	69.0
MemNN (Hill et al., 2016)	63.4	6.8	-	-
AS Reader (Kadlec et al., 2016)	68.6	69.5	75.0	73.9
DER Network (Kobayashi et al., 2016)	71.3	72.9	-	-
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	-	-
EpiReader (Trischler et al., 2016)	73.4	74.0	-	-
Stanford AR (Chen et al., 2016)	73.8	73.6	77.6	76.6
GAReader (Dhingra et al., 2016)	73.0	73.8	76.7	75.7
AoA Reader (Cui et al., 2016)	73.1	74.4	-	-
ReasoNet (Shen et al., 2016)	72.9	74.7	77.6	76.6
BIDAF (Ours)	**76.3**	**76.9**	**80.3**	**79.6**
MemNN* (Hill et al., 2016)	66.2	69.4	-	-
ASReader* (Kadlec et al., 2016)	73.9	75.4	78.7	77.7
Iterative Attention* (Sordoni et al., 2016)	74.5	75.7	-	-
GA Reader* (Dhingra et al., 2016)	76.4	77.4	79.1	78.1
Stanford AR* (Chen et al., 2016)	77.2	77.6	80.2	79.2

[21] **Bidirectional Attention Flow for Machine Comprehension**, Seo et al, 2016

Extractive models
Google QANet

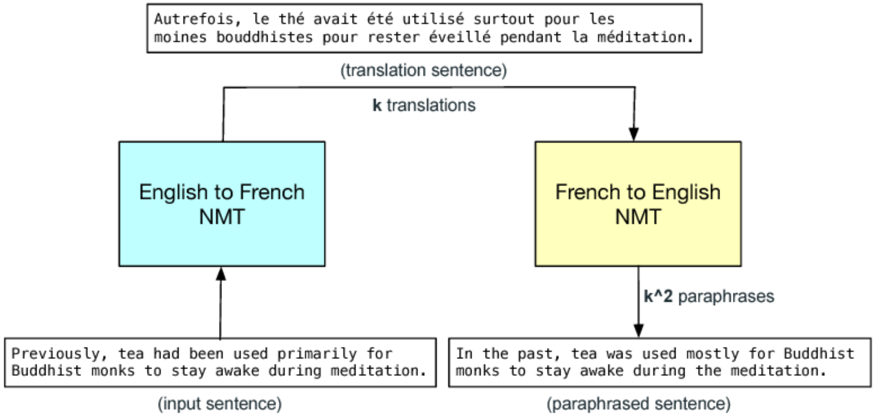
- Extractive model
- Fully differentiable
- Non-autoregressive model
- Language independant
- « Attention is All you Need »



[22] **Combining Local Convolution with Global Self-Attention for Reading Comprehension**, Google Research, 2017

Extractive models

Google QANet – Data augmentation with backtranslation



	EM / F1	Difference to Base Model EM / F1
Base Model	73.6 / 82.7	
- convolution in encoders	70.8 / 80.0	-2.8 / -2.7
- self-attention in encoders	72.2 / 81.4	-1.4 / -1.3
replace sep convolution with normal convolution	72.9 / 82.0	- 0.7 / -0.7
+ data augmentation $\times 2$ (1:1:0)	74.5 / 83.2	+0.9 / +0.5
+ data augmentation $\times 3$ (1:1:1)	74.8 / 83.4	+1.2 / +0.7
+ data augmentation $\times 3$ (1:2:1)	74.3 / 83.1	+0.7 / +0.4
+ data augmentation $\times 3$ (2:2:1)	74.9 / 83.6	+1.3 / +0.9
+ data augmentation $\times 3$ (2:1:1)	75.0 / 83.6	+1.4 / +0.9
+ data augmentation $\times 3$ (3:1:1)	**75.1 / 83.8**	**+1.5 / +1.1**
+ data augmentation $\times 3$ (4:1:1)	75.0 / 83.6	+1.4 / +0.9
+ data augmentation $\times 3$ (5:1:1)	74.9 / 83.5	+1.3 / +0.8

Extractive models

Google QANet

	Published[12]	LeaderBoard[13]
Single Model	EM / F1	EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.0	62.5 / 71.0
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016)	64.7 / 73.7	64.7 / 73.7
Multi-Perspective Matching (Wang et al., 2016)	65.5 / 75.1	70.4 / 78.8
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5	68.5 / 78.0
RaSoR (Lee et al., 2016)	70.8 / 78.7	69.6 / 77.7
FastQAExt (Weissenborn et al., 2017)	70.8 / 78.9	70.8 / 78.9
ReasoNet (Shen et al., 2017b)	69.1 / 78.9	70.6 / 79.4
Document Reader (Chen et al., 2017)	70.0 / 79.0	70.7 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	70.6 / 79.5	70.6 / 79.5
jNet (Zhang et al., 2017)	70.6 / 79.8	70.6 / 79.8
Conductor-net	N/A	72.6 / 81.4
Interactive AoA Reader (Cui et al., 2017)	N/A	73.6 / 81.9
Reg-RaSoR	N/A	75.8 / 83.3
DCN+	N/A	74.9 / 82.8
AIR-FusionNet	N/A	76.0 / 83.9
R-Net (Wang et al., 2017)	72.3 / 80.7	76.5 /84.3
BiDAF + Self Attention + ELMo	N/A	**77.9/ 85.3**
Reinforced Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8	73.2 / 81.8
Dev set: QANet	**73.6 / 82.7**	N/A
Dev set: QANet + data augmentation $\times 2$	**74.5 / 83.2**	N/A
Dev set: QANet + data augmentation $\times 3$	**75.1 / 83.8**	N/A
Test set: QANet + data augmentation $\times 3$	**76.2 / 84.6**	76.2 / 84.6

Table 2: The performances of different models on SQuAD dataset.

Extractive models

Error analysis

Error type	Ratio (%)	Example
Imprecise answer boundaries	50	**Context**: "The Free Movement of Workers Regulation articles 1 to 7 set out the main provisions on equal treatment of workers." **Question**: "Which articles of the Free Movement of Workers Regulation set out the primary provisions on equal treatment of workers?" **Prediction**: "1 to 7", **Answer**: "articles 1 to 7"
Syntactic complications and ambiguities	28	**Context**: "A piece of paper was later found on which Luther had written his last statement. " **Question**: "What was later discovered written by Luther?" **Prediction**: "A piece of paper", **Answer**: "his last statement"
Paraphrase problems	14	**Context**: "Generally, education in Australia follows the three-tier model which includes primary education (primary schools), followed by secondary education (secondary schools/high schools) and tertiary education (universities and/or TAFE colleges)." **Question**: "What is the first model of education, in the Australian system?" **Prediction**: "three-tier", **Answer**: "primary education"
External knowledge	4	**Context**: "On June 4, 2014, the NFL announced that the practice of branding Super Bowl games with Roman numerals, a practice established at Super Bowl V, would be temporarily suspended, and that the game would be named using Arabic numerals as Super Bowl 50 as opposed to Super Bowl L." **Question**: "If Roman numerals were used in the naming of the 50th Super Bowl, which one would have been used?' **Prediction**: "Super Bowl 50", **Answer**: "L"
Multi-sentence	2	**Context**: "Over the next several years in addition to host to host interactive connections the network was enhanced to support terminal to host connections, host to host batch connections (remote job submission, remote printing, batch file transfer), interactive file transfer, gateways to the Tymnet and Telenet public data networks, X.25 host attachments, gateways to X.25 data networks, Ethernet attached hosts, and eventually TCP/IP and additional public universities in Michigan join the network. All of this set the stage for Merit's role in the NSFNET project starting in the mid-1980s." **Question**: "What set the stage for Merits role in NSFNET" **Prediction**: "All of this set the stage for Merit 's role in the NSFNET project starting in the mid-1980s", **Answer**: "Ethernet attached hosts, and eventually TCP/IP and additional public universities in Michigan join the network"
Incorrect preprocessing	2	**Context**: "English chemist John Mayow (1641-1679) refined this work by showing that fire requires only a part of air that he called spiritus nitroaereus or just nitroaereus." **Question**: "John Mayow died in what year?" **Prediction**: "1641-1679", **Answer**: "1679"

