NLP Applications II

Part2: Question Answering

About myself

ander.barrena@ehu.eus
@4nderB

- PhD in Natural Language Processing in 2017 (UPV/EHU) Machine Learning!
  - PGM and Generative Statistical Models
  - Personalized Page Rank
- Enjoyed 2 post-Doctoral grants from UPV/EHU
- Assistant Professor & HAP/LAP & DL4nlp
- Research interest:
  - Multilingual Named Entity Disambiguation
  - Multilingual Word Sense Disambiguation
  - Sequence Labeling
  - Language Modeling
  - Question Answering
  - Semantic Textual Similarity
  - Information Extraction…
  - Prompting
  - Brain Image Classification for Dyslexia

Course Outline

- Question Answering systems (Q.A.) +lab (Q.A. fine-tune BERT model)
- Multilingual and Multimodal Q.A. +lab (Q.A. test BERT models)
- Information Retrieval (I.R.) +lab (I.R. train and test BM25 model)
- Open Domain Q.A. +assingment (Open Domain Q.A)

Course Outline

- **Question Answering systems (Q.A.)** +lab
  - Based on Danqi Chen (Princeton University) & Christopher Manning (Stanford University) & Jon Ander Campos (UPV/EHU) slides.
- Multilingual and Multimodal Q.A. +lab
- Information Retrieval (I.R.) +lab
- Open Domain Q.A. +assignment

Course Outline

- **Question Answering systems (Q.A.)**
  - **Introduction**
  - QA practical applications
  - Reading comprehension
  - SQuAD dataset
  - QA models
  - Is reading comprehension solved?

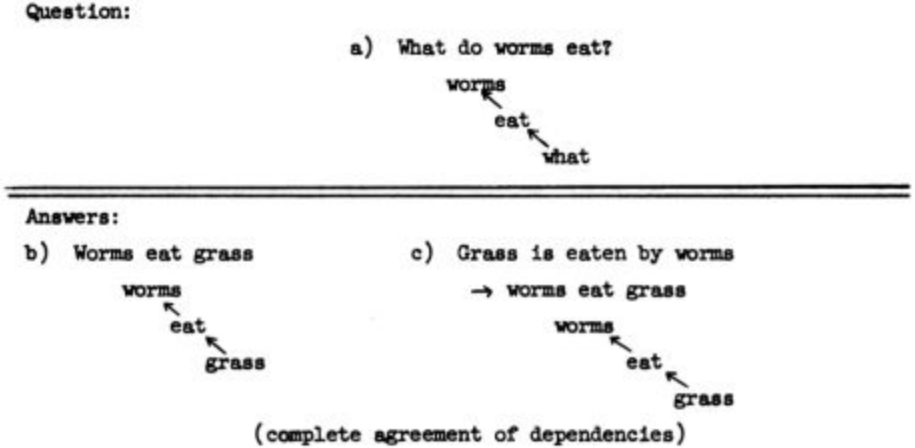
What is question answering?

Question (Q)
Answer (A)

The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language

What is question answering?

Question (Q)
Answer (A)



Question:

a) What do worms eat?

worms ← eat ← what

Answers:

b) Worms eat grass

worms ← eat ← grass

c) Grass is eaten by worms

→ worms eat grass

worms ← eat ← grass

(complete agreement of dependencies)

The earliest QA systems dated back to 1960s!

(Simmons et al., 1964)

Q.A.: a taxonomy

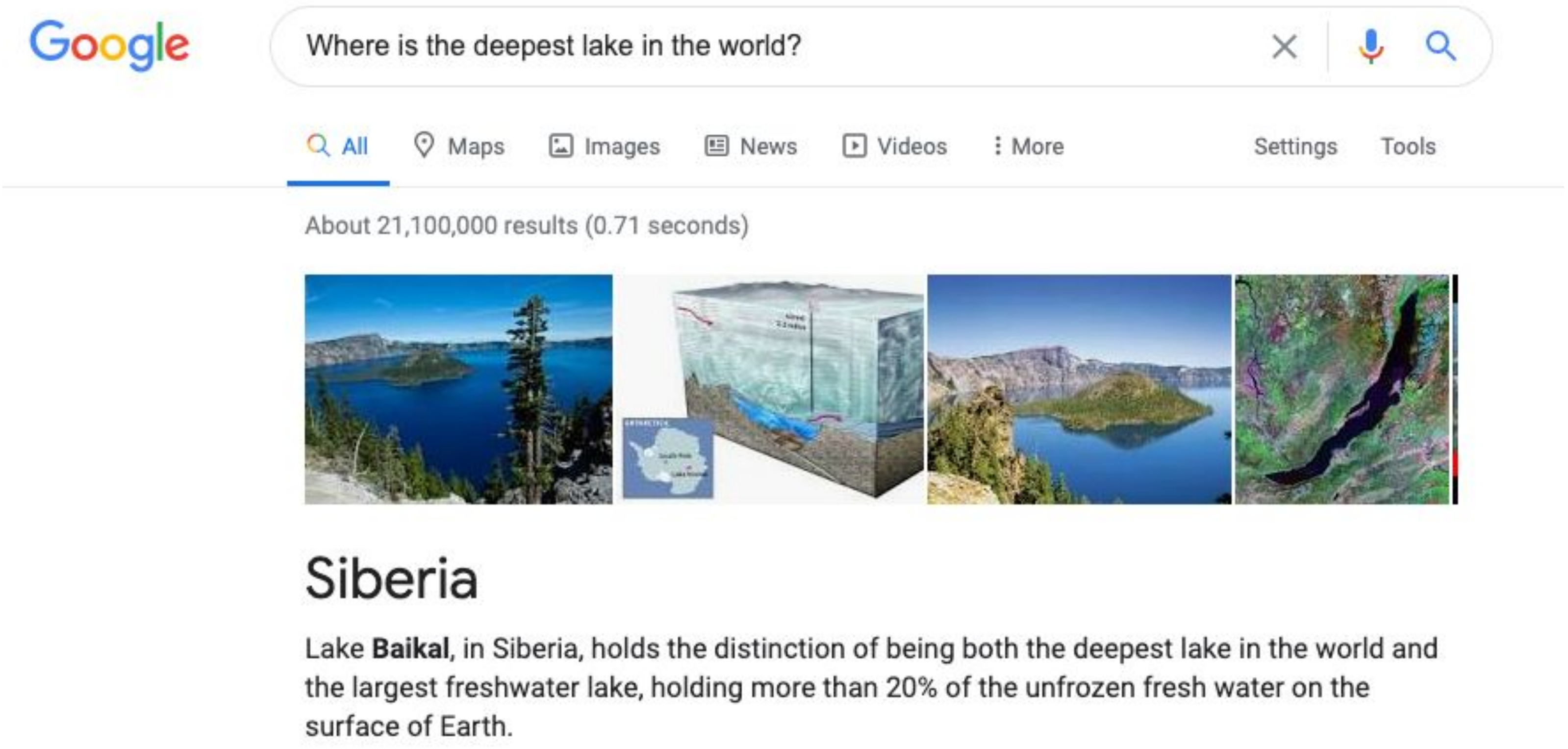
Question (Q)
Answer (A)

- What information source does a system build on?
  - A text passage, all Web documents, knowledge bases, tables, images..
- Question type
  - Factoid vs non-factoid, open-domain vs closed-domain, simple vs compositional, ..
  - “Who discovered electricity?” Vs “How to cook burgers?”
- Answer type
  - A short segment of text, a paragraph, a list, yes/no, ...

Course Outline

- **Question Answering systems (Q.A.)**
  - Introduction
  - **QA practical applications**
  - Reading comprehension
  - SQuAD dataset
  - QA models
  - Is reading comprehension solved?

Q.A.: Lots of practical application



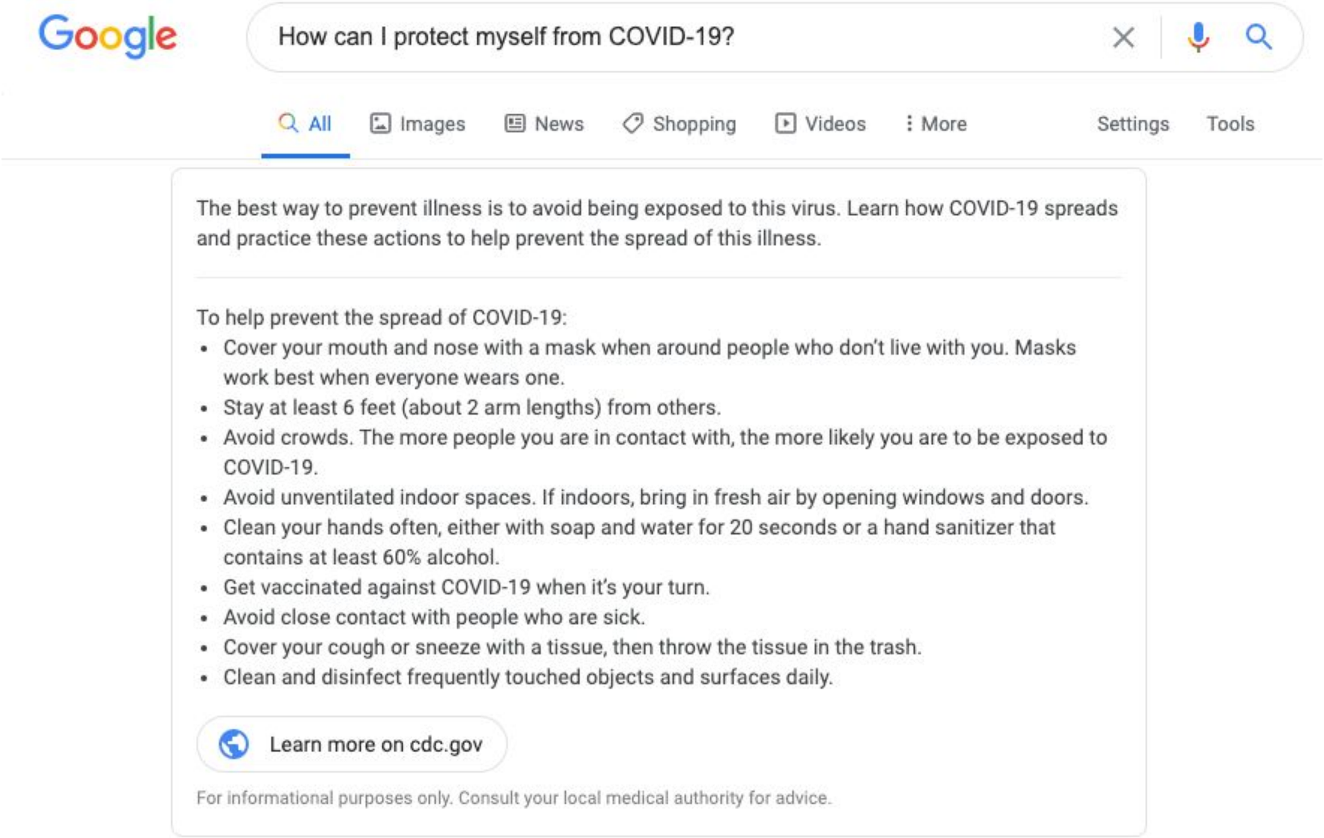
Where is the deepest lake in the world?

About 21,100,000 results (0.71 seconds)

Siberia

Lake **Baikal**, in Siberia, holds the distinction of being both the deepest lake in the world and the largest freshwater lake, holding more than 20% of the unfrozen fresh water on the surface of Earth.

Q.A.: Lots of practical application



The best way to prevent illness is to avoid being exposed to this virus. Learn how COVID-19 spreads and practice these actions to help prevent the spread of this illness.

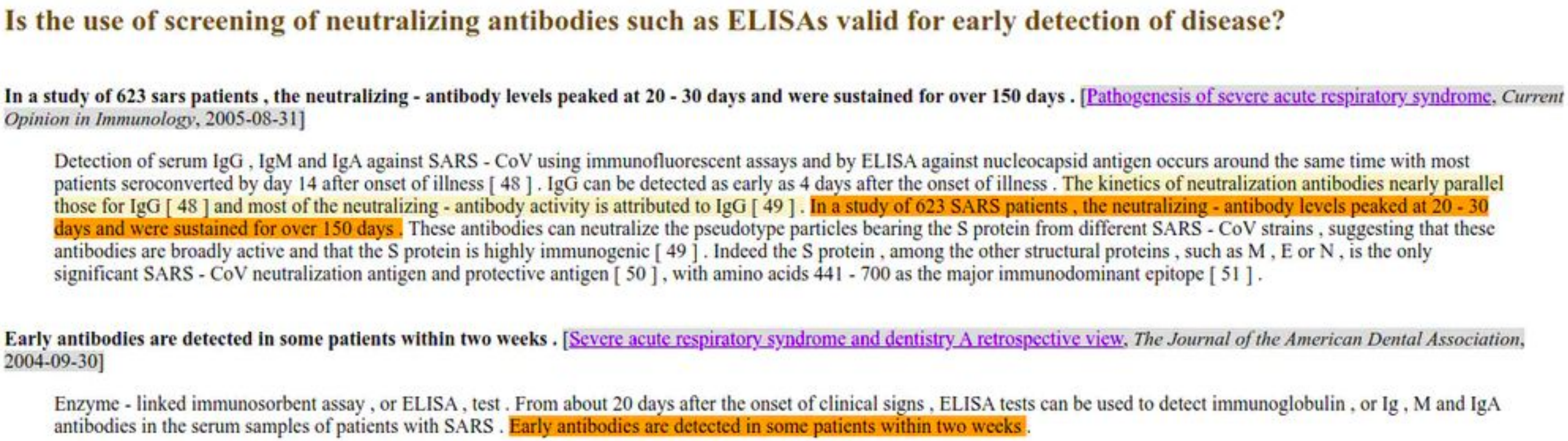
To help prevent the spread of COVID-19:

- Cover your mouth and nose with a mask when around people who don't live with you. Masks work best when everyone wears one.
- Stay at least 6 feet (about 2 arm lengths) from others.
- Avoid crowds. The more people you are in contact with, the more likely you are to be exposed to COVID-19.
- Avoid unventilated indoor spaces. If indoors, bring in fresh air by opening windows and doors.
- Clean your hands often, either with soap and water for 20 seconds or a hand sanitizer that contains at least 60% alcohol.
- Get vaccinated against COVID-19 when it's your turn.
- Avoid close contact with people who are sick.
- Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
- Clean and disinfect frequently touched objects and surfaces daily.

Learn more on cdc.gov

For informational purposes only. Consult your local medical authority for advice.

Q.A.: Lots of practical application



Is the use of screening of neutralizing antibodies such as ELISAs valid for early detection of disease?

In a study of 623 sars patients , the neutralizing - antibody levels peaked at 20 - 30 days and were sustained for over 150 days . [Pathogenesis of severe acute respiratory syndrome, *Current Opinion in Immunology*, 2005-08-31]

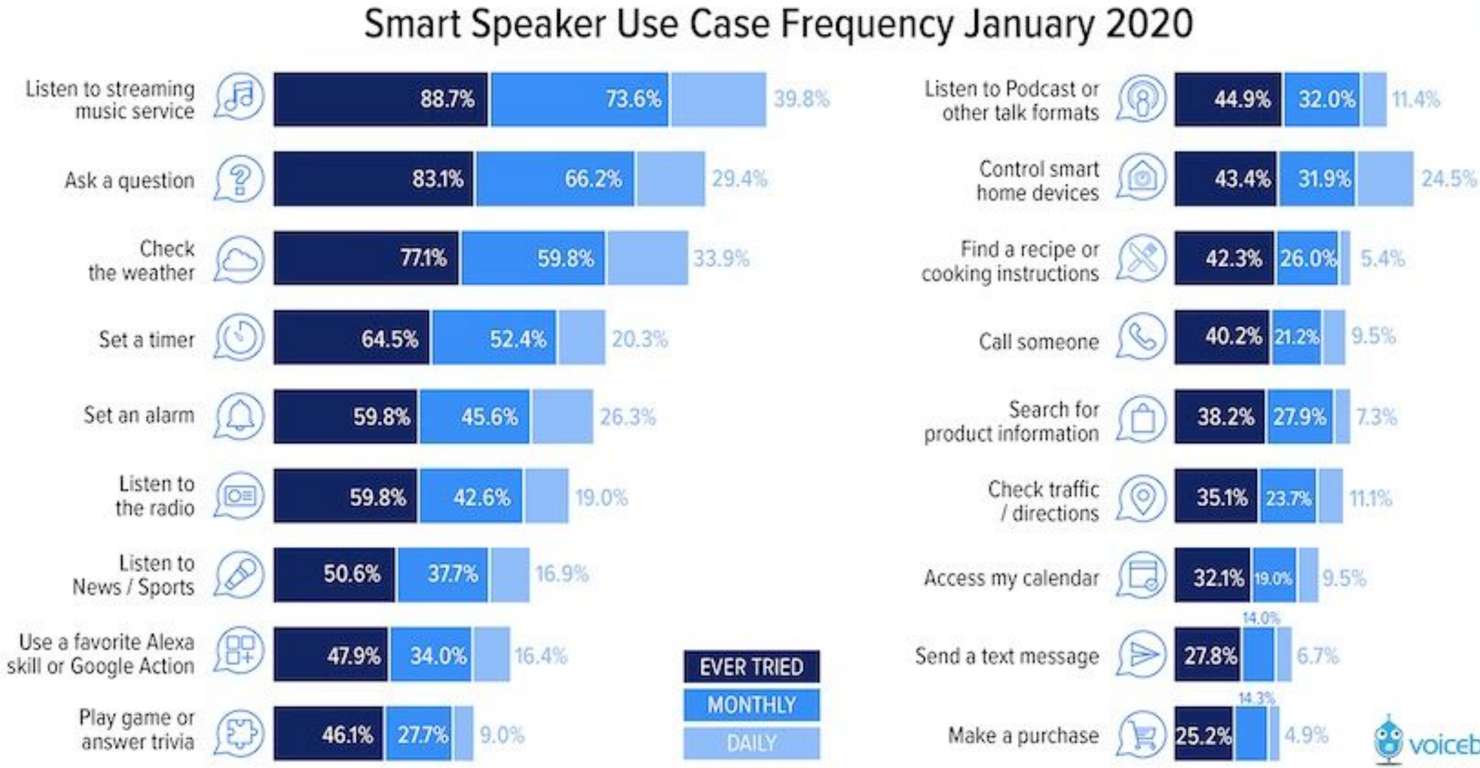
> Detection of serum IgG , IgM and IgA against SARS - CoV using immunofluorescent assays and by ELISA against nucleocapsid antigen occurs around the same time with most patients seroconverted by day 14 after onset of illness [48] . IgG can be detected as early as 4 days after the onset of illness . The kinetics of neutralization antibodies nearly parallel those for IgG [48] and most of the neutralizing - antibody activity is attributed to IgG [49] . In a study of 623 SARS patients , the neutralizing - antibody levels peaked at 20 - 30 days and were sustained for over 150 days . These antibodies can neutralize the pseudotype particles bearing the S protein from different SARS - CoV strains , suggesting that these antibodies are broadly active and that the S protein is highly immunogenic [49] . Indeed the S protein , among the other structural proteins , such as M , E or N , is the only significant SARS - CoV neutralization antigen and protective antigen [50] , with amino acids 441 - 700 as the major immunodominant epitope [51] .

Early antibodies are detected in some patients within two weeks . [Severe acute respiratory syndrome and dentistry A retrospective view, *The Journal of the American Dental Association*, 2004-09-30]

> Enzyme - linked immunosorbent assay , or ELISA , test . From about 20 days after the onset of clinical signs , ELISA tests can be used to detect immunoglobulin , or Ig , M and IgA antibodies in the serum samples of patients with SARS . Early antibodies are detected in some patients within two weeks .