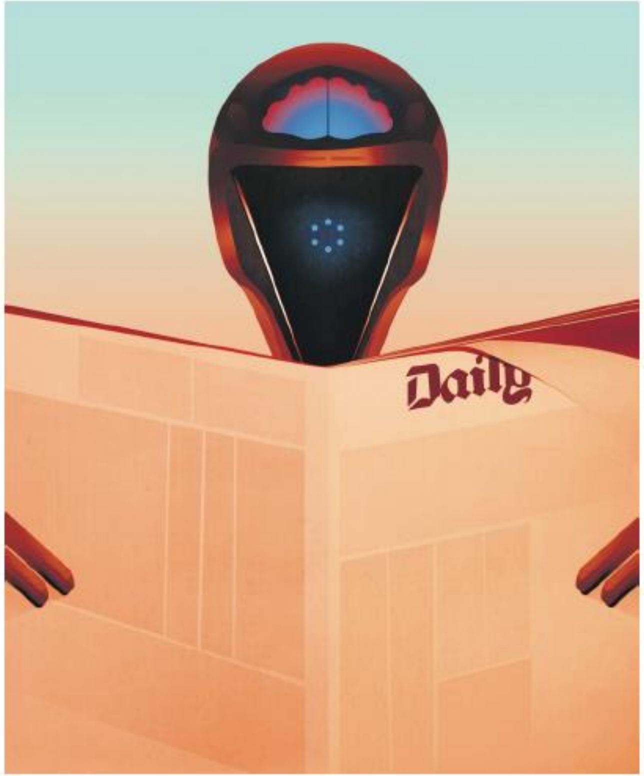
Content

1. Machine reading tasks

2. Models of reading
   1. Building blocks
   2. Retrieval models
   3. Reasoning models

3. Applications

4. Open Questions



Courtesy of Phil Blunsom

Reasoning models

Competent statistical NLP

Featured Logistic Regression

- Whether *e* is in the passage
- Whether *e* is in the question
- Frequency of *e* in passage
- First position of *e* in passage
- n-gram exact match
- Syntactic dependency around *e*

System	CNN Dev	CNN Test	Daily Mail Dev	Daily Mail Test
Frame-semantic model	36.3	40.2	35.5	35.5
Impatient Reader	61.8	63.8	69.0	68.0
Competent statistical NLP	**67.1**	**67.9**	**69.1**	**68.3**
MemNN window + self sup	63.4	66.8		
MemNN win, ss, ens, no-c	66.2	**69.4**		

- *The required reasoning and inference level is **can be limited***
- *There isn't much room left for improvement*
- *However, the scale and ease of data production is appealing*

[23] **Texts as Knowledge Bases**, Manning et al, 2016

Machine reading

Reasoning over knowledge extraction

- Textual data can specify reasoning capabilities

- **Goal**: build machines that can "understand " textual information, *i.e.* converting it into interpretable structured knowledge to be leveraged by humans and other machines alike.

- Optimized with categorical cross-entropy loss

$$CCE = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{J} y_j \cdot log(\hat{y}_j) + (1 - y_j) \cdot log(1 - \hat{y}_j)$$

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

Task 7: Counting
Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets
Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation
Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge
John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

[24] **Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks,** Weston and al

Reasoning models

Memory networks

- Class of models that combine large memory with learning component that can read and write to it.
- Most ML has limited memory which is more-or-less all that's needed for "low level" tasks e.g. object detection.
- Incorporates **reasoning** with **attention** over **memory**.

Reasoning models

End-to-end memory networks

Model

$$\boldsymbol{m}_i = \boldsymbol{A}\Phi(x_i) \qquad \boldsymbol{u} = \boldsymbol{B}\Phi(q)$$

$$\boldsymbol{c}_i = \boldsymbol{C}\Phi(x_i)$$

$$p_i = \mathrm{softmax}(\boldsymbol{u}^\top \boldsymbol{m}_i)$$

$$\boldsymbol{o} = \sum_i p_i \boldsymbol{c}_i$$

$$\boldsymbol{u}^{k+1} = \boldsymbol{o}^k + \boldsymbol{u}^k$$

$$\hat{\boldsymbol{a}} = \mathrm{softmax}(\boldsymbol{u}^\top \boldsymbol{W}' \Phi(\boldsymbol{y}_1), \ldots, \boldsymbol{u}^\top \boldsymbol{W}' \Phi(\boldsymbol{y}_{|C|}))$$

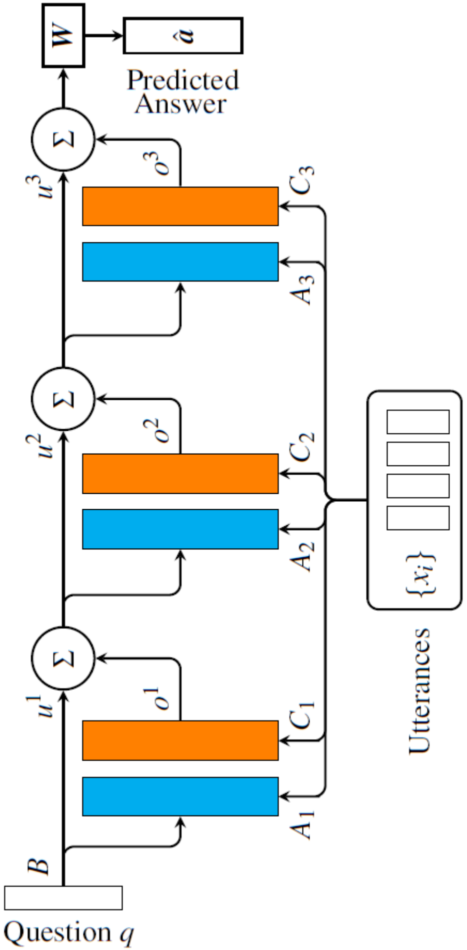
Optimization task

- Categorical cross-entropy
- Stochastic Gradient Descent with clipping
- Grid-searched Hyper Parameters

Joe went to the kitchen.
Fred went to the kitchen.
Joe picked up the milk.
Joe travelled to his office.
Joe left the milk.
Joe went to the bathroom.

Where is the milk now?

Office



Reasoning models

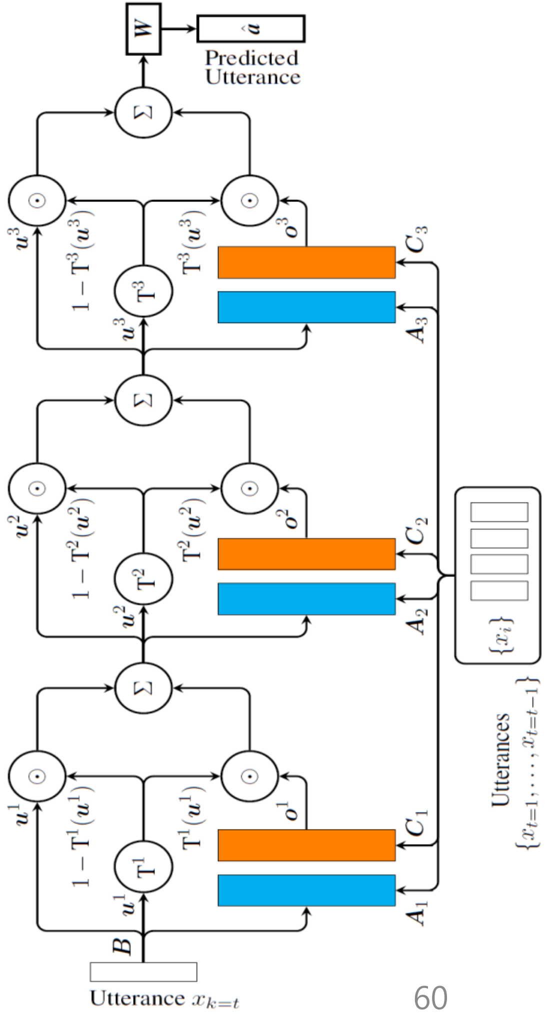
Gated End-to-end memory networks

$$m_i = A\Phi(x_i) \qquad u = B\Phi(q)$$
$$c_i = C\Phi(x_i)$$
$$p_i = \text{softmax}(u^\top m_i)$$
$$o = \sum_i p_i c_i$$

gated controller update

$$\mathrm{T}^k(u^k) = \sigma(W_T^k u^k + b_T^k)$$
$$u^{k+1} = o^k \odot \mathrm{T}^k(u^k) + u^k \odot (1 - \mathrm{T}^k(u^k))$$

$$\hat{a} = \text{softmax}(u^\top W' \Phi(y_1), \ldots, u^\top W' \Phi(y_{|C|}))$$



Properties

- End-to-End memory access regulation
- Close to Highway Network and Residual Network

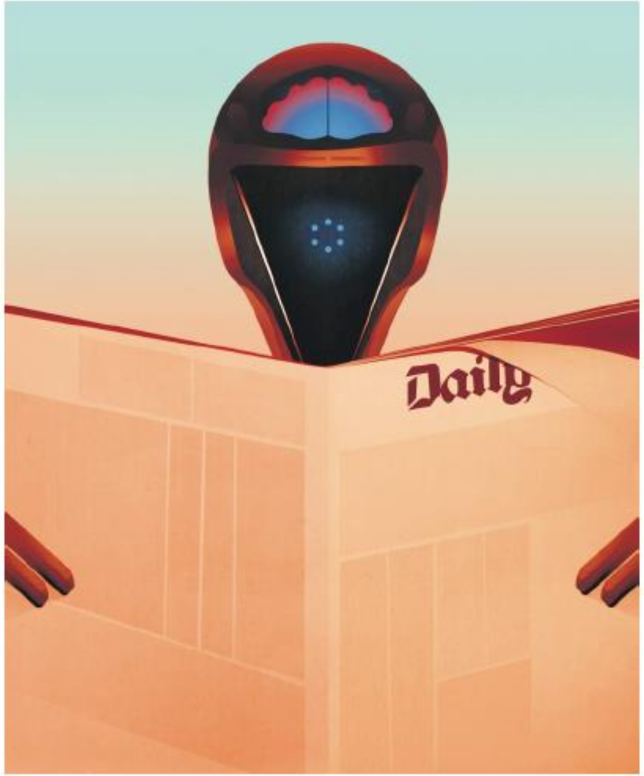
[25] **Gated End-to-End Memory Network**, Liu and Perez, 2017