System to collect answers of Covid-related questions in scientific publications
Winner in two competitions (White House, NIH)

Q.A.: Lots of practical application


Smart Speaker Use Case Frequency January 2020
Listen to streaming music service 88.7% 73.6% 39.8%
Ask a question 83.1% 66.2% 29.4%
Check the weather 77.1% 59.8% 33.9%
Set a timer 64.5% 52.4% 20.3%
Set an alarm 59.8% 45.6% 26.3%
Listen to the radio 59.8% 42.6% 19.0%
Listen to News / Sports 50.6% 37.7% 16.9%
Use a favorite Alexa skill or Google Action 47.9% 34.0% 16.4%
Play game or answer trivia 46.1% 27.7% 9.0%
Listen to Podcast or other talk formats 44.9% 32.0% 11.4%
Control smart home devices 43.4% 31.9% 24.5%
Find a recipe or cooking instructions 42.3% 26.0% 5.4%
Call someone 40.2% 21.2% 9.5%
Search for product information 38.2% 27.9% 7.3%
Check traffic / directions 35.1% 23.7% 11.1%
Access my calendar 32.1% 19.0% 9.5%
Send a text message 27.8% 14.0% 6.7%
Make a purchase 25.2% 14.3% 4.9%
EVER TRIED
MONTHLY
DAILY

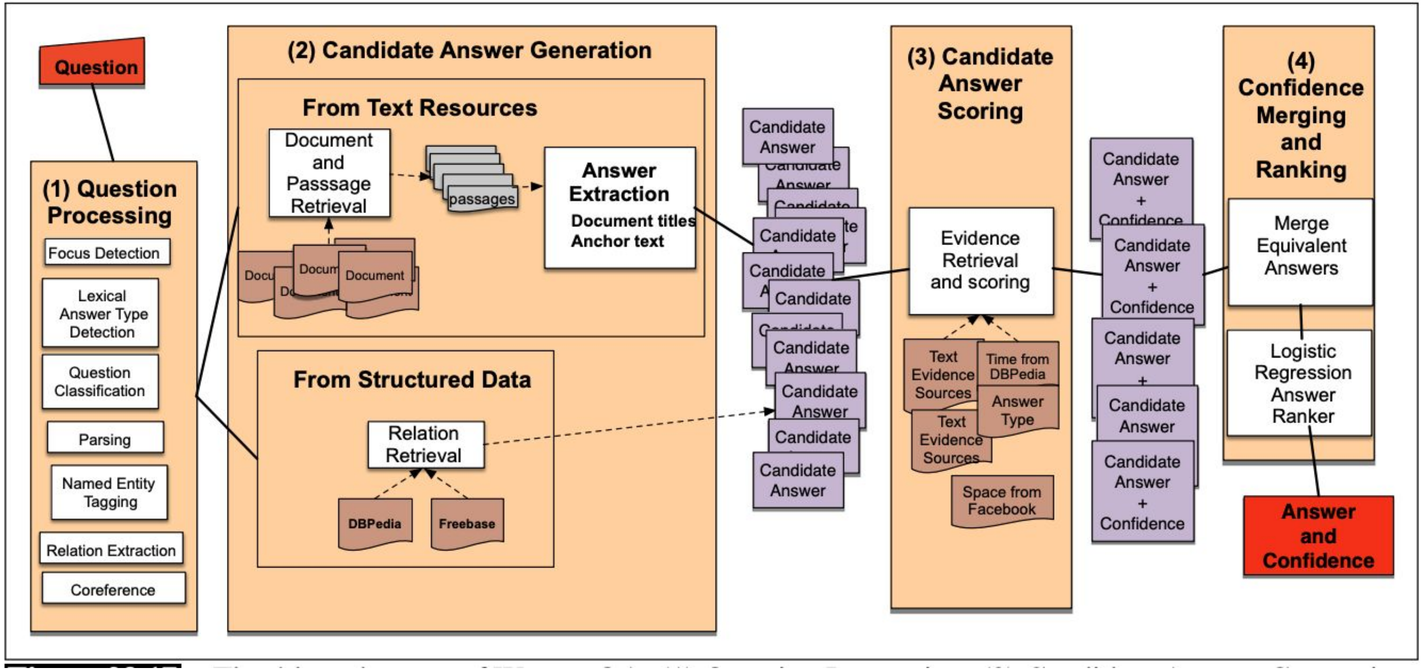
voicebot.ai

Source: Voicebot.ai 2020

Q.A.: Lots of practical application

IBM Watson defeated two of Jeopardy's greatest champions in 2011

Q.A.: Lots of practical application (IBM watson 2011)



(1) Question processing, (2) Candidate answer generation
(3) Candidate answer scoring, and (4) Confidence merging and ranking.

Q.A.: Lots of practical application (Deep Learning era)

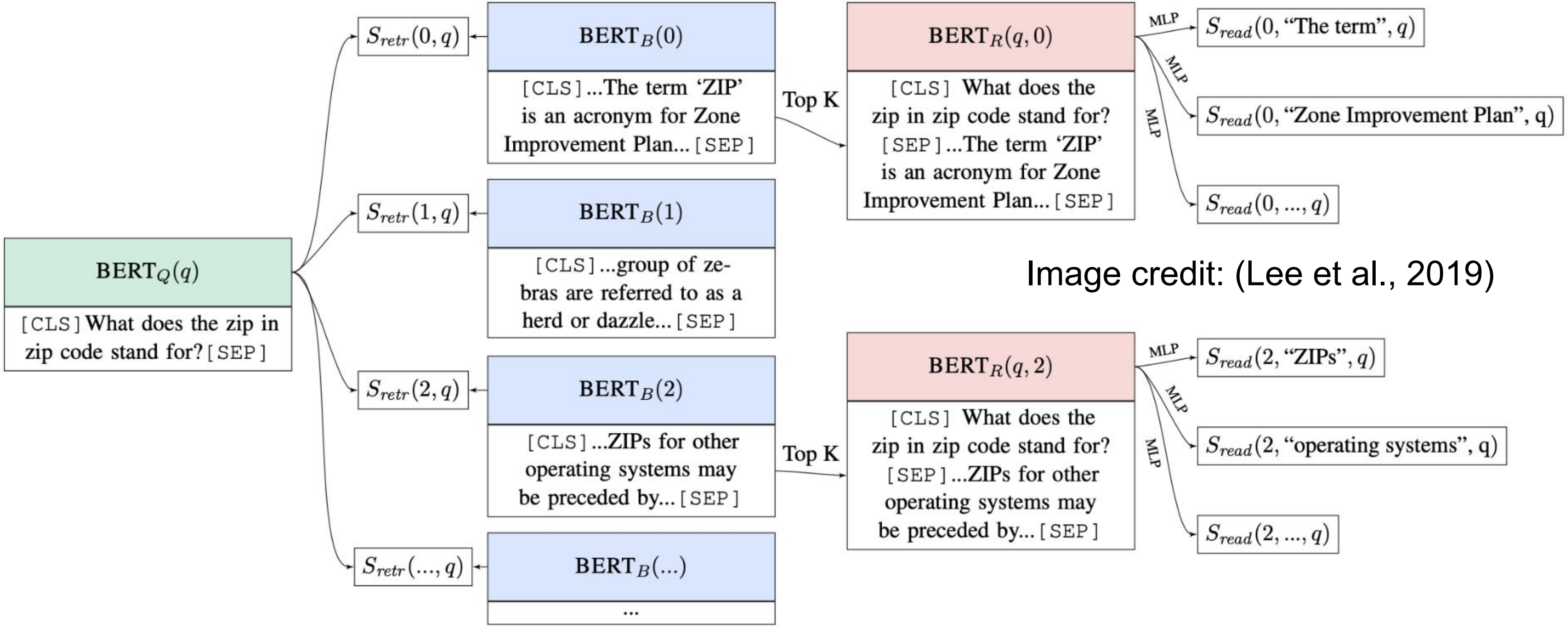
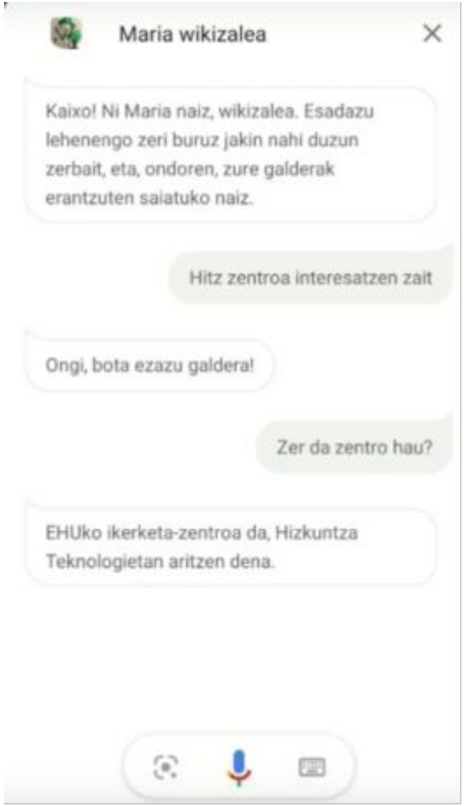

$S_{retr}(0, q)$
BERT$_B$(0)
[CLS]...The term 'ZIP' is an acronym for Zone Improvement Plan...[SEP]
Top K
BERT$_R$(q, 0)
[CLS] What does the zip in zip code stand for? [SEP]...The term 'ZIP' is an acronym for Zone Improvement Plan...[SEP]
MLP
S_{read}(0, "The term", q)
MLP
S_{read}(0, "Zone Improvement Plan", q)
MLP
S_{read}(0, ..., q)
$S_{retr}(1, q)$
BERT$_B$(1)
[CLS]...group of zebras are referred to as a herd or dazzle...[SEP]
BERT$_Q$(q)
[CLS]What does the zip in zip code stand for?[SEP]
$S_{retr}(2, q)$
BERT$_B$(2)
[CLS]...ZIPs for other operating systems may be preceded by...[SEP]
Top K
BERT$_R$(q, 2)
[CLS] What does the zip in zip code stand for? [SEP]...ZIPs for other operating systems may be preceded by...[SEP]
MLP
S_{read}(2, "ZIPs", q)
MLP
S_{read}(2, "operating systems", q)
MLP
S_{read}(2, ..., q)
$S_{retr}(..., q)$
BERT$_B$(...)
...

Image credit: (Lee et al., 2019)

Almost all the state-of-the-art question answering systems are built on top of end-to-end training and pre-trained language models (e.g., BERT)!

Q.A.: Lots of practical application (Deep Learning era)



Maria wikizalea

Kaixo! Ni Maria naiz, wikizalea. Esadazu lehenengo zeri buruz jakin nahi duzun zerbait, eta, ondoren, zure galderak erantzuten saiatuko naiz.

Hitz zentroa interesatzen zait

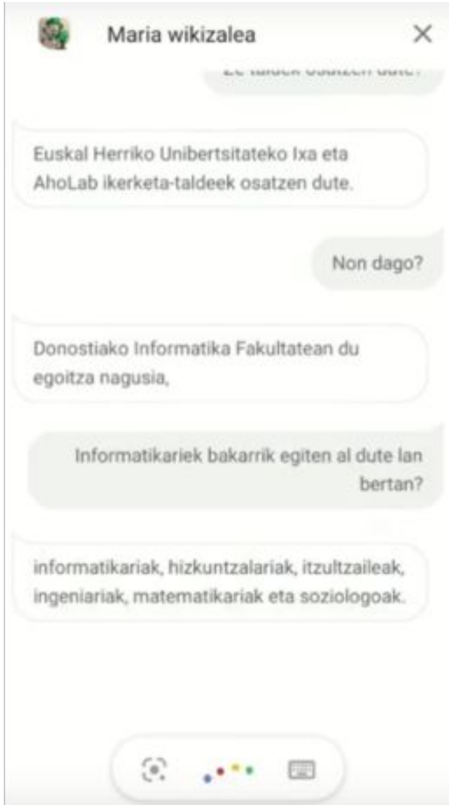
Ongi, bota ezazu galdera!

Zer da zentro hau?

EHUko ikerketa-zentroa da, Hizkuntza Teknologietan aritzen dena.