20 bAbi tasks: Benchmark results

	Baseline			MemN2N									
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	PE LS	PE LS RN	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint	PE LS RN joint	PE LS LW joint
1: 1 supporting fact	0.0	50.0	0.1	0.6	0.1	0.2	0.0	0.8	0.0	0.1	0.0	0.1
2: 2 supporting facts	0.0	80.0	42.8	17.6	21.6	12.8	8.3	62.0	15.6	14.0	11.4	18.8
3: 3 supporting facts	0.0	80.0	76.4	71.0	64.2	58.8	40.3	76.9	31.6	33.1	21.9	31.7
4: 2 argument relations	0.0	39.0	40.3	32.0	3.8	11.6	2.8	22.8	2.2	5.7	13.4	17.5
5: 3 argument relations	2.0	30.0	16.3	18.3	14.1	15.7	13.1	11.0	13.4	14.8	14.4	12.9
6: yes/no questions	0.0	52.0	51.0	8.7	7.9	8.7	7.6	7.2	2.3	3.3	2.8	2.0
7: counting	15.0	51.0	36.1	23.5	21.6	20.3	17.3	15.9	25.4	17.9	18.3	10.1
8: lists/sets	9.0	55.0	37.8	11.4	12.6	12.7	10.0	13.2	11.7	10.1	9.3	6.1
9: simple negation	0.0	36.0	35.9	21.1	23.3	17.0	13.2	5.1	2.0	3.1	1.9	1.5
10: indefinite knowledge	2.0	56.0	68.7	22.8	17.4	18.6	15.1	10.6	5.0	6.6	6.5	2.6
11: basic coreference	0.0	38.0	30.0	4.1	4.3	0.0	0.9	8.4	1.2	0.9	0.3	3.3
12: conjunction	0.0	26.0	10.1	0.3	0.3	0.1	0.2	0.4	0.0	0.3	0.1	0.0
13: compound coreference	0.0	6.0	19.7	10.5	9.9	0.3	0.4	6.3	0.2	1.4	0.2	0.5
14: time reasoning	1.0	73.0	18.3	1.3	1.8	2.0	1.7	36.9	8.1	8.2	6.9	2.0
15: basic deduction	0.0	79.0	64.8	24.3	0.0	0.0	0.0	46.4	0.5	0.0	0.0	1.8
16: basic induction	0.0	77.0	50.5	52.0	52.1	1.6	1.3	47.4	51.3	3.5	2.7	51.0
17: positional reasoning	35.0	49.0	50.9	45.4	50.1	49.0	51.0	44.4	41.2	44.5	40.4	42.6
18: size reasoning	5.0	48.0	51.3	48.1	13.6	10.1	11.1	9.6	10.3	9.2	9.4	9.2
19: path finding	64.0	92.0	100.0	89.7	87.4	85.6	82.8	90.7	89.9	90.2	88.0	90.6
20: agent's motivation	0.0	9.0	3.6	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2
Mean error (%)	6.7	51.3	40.2	25.1	20.3	16.3	13.9	25.8	15.6	13.3	12.4	15.2
Failed tasks (err. $> 5\%$)	4	20	18	15	13	12	11	17	11	11	11	10
On 10k training data												
Mean error (%)	3.2	36.4	39.2	15.4	9.4	7.2	6.6	24.5	10.9	7.9	7.5	11.0
Failed tasks (err. $> 5\%$)	2	16	17	9	6	4	4	16	7	6	6	6

Table 1: Test error rates (%) on the 20 QA tasks for models using 1k training examples (mean test errors for 10k training examples are shown at the bottom). Key: BoW = bag-of-words representation; PE = position encoding representation; LS = linear start training; RN = random injection of time index noise; LW = RNN-style layer-wise weight tying (if not stated, adjacent weight tying is used); joint = joint training on all tasks (as opposed to per-task training).

Content



Courtesy of Phil Blunsom

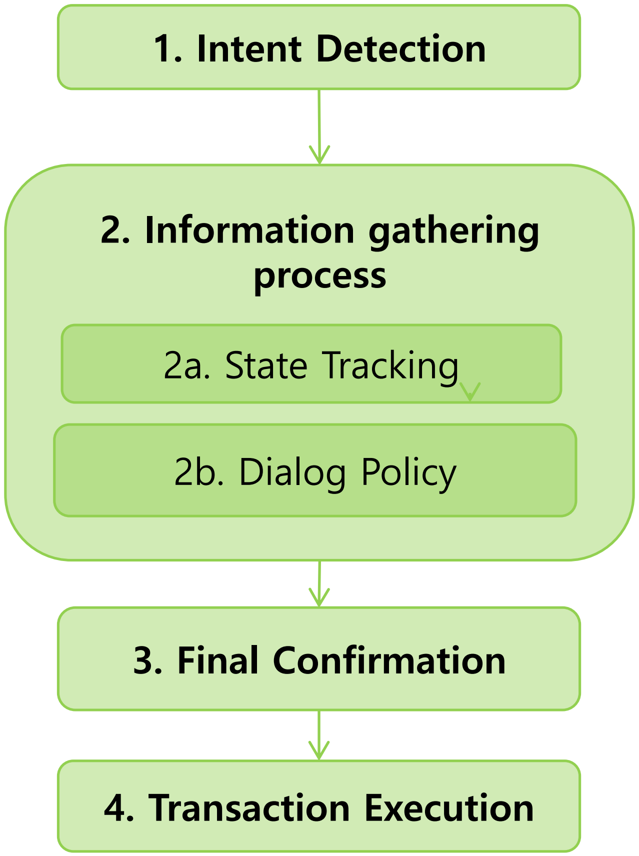
Dialog systems design

Modularity is the current solution

- Divide and Conquer approach
- Annotation processes are required
- Hand-crafted models, case-by-case adaptation

End-to-End opportunities

- Leveraging raw dialogs
- Can be (automatically) enriched with meta-data
- Seamless integration of back-end access



Dialog State tracking

Examples

	Utterance	Food
S	Hello, How may I help you?	
U	I need a Persian restaurant in the south part of town.	0.2 Persian
S	What kind of food would you like?	
U	Persian.	0.8 Persian
S	I'm sorry but there is no restaurant serving persian food	
U	How about Portuguese food?	0.4 Persian 0.6 Portuguese
S	Are you looking for Portuguese food?	
U	Yes.	0.1 Persian 0.9 Portuguese
S	Nandos is a nice place in the south of town serving tasty Portuguese food.	

Slot	User may give as a constraint?
area	Yes, 15 possible values
children allowed	Yes, 2 possible values
food	Yes, 28 possible values
has internet	Yes, 2 possible values
has tv	Yes, 2 possible values
name	Yes, 163 possible values
near	Yes, 52 possible values
pricerange	Yes, 4 possible values
type	Yes, 3 possible values (restaurant, pub, coffeeshop)
addr	No
phone	No
postcode	No
price	No

Informable slots in DSTC3 (Tourist Information Domain)

Slot	User may give as a constraint?
area	Yes, 5 possible values
food	Yes, 91 possible values
name	Yes, 113 possible values
pricerange	Yes, 3 possible values
addr	No
phone	No
postcode	No
signature	No

Informable slots in DSTC2 (Restaurant Information Domain)

[26] **The third Dialog State Tracking Challenge**, Henderson et al, 2016

Dialogue State Tracking

State of the art

Generative

- {Factorial} HMM
- Particle Filter

Discriminative

- Rule-based
- CRF/Max Entropy
- Deep Neural Network

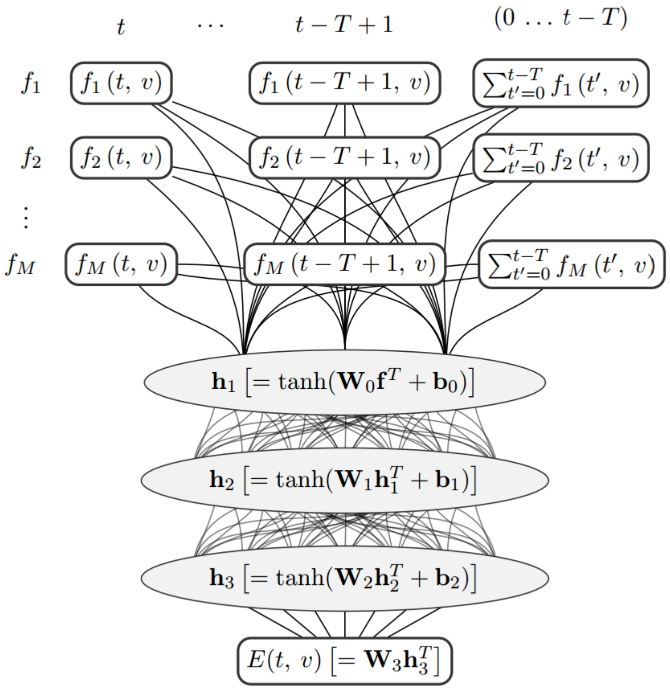


Figure 1: The Neural Network structure for computing $E(t, v) \in \mathbb{R}$ for each possible value v in the set $S_{t,s}$. The vector $\mathbf{f}$ is a concatenation of all the input nodes.

[27] **A generalized rule based tracker for dialogue state tracking**, Yu et al, 2014

[28] **Deep Neural Network Approach for the Dialog State Tracking Challenge**, Henderson et al, 2014

Dialog State Tracking

Open Challenges

1. Longer context
2. Looser supervision schema
3. Reasoning capability
4. Minimize intermediary reps
   - Fixed Ontology
   - Fixed KB

Good Morning, how can I help you

I need a car for March 10th to go to Paris

Ok, I'm checking this

and find me a cheap hotel for *the day after*

(-_-) "

Dialog State Tracking

Machine reading approach

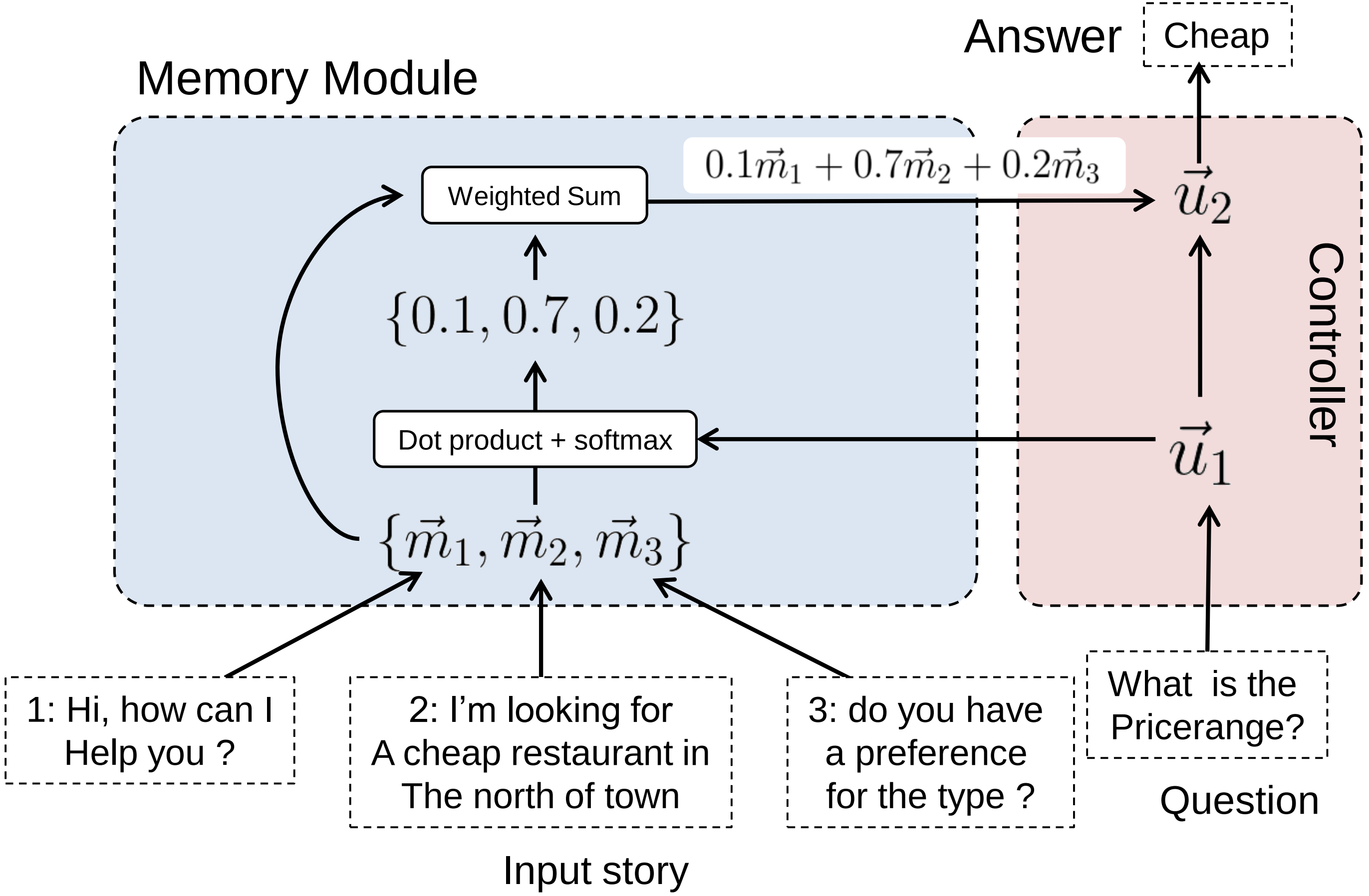
Index	Actor	Utterance
1	Cust	Im looking for a cheap restaurant in the west or east part of town.
2	Agent	Thanh Binh is a nice restaurant in the west of town in the cheap price range.
3	Cust	What is the address and post code.
4	Agent	Thanh Binh is on magdalene street city centre.
5	Cust	Thank you goodbye.
6	**Factoid Question** What is the pricerange ? Answer: {Cheap}	
7	**Yes/No Question** Is the Pricerange Expensive ? Answer: {No}	
8	**Indefinite Knowledge** Is the FoodType chinese ? Answer: {Maybe}	
8	**Listing task** What are the areas ? Answer: {West,East}	

Table 1. State tracking as machine reading task

[29] **Dialog State Tracking, a machine reading approach using memory networks**, Perez and Liu, EACL 2017

Dialog State tracking

with End-to-End Memory Network



End-to-End Memory Network

Results on DSTC-2 – Goal Tracking and Reasoning

[24] Dialog State Tracking, a machine reading approach using deep memory networks, Perez et Liu, EACL 2017

Variable	d	Yes-No	I.K.	Count.	List.
Food	20	**0.85**	0.79	0.89	0.41
	40	0.83	**0.84**	0.88	**0.42**
	60	0.82	0.82	**0.90**	0.39
Area	20	0.86	0.83	0.94	**0.79**
	40	**0.90**	0.89	**0.96**	0.75
	60	0.88	**0.90**	0.95	0.78
PriceRange	20	**0.93**	**0.86**	**0.93**	**0.83**
	40	0.92	0.85	0.90	0.80
	60	0.91	0.85	0.91	0.81

Model	Area	Food	Price	Joint
RNN - no dict.	0.92	0.86	0.86	0.69
RNN + sem. dict.	0.91	0.86	0.93	0.73
NBT-DNN	0.90	0.84	0.94	0.72
NBT-CNN	0.90	0.83	0.93	0.72
MemN2N($d = 40$)	**0.89**	**0.88**	**0.95**	**0.74**

Dialog state tracking

Machine reading approach

On "one supporting fact" task (DSTC-2 dataset): 83% acc vs 79% for the sota.

Table 11: Attention shifting example for the *PriceRange* slot from *DSTC2* dataset

Actor	Utterance	Hop 1	Hop 2	Hop 3	Hop 4	Hop 5
Cust	Im looking for a cheap restaurant that serves chinese food	0.00	0.14	0.01	0.00	0.00
Agent	What part of town do you have in mind	0.02	0.17	0.05	0.00	0.00
Cust	I dont care	0.00	0.00	0.14	0.00	0.00
Agent	Rice house serves chinese food in the cheap price range	0.00	0.02	0.03	0.98	1.00
Cust	What is the address and telephone number	0.57	0.07	0.15	0.00	0.00
Agent	Sure rice house is on mill road city centre	0.03	0.01	0.13	0.02	0.00
Cust	Phone number	0.00	0.01	0.03	0.00	0.00
Agent	The phone number of rice house is 765-239-09	0.37	0.58	0.45	0.00	0.00
Cust	Thank you good bye	0.00	0.00	0.00	0.00	0.00
What is the pricerange ? Answer: cheap						

[29] **Dialog State Tracking, a machine reading approach using memory networks**, Perez and Liu, 2017