Conversational system that replies to questions about Wikipedia, in any language

Conversational QA



Maria wikizalea

Euskal Herriko Unibertsitateko Ixa eta AhoLab ikerketa-taldeek osatzen dute.

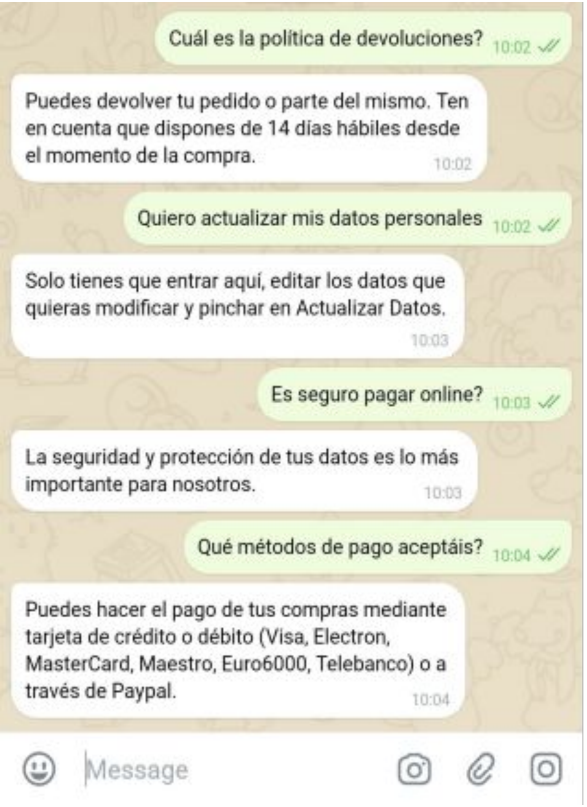
Non dago?

Donostiako Informatika Fakultatean du egoitza nagusia,

Informatikariek bakarrik egiten al dute lan bertan?

informatikariak, hizkuntzalariak, itzultzaileak, ingeniariak, matematikariak eta soziologoak.

Q.A.: Lots of practical application (Deep Learning era)



Cuál es la política de devoluciones? 10:02

Puedes devolver tu pedido o parte del mismo. Ten en cuenta que dispones de 14 días hábiles desde el momento de la compra. 10:02

Quiero actualizar mis datos personales 10:02

Solo tienes que entrar aquí, editar los datos que quieras modificar y pinchar en Actualizar Datos. 10:03

Es seguro pagar online? 10:03

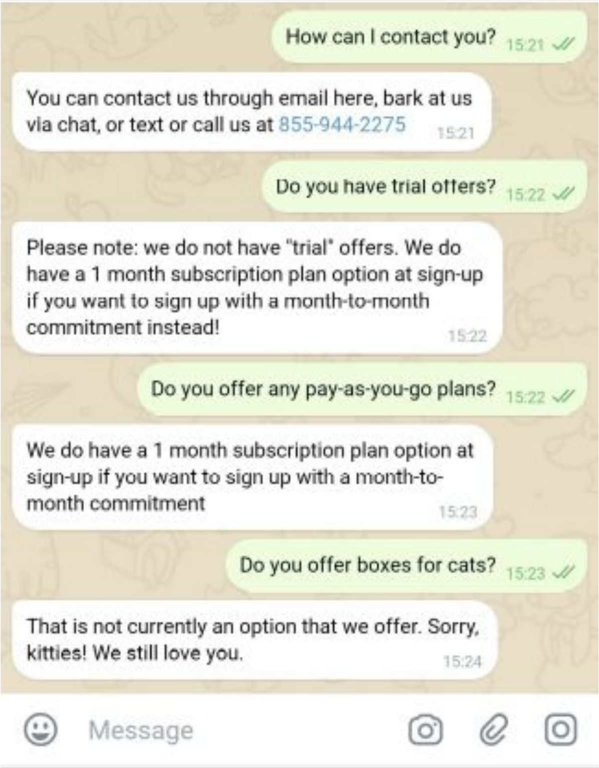
La seguridad y protección de tus datos es lo más importante para nosotros. 10:03

Qué métodos de pago aceptáis? 10:04

Puedes hacer el pago de tus compras mediante tarjeta de crédito o débito (Visa, Electron, MasterCard, Maestro, Euro6000, Telebanco) o a través de Paypal. 10:04

QA system that replies to questions using FAQ repositories, in any language.

Question Answering



How can I contact you? 15:21

You can contact us through email here, bark at us via chat, or text or call us at 855-944-2275 15:21

Do you have trial offers? 15:22

Please note: we do not have "trial" offers. We do have a 1 month subscription plan option at sign-up if you want to sign up with a month-to-month commitment instead! 15:22

Do you offer any pay-as-you-go plans? 15:22

We do have a 1 month subscription plan option at sign-up if you want to sign up with a month-to-month commitment 15:23

Do you offer boxes for cats? 15:23

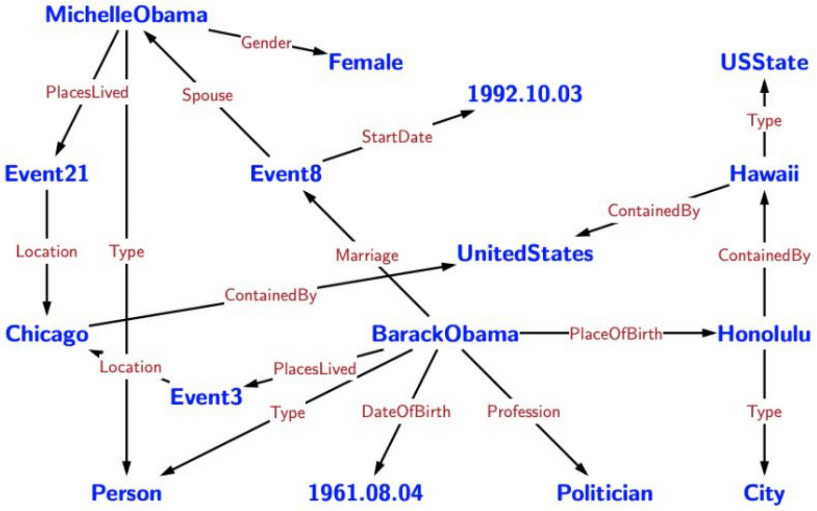
That is not currently an option that we offer. Sorry, kitties! We still love you. 15:24

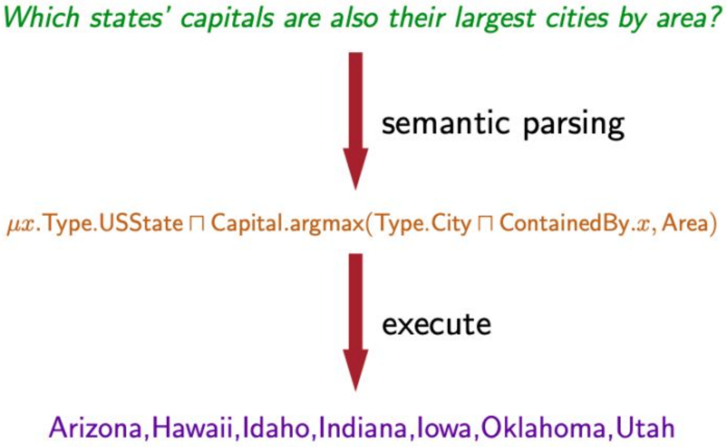
Beyond textual Q.A. problems

Today, we will mostly focus on how to answer questions based on **unstructured text.**

Freebase

100M entities (nodes) 1B assertions (edges)





Which states' capitals are also their largest cities by area?

semantic parsing

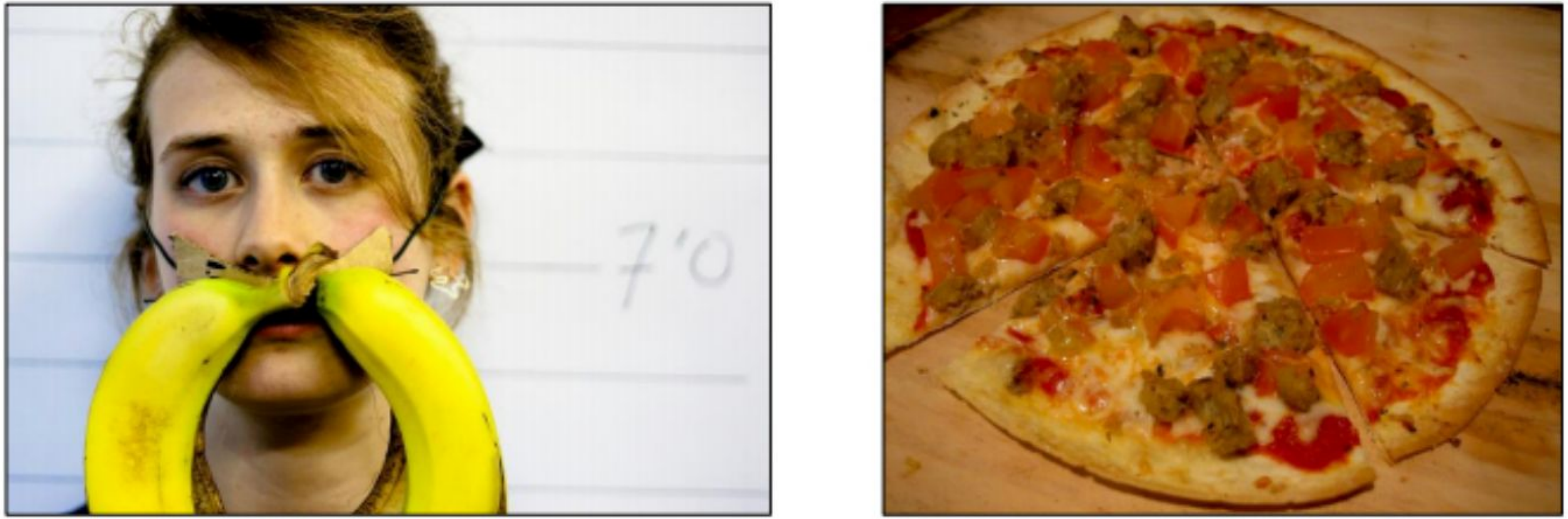
μx.Type.USState ⊓ Capital.argmax(Type.City ⊓ ContainedBy.x, Area)

execute

Arizona,Hawaii,Idaho,Indiana,Iowa,Oklahoma,Utah

Beyond textual Q.A. problems

Today, we will mostly focus on how to answer questions based on **unstructured text.**