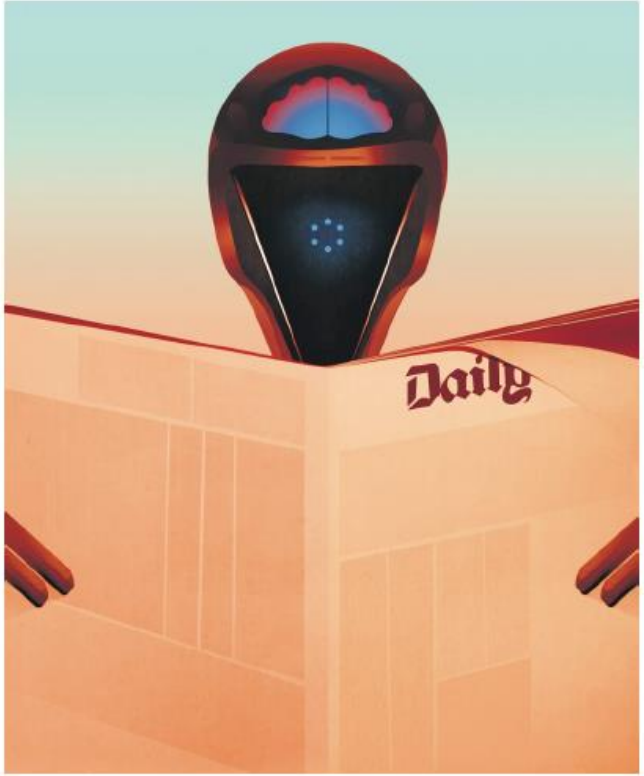
Content

1. Machine reading tasks

2. Models of reading

3. Applications
    1. Dialog State Tracking
    2. Dialog management
    3. User review understanding
    4. Fact checking

4. Open Questions



Courtesy of Phil Blunsom

End-to-End Dialog learning

Goal-oriented dialog tasks

Learning dialog from dialogs

- Simulated dialogs
- Emphasise each step of transaction
- Include some common learning challenges

Goal oriented dialog

- Backed with a Knowledge Base
- KB interactions are included in the decision set

A testbed for deep learning

- End-to-End learnable and flexible
- Attention with Non-parametric memory
- KB-fact and utterance support of the decision
- Dialog learning as Machine Reading

Time	Locutor	Dialog history	Hop #1	Hop #2	Hop #3
3	User	may i have a table in paris	.061	.072	.040
4	Bot	i'm on it	.026	.012	.001
5	User	<silence>	.087	.042	.012
6	Bot	any preference on a type of cuisine	.026	.023	.001
7	User	i love indian food	.081	.070	.055
8	Bot	how many people would be in your party	.025	.006	.001
9	User	we will be six	.059	.051	.018
10	Bot	which price range are looking for	.038	.043	.004
11	User	in a moderate price range please	.080	.095	.096
12	Bot	ok let me look into some options for you	.025	.042	.003
13	User	<silence>	.127	.069	.032
14	Bot	api_call indian paris six moderate	.062	.113	.043
15	User	instead could it be with french food	.188	.311	.683
16	Bot	sure is there anything else to update	.016	.007	.001
17	User	no	.028	.013	.007
18	Bot	ok let me look into some options for you	.011	.006	.000
User input		<silence>			
Correct answer		api_call french paris six moderate			
Predicted answer		api_call french paris six moderate **[Correct]**			

Time	Locutor	Dialog history	Hop #1	Hop #2	Hop #3
14	Bot	api_call indian paris six moderate	.012	.000	.000
15	User	instead could it be with french food	.067	.103	.147
20	Bot	api_call french paris six moderate	.012	.000	.000
21	User	resto_1 r_phone rest_1_phone	.018	.004	.000
23	User	resto_1 r_cuisine french	.029	.005	.000
24	User	resto_1 r_location paris	.060	.292	.094
25	User	resto_1 r_number six	.050	.298	.745
26	User	resto_1 r_price moderate	.060	.090	.002
27	User	resto_1 r_rating 6	.016	.002	.000
30	User	resto_2 r_cuisine french	.031	.007	.000
31	User	resto_2 r_location paris	.040	.081	.004
32	User	resto_2 r_number six	.020	.012	.000
33	User	resto_2 r_price moderate	.029	.009	.000
37	User	resto_3 r_cuisine french	.014	.001	.000
38	User	resto_3 r_location paris	.028	.016	.001
39	User	resto_3 r_number six	.024	.022	.004
40	User	resto_3 r_price moderate	.039	.015	.001
User input		<silence>			
Correct answer		what do you think of this option: resto_1			
Predicted answer		what do you think of this option: resto_1 **[Correct]**			

[30] **Dialog System and Technology Challenge 6th - Task 1 - End-to-End Goal-Oriented Dialog,** Perez, Bourreau and Bordes, 2017

End-to-End Dialog learning

Dialog System and Technology Challenge 6th - Task 1

Organization

- Task 1: Issuing API calls.
- Task 2: Updating API calls.
- Task 3: Displaying options.
- Task 4: Providing extra information.
- Task 5: Conducting full dialogs.

Corpora

- 2 corpus with/without OOV
- 2 corpus with a new slot
- 2 Knowledge Bases

Objectives

- Emphasise challenges of real world transactional dialog
- Compare the models and learning algorithms

Time	Locutor	Dialog history	Hop #1	Hop #2	Hop #3
3	User	may i have a table in paris	.061	.072	.040
4	Bot	i'm on it	.026	.012	.001
5	User	<silence>	.087	.042	.012
6	Bot	any preference on a type of cuisine	.026	.023	.001
7	User	i love indian food	.081	.070	.055
8	Bot	how many people would be in your party	.025	.006	.001
9	User	we will be six	.059	.051	.018
10	Bot	which price range are looking for	.038	.043	.004
11	User	in a moderate price range please	.080	.095	.096
12	Bot	ok let me look into some options for you	.025	.042	.003
13	User	<silence>	.127	.069	.032
14	Bot	api_call indian paris six moderate	.062	.113	.043
15	User	instead could it be with french food	.188	.311	.683
16	Bot	sure is there anything else to update	.016	.007	.001
17	User	no	.028	.013	.007
18	Bot	ok let me look into some options for you	.011	.006	.000
User input		<silence>			
Correct answer		api_call french paris six moderate			
Predicted answer		api_call french paris six moderate	**[Correct]**		

Time	Locutor	Dialog history	Hop #1	Hop #2	Hop #3
14	Bot	api_call indian paris six moderate	.012	.000	.000
15	User	instead could it be with french food	.067	.103	.147
20	Bot	api_call french paris six moderate	.012	.000	.000
21	User	resto_1 r_phone rest_1_phone	.018	.004	.000
23	User	resto_1 r_cuisine french	.029	.005	.000
24	User	resto_1 r_location paris	.060	.292	.094
25	User	resto_1 r_number six	.050	.298	.745
26	User	resto_1 r_price moderate	.060	.090	.002
27	User	resto_1 r_rating 6	.016	.002	.000
30	User	resto_2 r_cuisine french	.031	.007	.000
31	User	resto_2 r_location paris	.040	.081	.004
32	User	resto_2 r_number six	.020	.012	.000
33	User	resto_2 r_price moderate	.029	.009	.000
37	User	resto_3 r_cuisine french	.014	.001	.000
38	User	resto_3 r_location paris	.028	.016	.001
39	User	resto_3 r_number six	.024	.022	.004
40	User	resto_3 r_price moderate	.039	.015	.001
User input		<silence>			
Correct answer		what do you think of this option: resto_1			
Predicted answer		what do you think of this option: resto_1	**[Correct]**		

[30] **Dialog System and Technology Challenge 6th - Task 1 - End-to-End Goal-Oriented Dialog,** Perez, Bourreau and Bordes, 2016

Systems and results

Decision models

- (Dynamic) Memory Networks [1,2]
- LSTMs [3]
- Hybrid Code Networks [4]
- Recurrent Entity Networks [5]
- Quantitazed Language Model

Entity/Slot resolution strategies

- Dictionary and Heuristics
- Dedicated models (CRF, LSTMs)
- Delexicalization

Losses

- Categorical Cross-Entropy
- Ranking loss over similarity measure

Optimizers

- Mometum based SGD
- Gradient clipping
- Early stopping strategy

[31] Long Short Term Memory, Hochreiter and Schmidhuber, 1997
[32] Ask Me Anything: Dynamic Memory Networks for Natural Language Processing, Socher et al, 2015
[33] End-to-end Memory Network, Sukhbaatar et al, 2015
[34] Hybrid Code Networks, Williams et al, 2017
[35] Tracking the World State with Recurrent Entity Networks, Henalf et al, 2017