What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

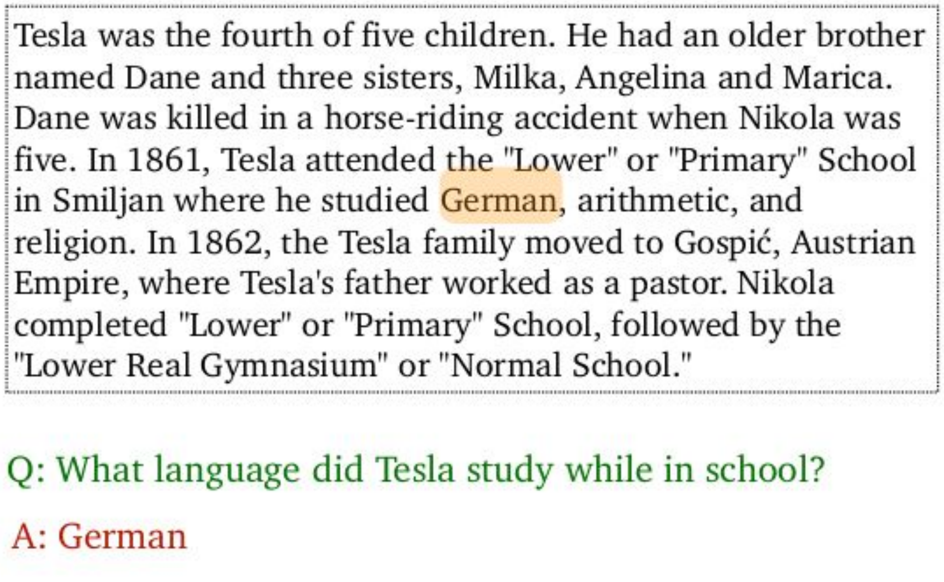
(Antol et al., 2015): Visual Question Answering

Course Outline

- **Question Answering systems (Q.A.)**
  - Introduction
  - QA practical applications
  - **Reading comprehension**
  - SQuAD dataset
  - QA models
  - Is reading comprehension solved?

Reading comprehension

Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) ⟶ A



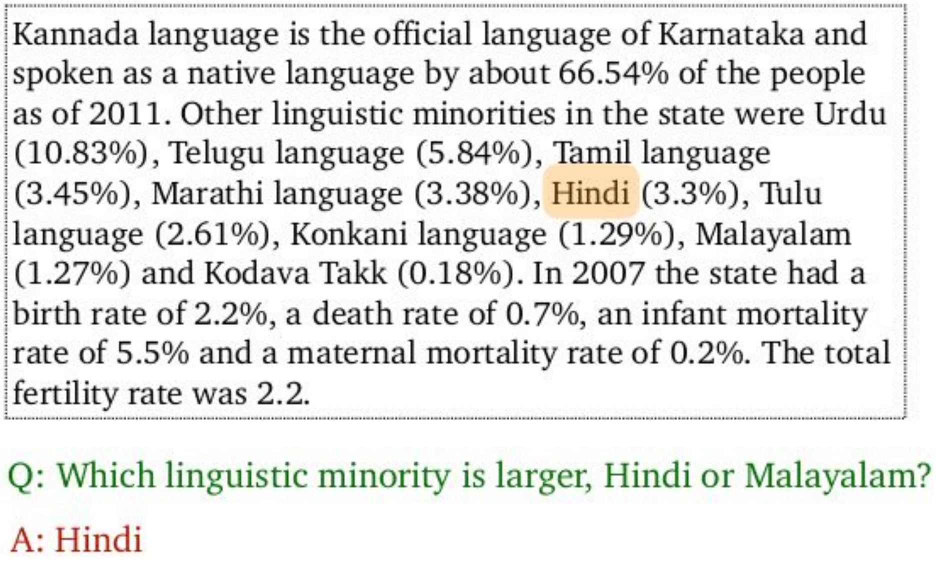
Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

Q: What language did Tesla study while in school?

A: German

Reading comprehension

Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) ⟶ A



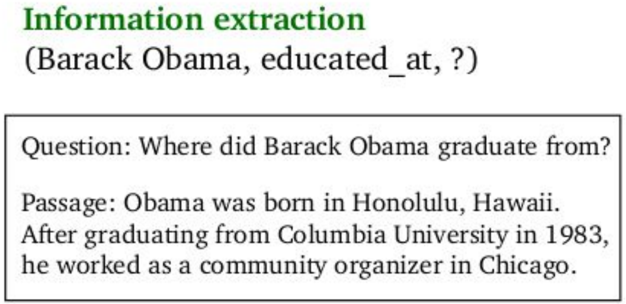
Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?

A: Hindi

Why do we care about this problem?

- Useful for many practical applications
- Reading comprehension is an important testbed for evaluating how well computer systems understand human language
  - Wendy Lehnert 1977: “Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding.”
- Many other NLP tasks can be reduced to a reading comprehension problem:



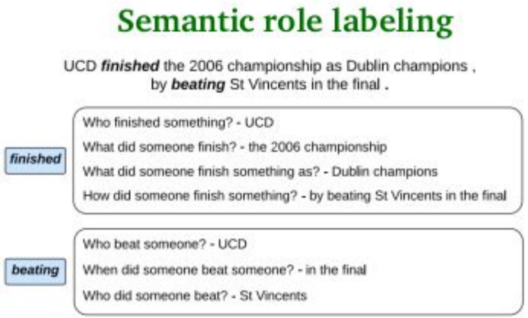
Information extraction

(Barack Obama, educated_at, ?)

Question: Where did Barack Obama graduate from?

Passage: Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.

(Levy et al., 2017)



Semantic role labeling

UCD ***finished*** the 2006 championship as Dublin champions , by ***beating*** St Vincents in the final .

finished

Who finished something? - UCD

What did someone finish? - the 2006 championship

What did someone finish something as? - Dublin champions

How did someone finish something? - by beating St Vincents in the final

beating

Who beat someone? - UCD

When did someone beat someone? - in the final

Who did someone beat? - St Vincents

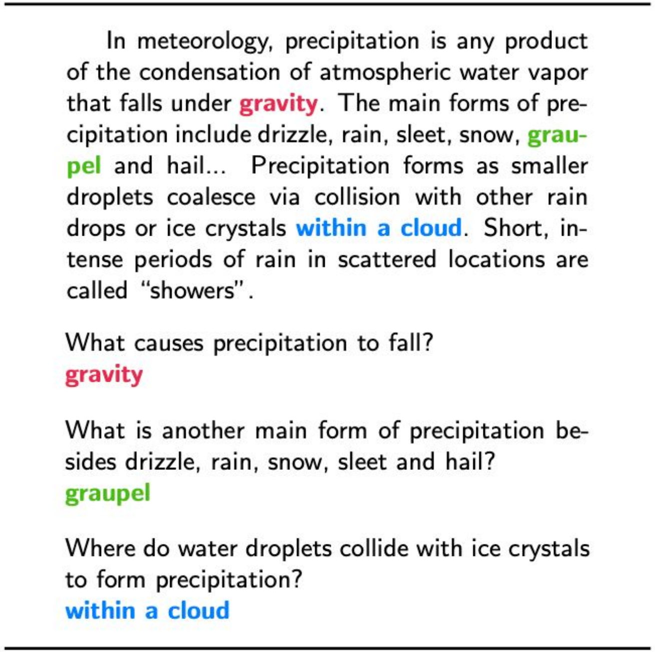
(He et al., 2015)

Course Outline

- **Question Answering systems (Q.A.)**
  - Introduction
  - QA practical applications
  - Reading comprehension
  - **SQuAD dataset**
  - QA models
  - Is reading comprehension solved?

Stanford question answering dataset (SQuAD)

- 100k annotated (passage, question, answer) triples
  - Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!
- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is a short segment of text (or span) in the passage.
  - This is a limitation— not all the questions can be
  - answered in this way!
- SQuAD still remains the most popular reading comprehension dataset; it is “almost solved” today and the state-of-the-art exceeds the estimated human performance.



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

(Rajpurkar et al., 2016): SQuAD: 100,000+ Questions for Machine Comprehension

Stanford question answering dataset (SQuAD)

- Evaluation: exact match (0 or 1) and F1 (partial credit).
- For development and testing sets, 3 gold answers are collected, because there could be multiple plausible answers.
- We compare the predicted answer to each gold answer (a, an, the, punctuations are removed) and take max scores. Finally, we take the average of all the examples for both exact match and F1.
- Estimated human performance: EM = 82.3, F1 = 91.2

Q: What did Tesla do in December 1878?
A: {left Graz, left Graz, left Graz and severed all relations with his family}
Prediction: {left Graz and served}

Exact match: max{0, 0, 0} = 0
F1: max{0.67, 0.67, 0.61} = 0.67

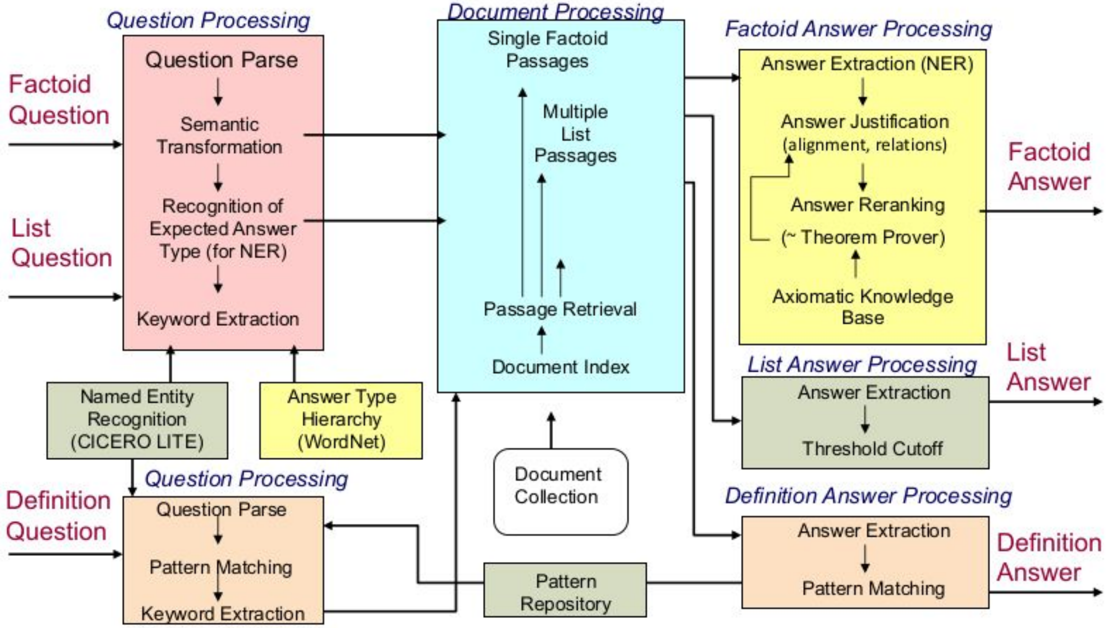
(Rajpurkar et al., 2016): SQuAD: 100,000+ Questions for Machine Comprehension

Course Outline

- **Question Answering systems (Q.A.)**
  - Introduction
  - QA practical applications
  - Reading comprehension
  - SQuAD dataset
  - **QA models (BiDAF)**
  - Is reading comprehension solved?

Pre deep learning era for reading comprehension

- Complex pipelined models but they did work fairly well on “factoid” questions
  - architecture of LCC (Harabagiu/Moldovan) QA system, circa 2003



Neural Models for reading comprehension

- **How can we build a model to solve SQuAD?**
  - Problem formulation
    - Input:

      $$C = (c_1, c_2, \ldots, c_N),\ Q = (q_1, q_2, \ldots, q_M),\ c_i, q_i \in V$$

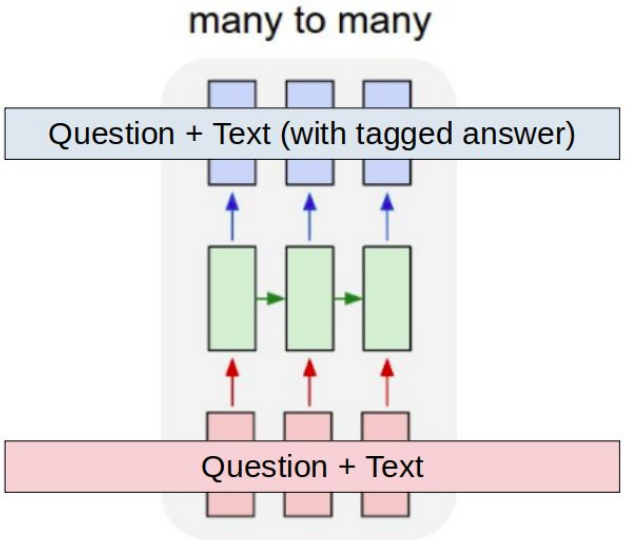
      - c: passage or document
      - q: question or query
    - Output: 1 ≤ start ≤ end ≤ N (answer is a spam in the passage)

Neural Models for reading comprehension

- **How can we build a model to solve SQuAD?**

start & end spams

o o o [sep] o o o o o o ***a a a a*** *o o o o ...*



Question + Text (with tagged answer)

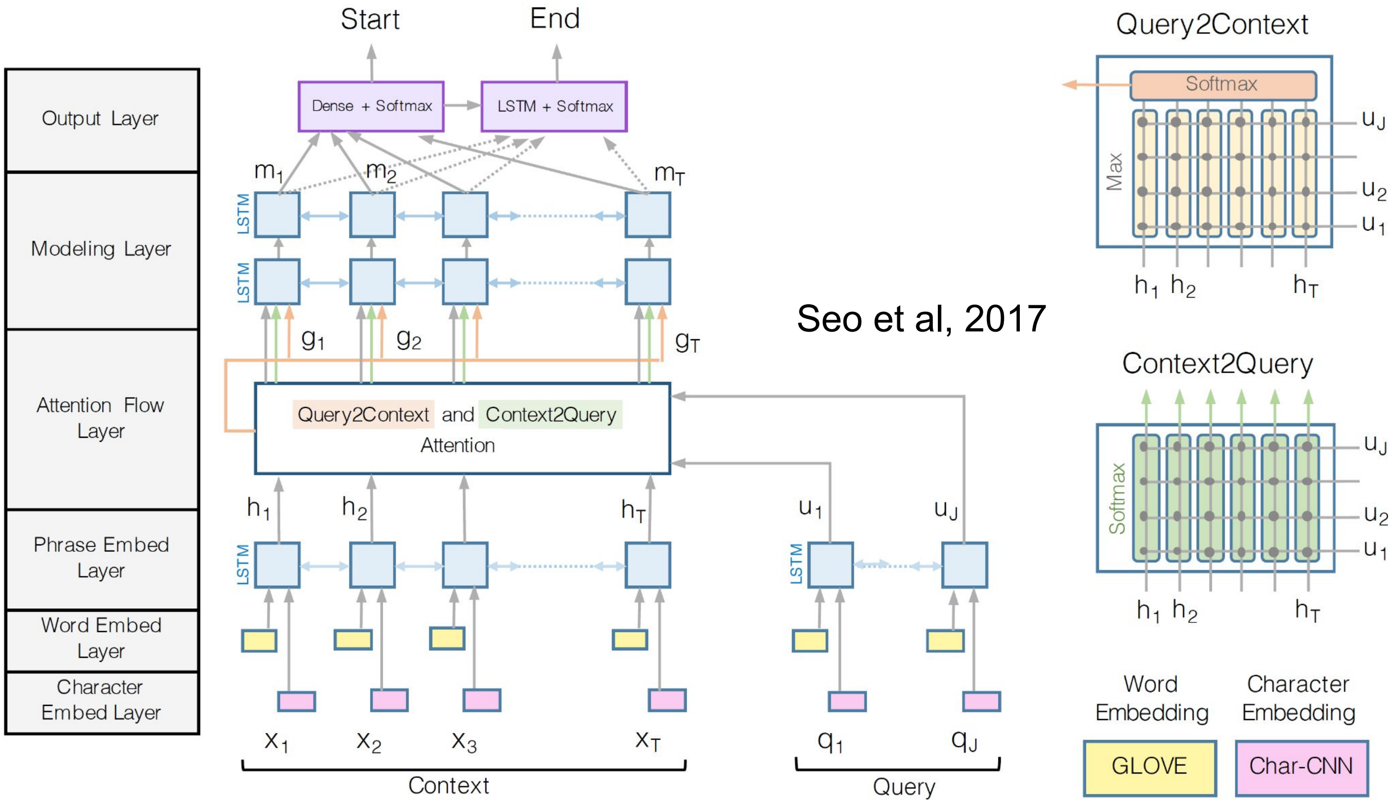
Question + Text

q q q [sep] w w w w w w w w w w w ...

Neural Models for reading comprehension

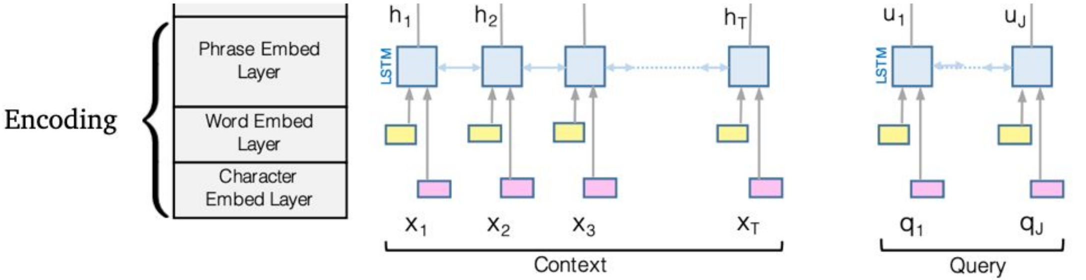
- **How can we build a model to solve SQuAD?**
  - Problem formulation
    - Input: C = (c1 , c2 , ..., cN) , Q = (q1 , q2 , ..., qM) , ci , qi ∈ V
      - c: passage or document
      - q: question or query
    - Output: 1 ≤ start ≤ end ≤ N (answer is a spam in the passage)
  - A family of LSTM-based models with attention (2016-2018)
    - Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), Match-LSTM (Wang et al., 2017), **BiDAF** (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)..
  - Fine-tuning **BERT-like** models for reading comprehension (2019+)

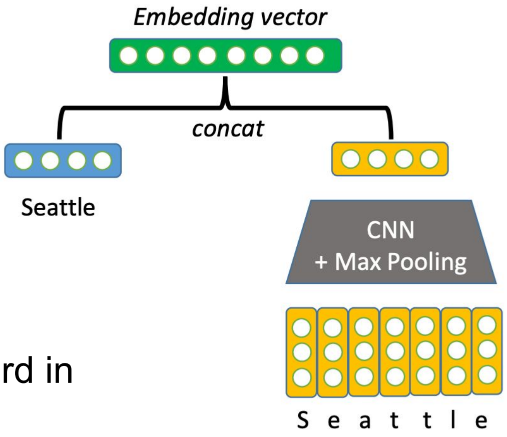
LSTM-based BiDAF: Bidirectional Attention Flow model



Seo et al, 2017

LSTM-based BiDAF: Encoding





- Concatenation of word and character embeddings for each word in context and query (Glove + CNN over chars)

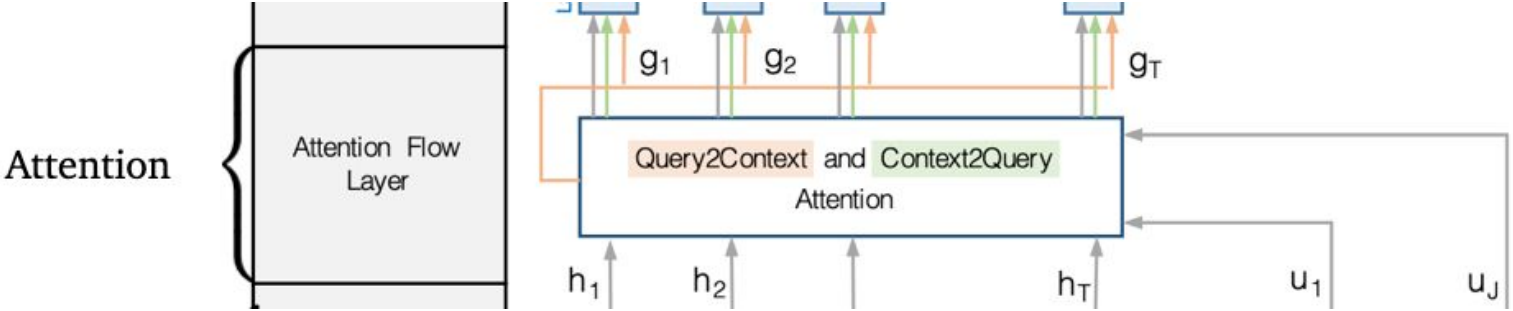
$$e(c_i) = f([\mathrm{GloVe}(c_i); \mathrm{charEmb}(c_i)]) \qquad e(q_i) = f([\mathrm{GloVe}(q_i); \mathrm{charEmb}(q_i)])$$

- 2x Bidirectional LSTMs separately to produce contextual embeddings for both context and query.