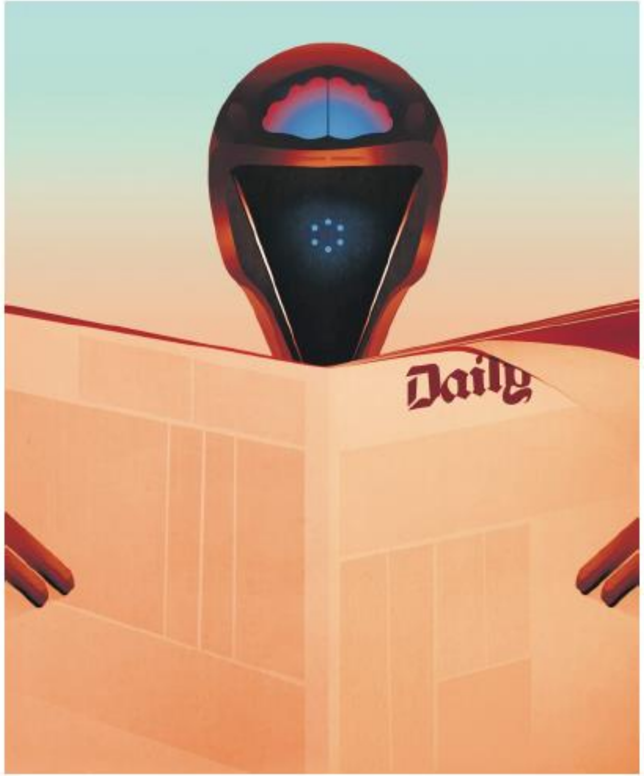
Content

1. Machine reading tasks

2. Models of reading

3. Applications
    1. Dialog State Tracking
    2. Dialog Management
    3. User review understanding
    4. Fact checking

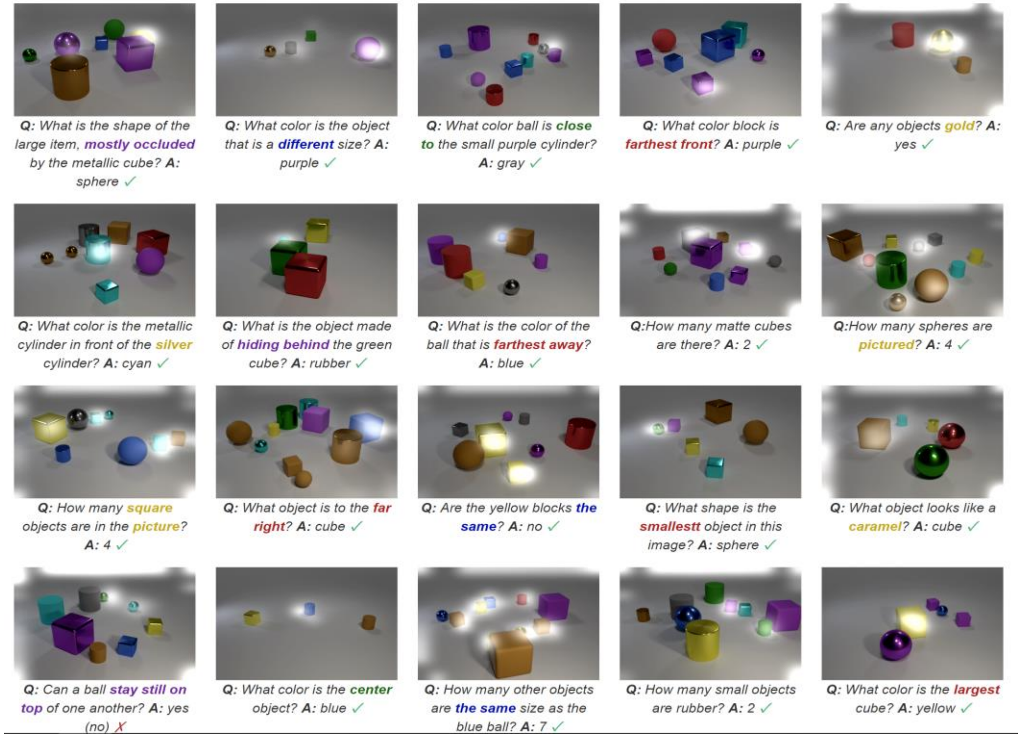
4. Open Questions



Courtesy of Phil Blunsom

Review reading

Inspiration from relational visual question answering [Johnson et al, 2017]



p.s. Here are some more examples of the model's predictions. See how the model correctly handle questions that involve **obstructions**, **object uniqueness**, **relative distances**, **superlatives**, **varied vocabulary**.

Review reading

ReviewQA: a relational aspect-based opinion reading dataset

Hotel: BEST WESTERN Corona
Title: Convenient Location. Helpful Staff.
Overall rating: ★★★★☆

Comment: I just needed a place to sleep and this place was ideally located for my meetings. Plimlico tube is only a few minutes walk. Room was small but clean. Staff very helpful. Breakfast OK.

Ratings

Service	★★★★☆	Location	★★★★★
Rooms	★★★☆☆	Cleanliness	★★★★☆

Task	**Natural Language Questions**	
5	What is the rating of service?	3
3	Is the client satisfied with the location?	Yes
7	Does the customer prefer the service or the room?	Service

	# documents	# queries
Train	90.000	528.665
Test	10.000	58.827
Total	100.000	587.492

Task id	Description/Comment	Example	Expected answer
1	**Detection of an aspect in a review.**	Is sleep quality mentioned in this review?	Yes/No
2	**Prediction of the customer general satisfaction.**	Is the client satisfy by this hotel?	Yes/No
3	**Prediction of the global trend of an aspect in a given review.**	Is the client satisfied with the cleanliness of the hotel?	Yes/No
4	**Prediction of whether the rating of a given aspect is above or under a given value.**	Is the rating of location under 4?	Yes/No
5	**Prediction of the exact rating of an aspect in a review.**	What is the rating of the aspect Value in this review?	A rating between 1 and 5
6	**Prediction of the list of all the positive/negative aspects mentioned in the review.**	Can you give me a list of all the positive aspects in this review?	a list of aspects
7.0	**Comparison between aspects.**	Is the sleep quality better than the service in this hotel?	Yes/No
7.1		Which one of these two aspects, service, location has the best rating?	an aspect
8	**Prediction of the strengths and weaknesses in a review.**	What is the best aspect rated in this comment?	an aspect

[36] **ReviewQA: a relational aspect-based opinion reading dataset**, Grail and Perez, 2018

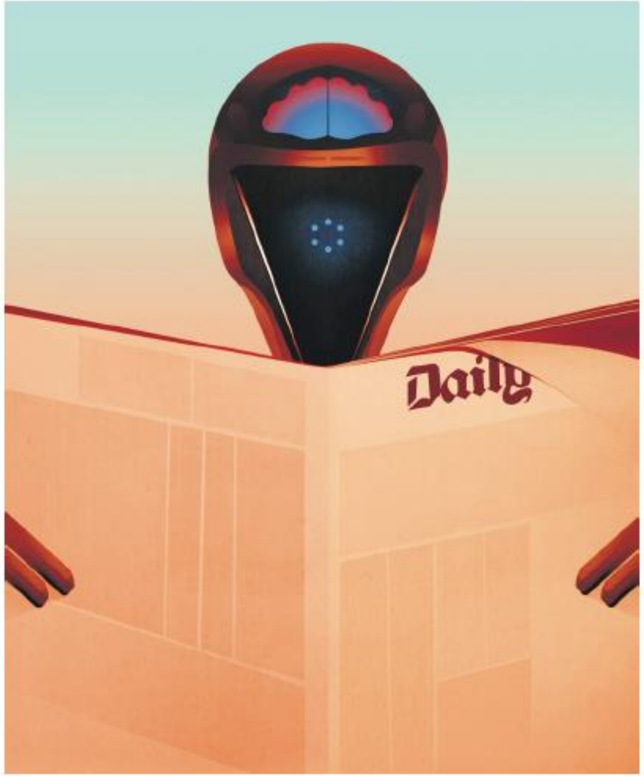
Content

1. Machine reading tasks

2. Models of reading

3. Applications
    1. Dialog State Tracking
    2. Dialog Management
    3. User review understanding
    4. Fact checking

4. Open Questions



Fact checking

- Given a claim, retrieve evidence documents for and against it

- Given evidence documents, find relevant paragraphs and sentences in it

- For claim and each evidence paragraph and sentence: detect stance of paragraph sentence towards a claim/target

Stance detection:
Tweet: Be prepared - if we continue the policies of the liberal left, we will be #Greece
Target: Donald Trump
Label: favor

Fake news detection:
Document: Dino Ferrari hooked the whopper wels catfish, (...), which could be the biggest in the world.
Headline: Fisherman lands 19 STONE catfish which could be the biggest in the world to be hooked
Label: agree

Natural language inference:
Premise: Fun for only children
Hypothesis: Fun for adults and children
Label: contradiction

# Headline-body pairs			49972
# Headlines			1648
# Bodies			1683
# Bodies in test set			169
# Headline-body pairs in test set			5025
Average # tokens of headline			12.6
Average # tokens of body			427.5
Unrelated	*Discuss*	*Agree*	*Disagree*
73.1%	17.8%	7.4%	1.7%

Table 1: Statistics of *FNC1* dataset

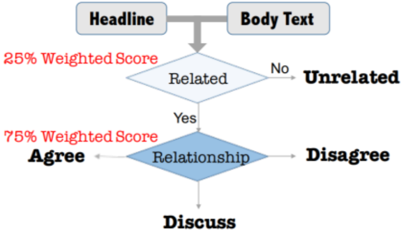


Figure 2: Score Metric for *FNC1*

[37] http://www.fakenewschallenge.org/ , 2017

Fact checking as Stance Detection

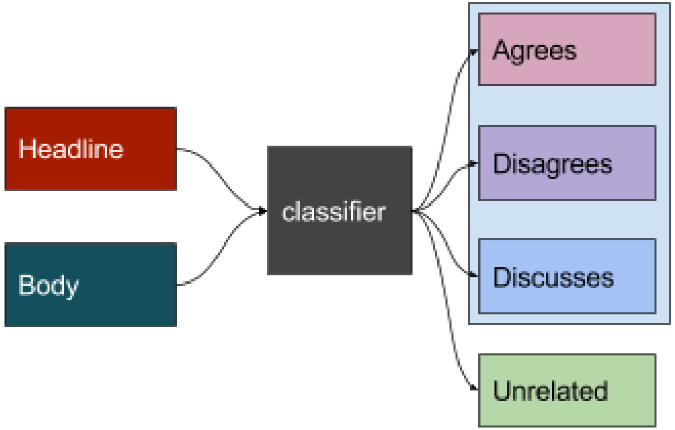
Determine attitude expressed in document and paragraph/sentence towards a topic, statement and target

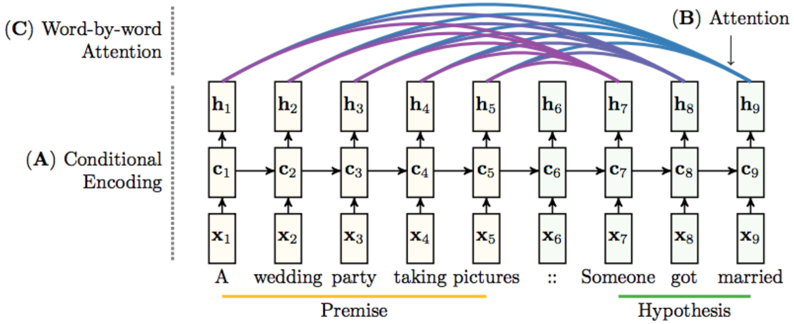
Different classification schemes

- positive, negative, neutral
(SemEval 2016 Task 6, RTE, SNLI)

- support, deny, query, comment
(SemEval 2017 Task 8 RumourEval)

- agree, disagree, discuss, unrelated
(Fake News Challenge)





[37] http://www.fakenewschallenge.org/ , 2017

Fact checking as Stance Detection

Deep LSTM reader

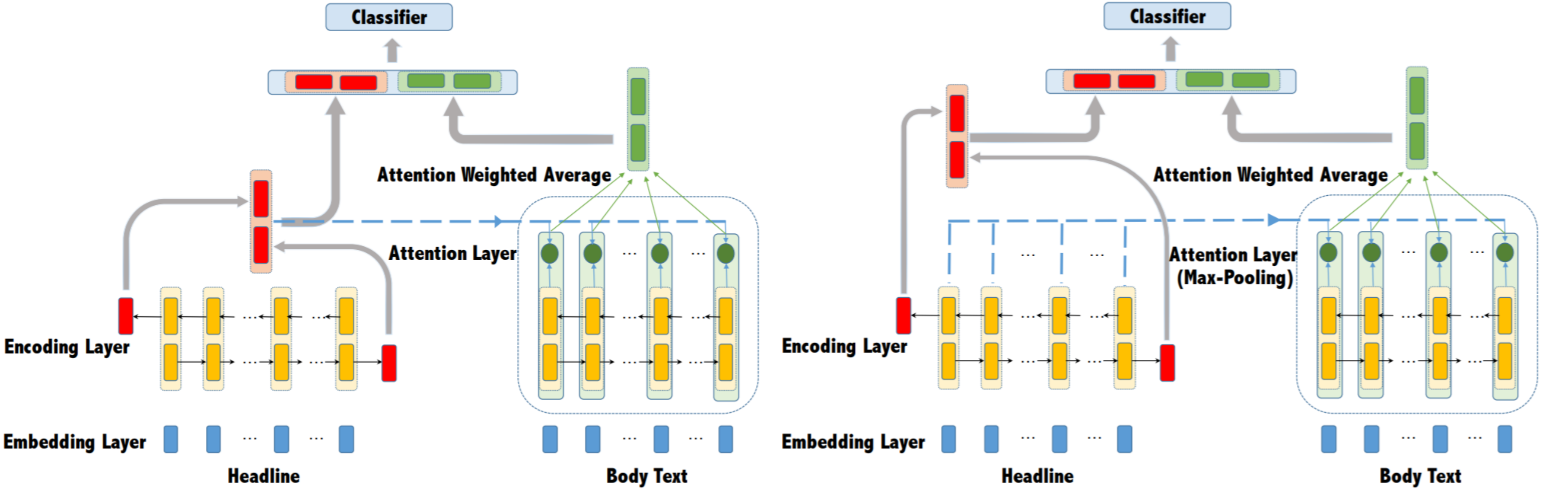


Figure 3: Illustration of Attentive Reader with simple attention (left) and full attention (right)

[38] **Neural Stance Detectors for Fake News Challenge**, Xu et al, 2017
[39] **Stance detection with bidirectional conditional encoding**, Augenstein et al. 2016

Fact checking as Stance Detection

Deep LSTM reader

Models	Ave. Dev. Score	Max Dev. Score	Ave. Test Score	Max Test Score
FNC Baseline	–	–	79.2%	–
Bidirectional Encoder (unconditional)	80.1%	80.5%	79.9%	80.1%
Bidirectional Encoder (conditional)	79.5%	81.2%	80.2%	82.0%
Bidirectional Encoder (concatenated)	82.7%	82.9%	82.0%	83.5%
Attentive Reader (simple attention)	82.4%	83.4%	81.4%	82.6%
Attentive Reader (full attention)	83.7%	85.4%	85.2%	86.5%
Bilateral Multiple Perspective Matching	84.1%	84.8%	84.6%	85.6%

Table 5: Evaluation results on both development set and test set for various models

[38] **Neural Stance Detectors for Fake News Challenge**, Xu et al, 2017

Fact checking as Stance Detection

Deep LSTM reader

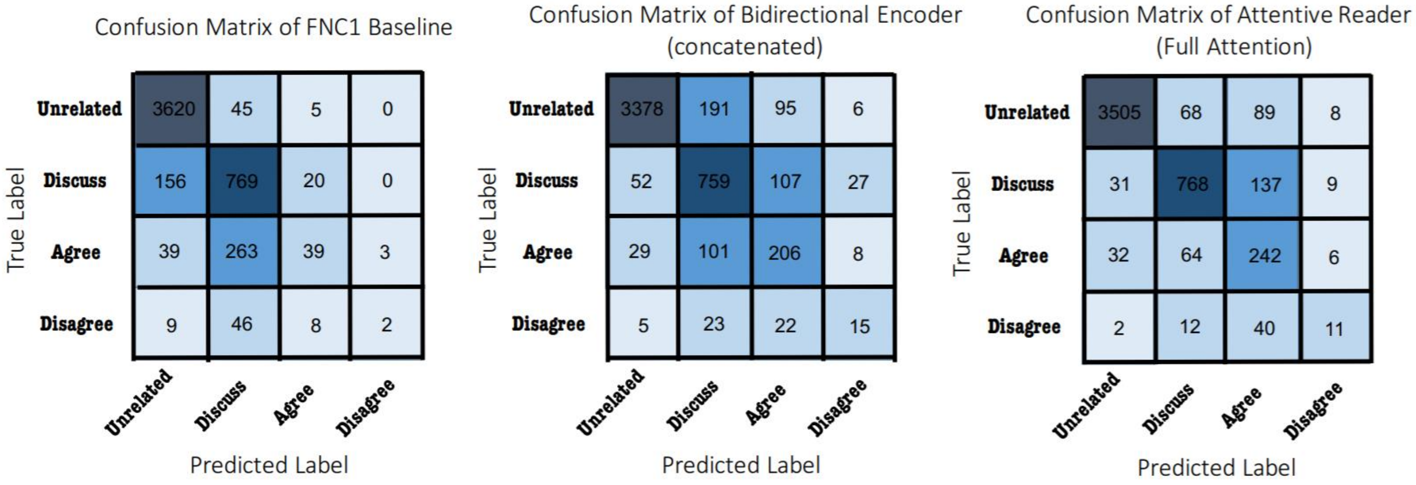


Figure 6: Confusion matrix on test set using *FNC1* Baseline (left), Bidirectional Encode (concatenated) (middle) and Attentive Reader with full attention (right)

[38] **Neural Stance Detectors for Fake News Challenge**, Xu et al, 2017

Fact checking as Stance Detection

“Relationship between sequences can be modelled effectively with deep neural models”

Many challenges

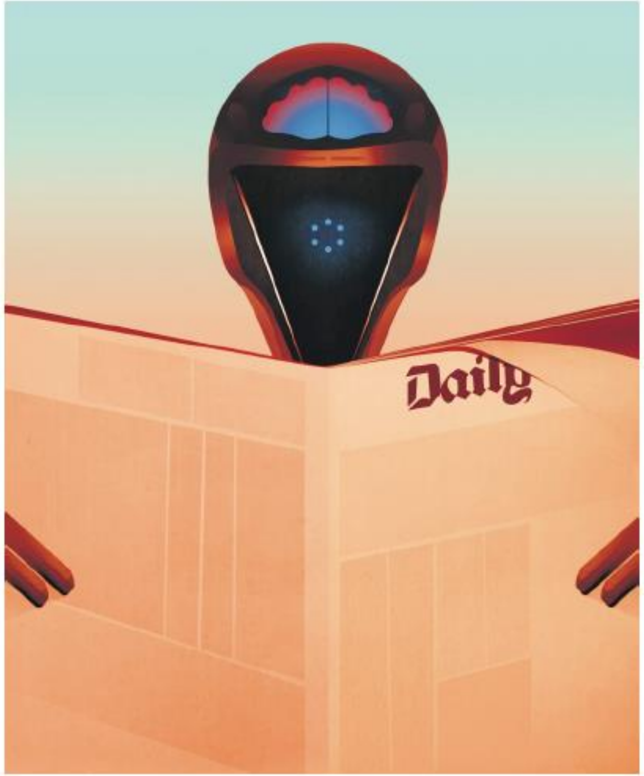
• Hard to collect data, especially with balanced labels (un/semi - supervised ?)