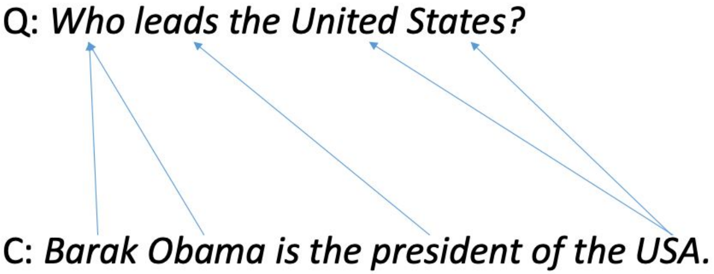
$$\begin{aligned}
\overrightarrow{\mathbf{c}}_i &= \mathrm{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H \\
\overleftarrow{\mathbf{c}}_i &= \mathrm{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H \\
\mathbf{c}_i &= [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H}
\end{aligned}$$

$$\begin{aligned}
\overrightarrow{\mathbf{q}}_i &= \mathrm{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H \\
\overleftarrow{\mathbf{q}}_i &= \mathrm{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H \\
\mathbf{q}_i &= [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}
\end{aligned}$$

LSTM-based BiDAF: Attention



- Context2query attention: For each context word, choose the most relevat words from the query words:

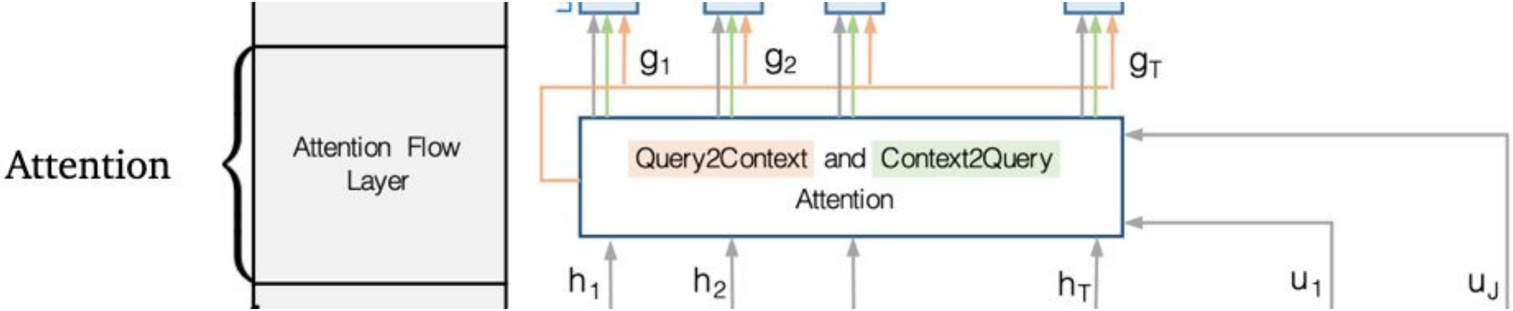


Q: *Who leads the United States?*

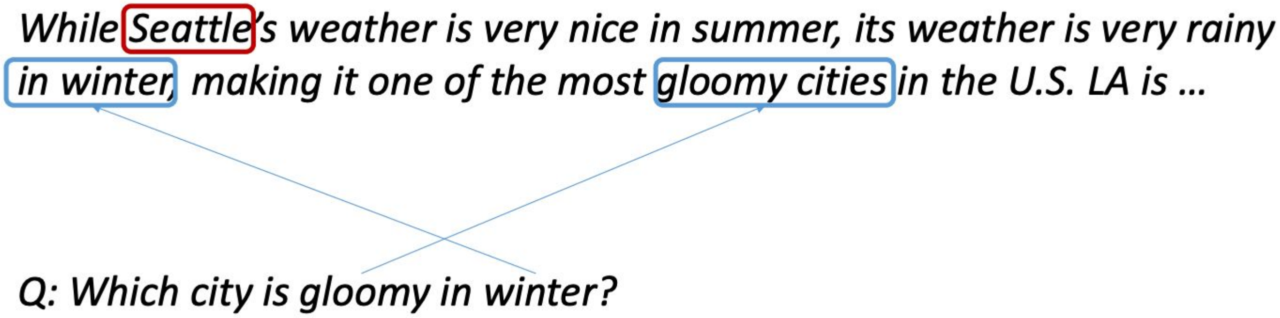
C: *Barak Obama is the president of the USA.*

For each context word, find the most relevant query word.

LSTM-based BiDAF: Attention



- Query2context attention: choose the context words that are most relevant to one of query words.



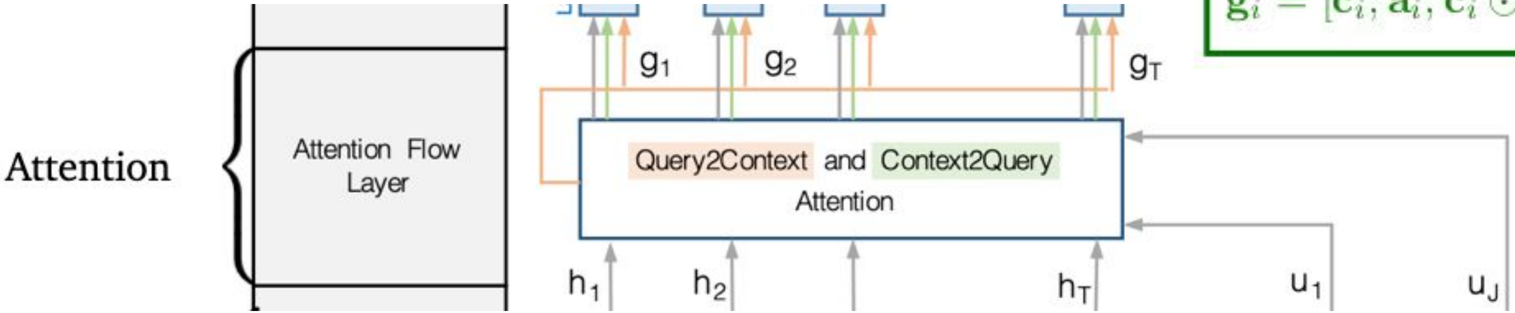
While Seattle's weather is very nice in summer, its weather is very rainy in winter, making it one of the most gloomy cities in the U.S. LA is …

Q: Which city is gloomy in winter?

LSTM-based BiDAF: Attention

The final output is

$$\mathbf{g}_i = [\mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}] \in \mathbb{R}^{8H}$$



- First, compute a similarity score for every pair of (c,q):

$$S_{i,j} = \mathbf{w}_{\text{sim}}^{\mathsf{T}}[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j] \in \mathbb{R} \qquad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6H}$$

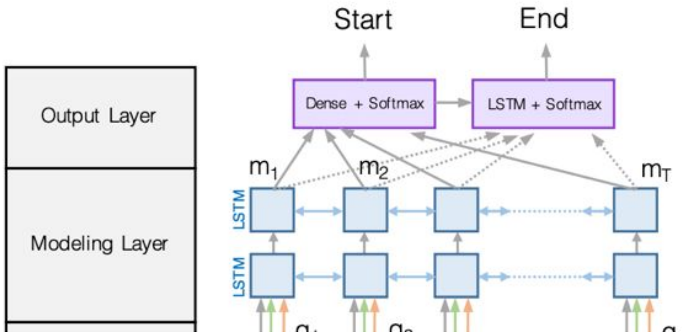
- Context2query attention:

$$\alpha_{i,j} = \text{softmax}_j(S_{i,j}) \in \mathbb{R} \qquad \mathbf{a}_i = \sum_{j=1}^{M} \alpha_{i,j}\mathbf{q}_j \in \mathbb{R}^{2H}$$

- Query2context attention:

$$\beta_i = \text{softmax}_i(\max_{j=1}^{M}(S_{i,j})) \in \mathbb{R}^N \qquad \mathbf{b} = \sum_{i=1}^{N} \beta_i \mathbf{c}_i \in \mathbb{R}^{2H}$$

LSTM-based BiDAF: Modeling and output layers



- **Modeling layer**: pass gi to another two layers of **bi-directional** LSTMs.
    - Attention layer is modeling interactions between query and context
    - Modeling layer is modeling interactions within context words

$$\mathbf{m}_i = \mathrm{BiLSTM}(\mathbf{g}_{\mathrm{i}}) \in \mathbb{R}^{2H}$$

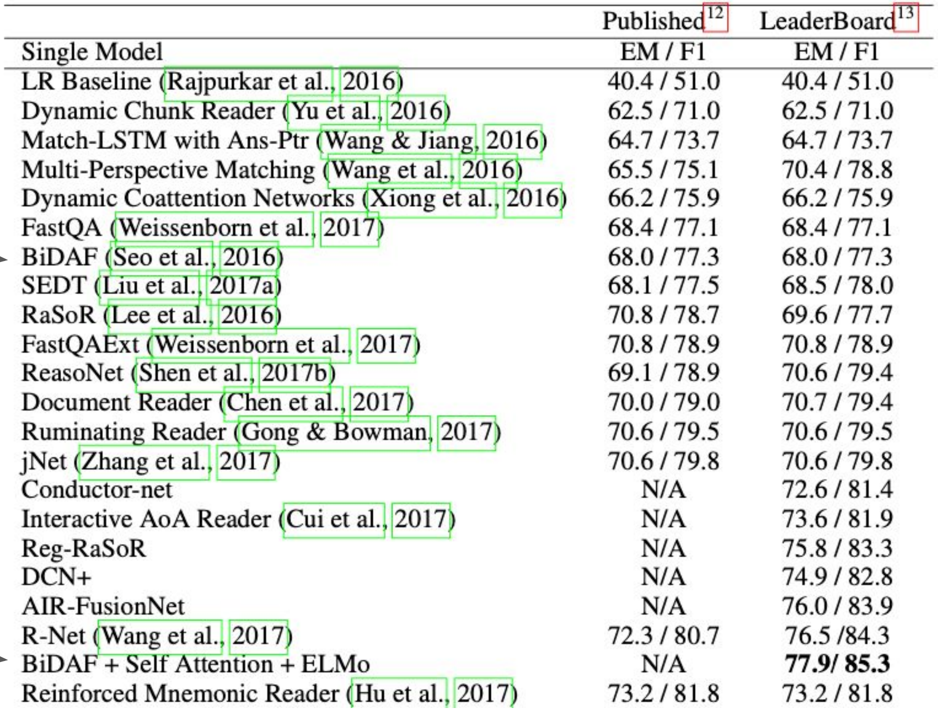
- Output layer: two classifiers to predict start and end positions:

$$p_{\text{start}} = \mathrm{softmax}(\mathbf{w}_{\text{start}}^{\mathsf{T}}[\mathbf{g}_i; \mathbf{m}_i]) \qquad p_{\text{end}} = \mathrm{softmax}(\mathbf{w}_{\text{end}}^{\mathsf{T}}[\mathbf{g}_i; \mathbf{m}'_i])$$

$$\mathbf{m}'_i = \mathrm{BiLSTM}(\mathbf{m}_i) \in \mathbb{R}^{2H} \quad \mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}} \in \mathbb{R}^{10H}$$