• Little and imbalanced data (multi-task ?)

• Explainable decisions are (often) needed

Content

1. Machine reading tasks
2. Models of reading
3. Applications
4. Open Questions



Courtesy of Phil Blunsom

Open Questions

Multi-document Open-Domain Question answering

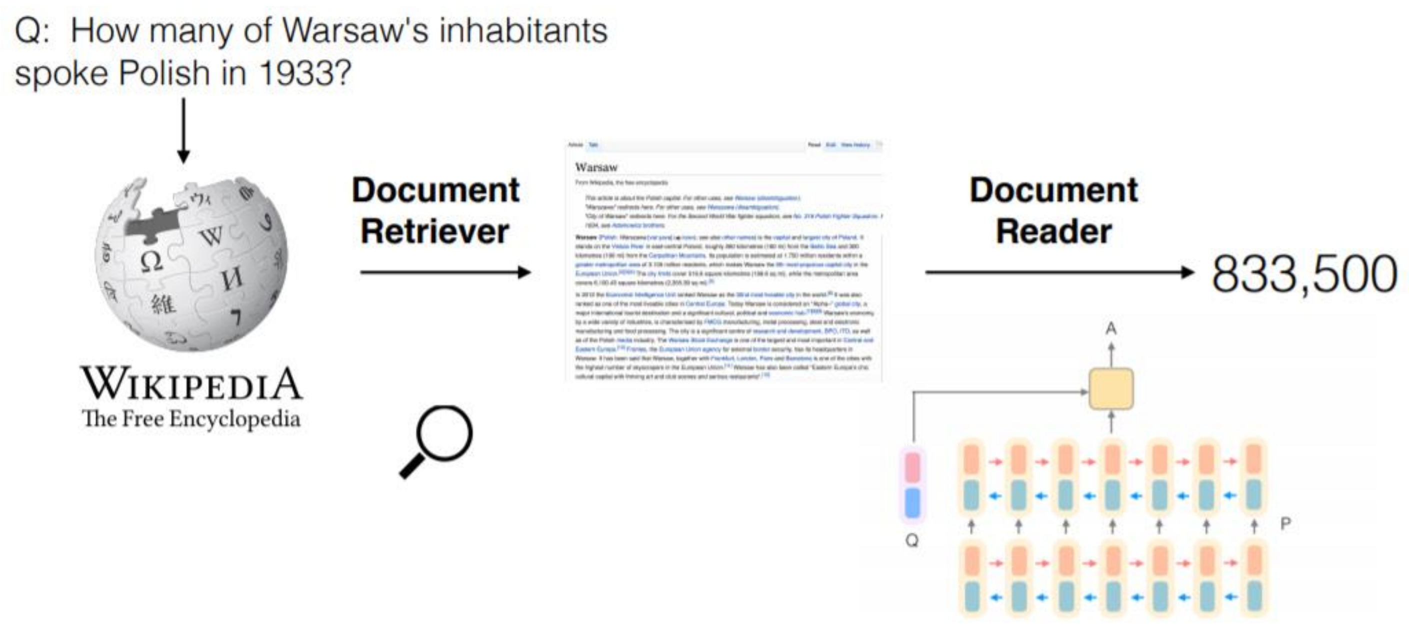


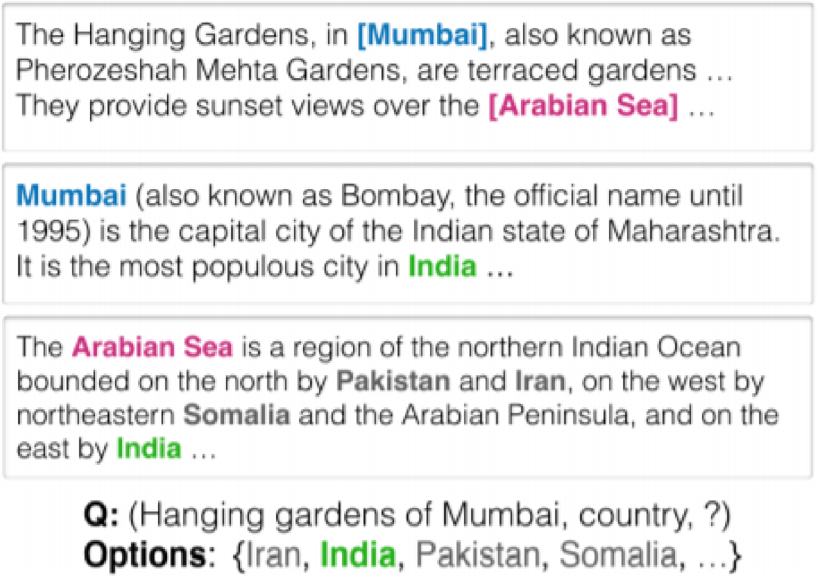
Figure 1: An overview of our question answering system DrQA.

[40] **Reading Wikipedia to Answer Open-Domain Questions**, Chen et al, 2017

Open Questions

Multi document reasoning

- Most Reading Comprehension methods limit themselves to queries which can be answered using a single sentence, paragraph, or document.
- Enabling models to combine disjoint pieces of textual evidence would extend the scope of machine comprehension
- Text understanding across multiple documents and to investigate the limits of existing methods.
- Toward ensemblist operations (union, intersection, selection …)

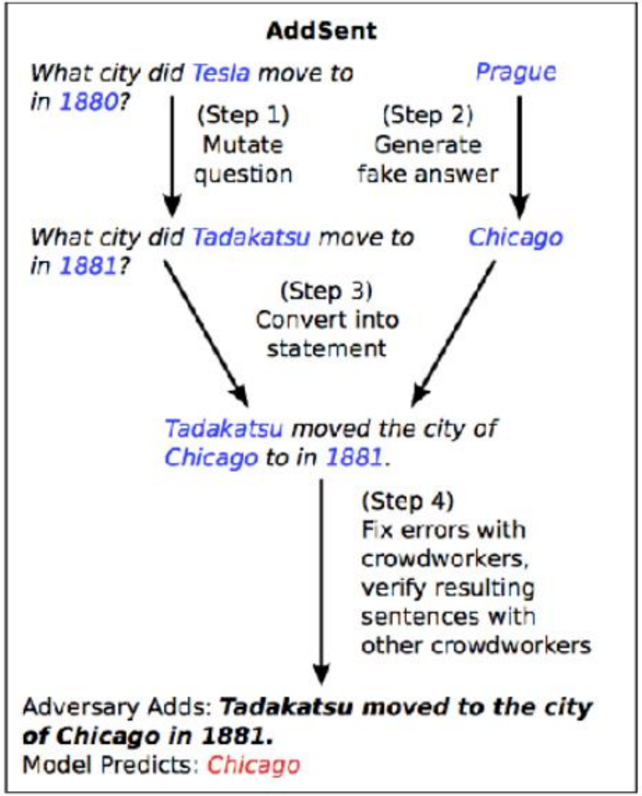


[41] **Constructing Datasets for Multi-hop Reading Comprehension Across Documents**, Riedel et al, 2017

Open Questions

Adversarial Examples

- Add a sentence or word string specifically designed to distract the model
- Drops accuracy of state-of-the-art models from 81% to 46% of Exact Match accuracy
- Current issue of deep models, already observed on image tasks



[42] **Adversarial Examples for Evaluating Reading Comprehension Systems**, Liang et al, 2017

Conclusions

Machine reading paradigm, a next step toward natural language comprehension

Promissing results are already available

Deep learning is (currently) a major enabler of this recent development

Machine reading is a playground for (deep) machine learning research

Very active community (Datasets, papers and codes)

A lot of challenges with numerous possible impacts

Thank you

europe.naverlabs.com