LSTM-based BiDAF: Performance on SQuAD

- This model achieved 77.3 F1 on SQuAD v1.1.
  - Without context-to-query attention
    - ⇒ 67.7 F1
  - Without query-to-context attention
    - ⇒ 73.7 F1
  - Without character embeddings
    - ⇒ 75.4 F1



	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

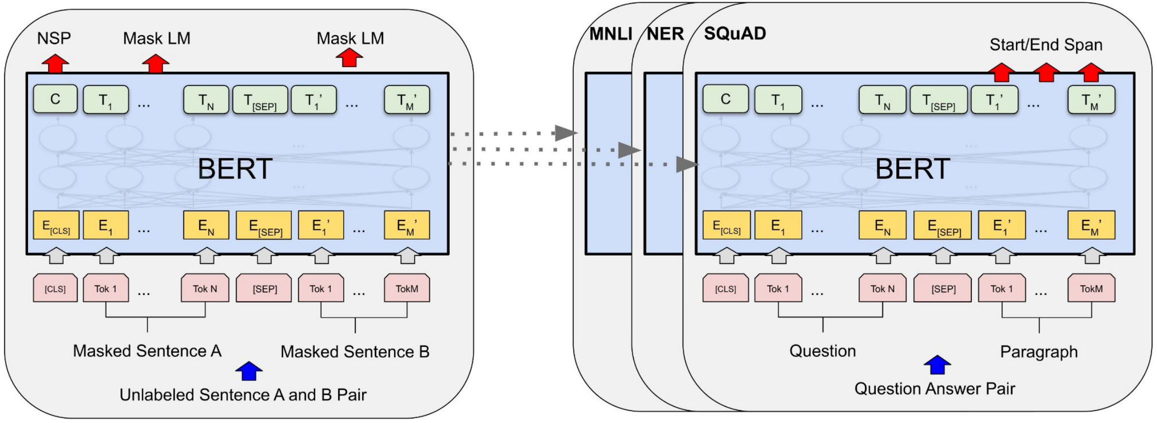
(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

Course Outline

- **Question Answering systems (Q.A.)**
  - Introduction
  - QA practical applications
  - Reading comprehension
  - SQuAD dataset
  - **QA models (BERT)**
  - Is reading comprehension solved?

BERT for reading comprehension

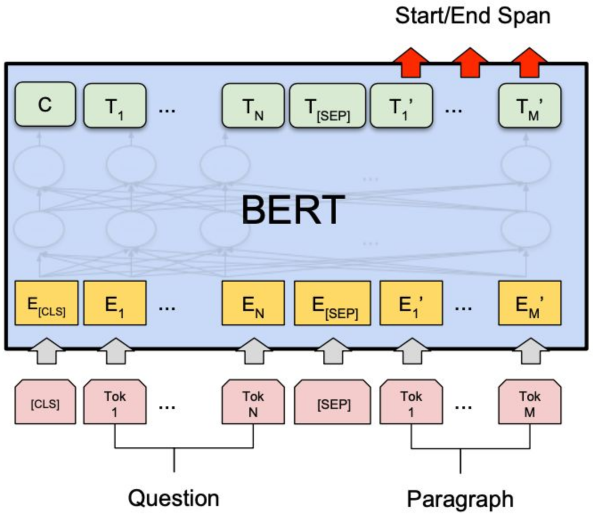
- **BERT** is a deep bidirectional Transformer encoder pre-trained on large amounts of text (Wikipedia + BooksCorpus)
- BERT is pre-trained on two training objectives:
  - Masked language model (MLM)
  - Next sentence prediction (NSP)
- BERT base has 12 layers and 110M parameters, large 24 layers and 330M parameters



Pre-training

Fine-Tuning

BERT for reading comprehension

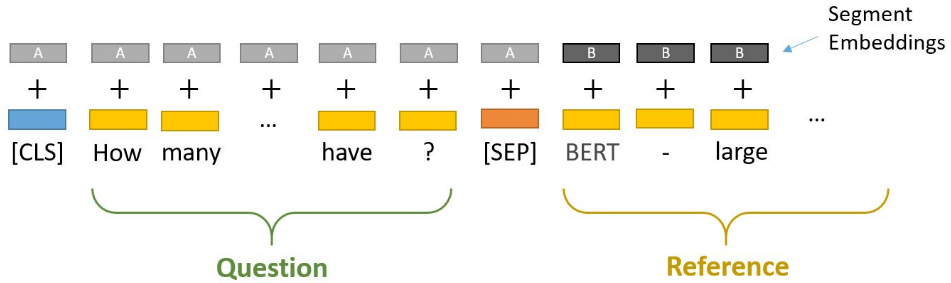


$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\intercal}\mathbf{H})$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\intercal}\mathbf{H})$$

where $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]$ are the hidden vectors of the paragraph, returned by BERT

Question = Segment A
Passage = Segment B
Answer = predicting two endpoints in segment B

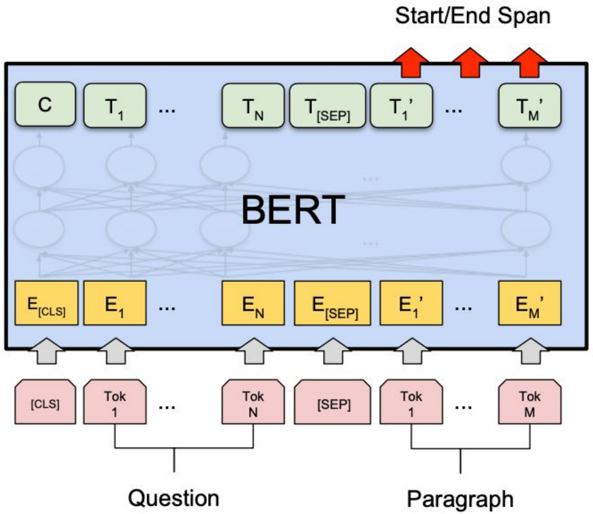


Question: How many parameters does BERT-large have?

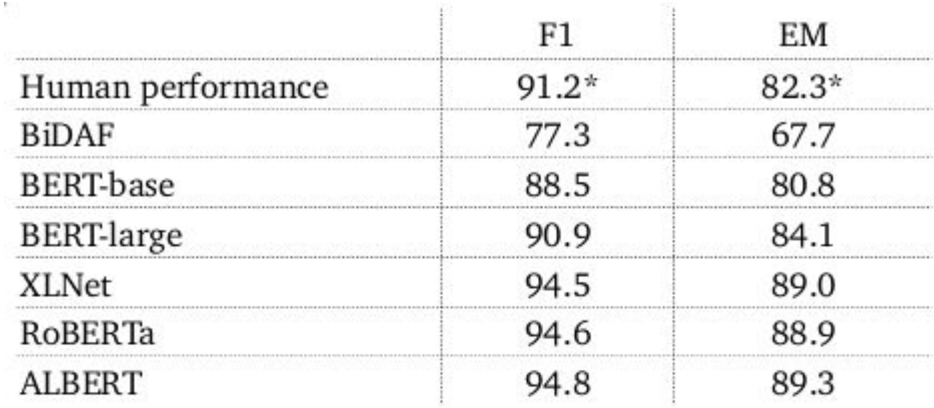
Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

Image credit: https://mccormickml.com/

BERT for reading comprehension



- All the BERT parameters (e.g., 110M) as well as the newly introduced parameters h start , h end (e.g., 768 x 2 = 1536) are optimized together
- **It works amazingly well.** Stronger pre-trained language models can lead to even better performance and SQuAD becomes a standard dataset for testing pre-trained models.



	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

+DistilBERT smaller, faster, cheaper and lighter (lab time)

BiDAF Vs BERT models

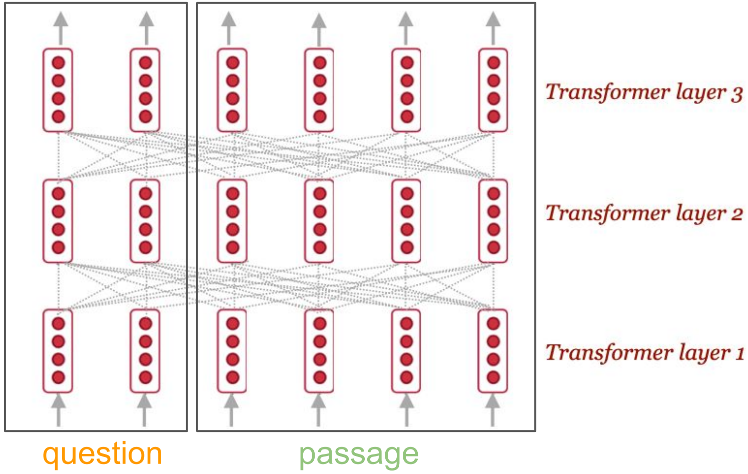
- **BERT** model has many many more parameters (110M or 330M) and BiDAF has ~2.5M parameters.

- **BiDAF** is built on top of several bidirectional LSTMs while BERT is built on top of Transformers (no recurrence architecture and easier to parallelize).

- **BERT** is pre-trained while BiDAF is only built on top of GloVe (and all the remaining parameters need to be learned from the supervision datasets).
    - Pre-training is clearly a game changer but it is expensive..

BiDAF Vs BERT models

- Are they really fundamentally different? Probably not.
  - BiDAF and other models aim to model the interactions between question and passage.
  - BERT uses self-attention between the concatenation of question and passage
  - (Clark and Gardner, 2018) shows that adding a self-attention layer for the passage attention(Q, P) to BiDAF also improves performance.



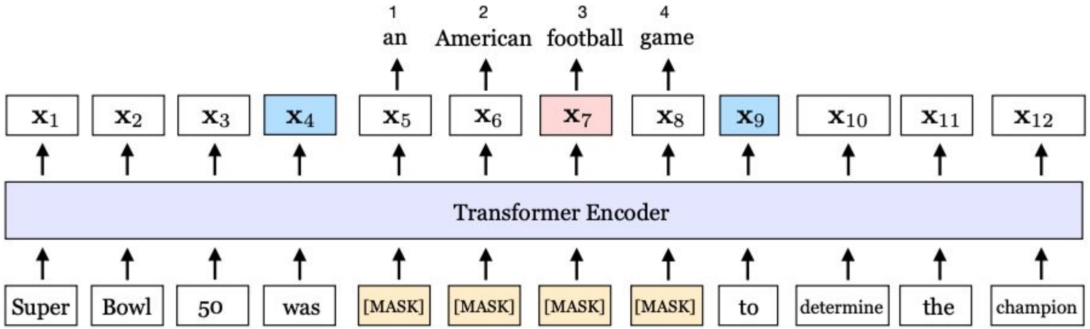
Course Outline

- **Question Answering systems (Q.A.)**
    - Introduction
    - QA practical applications
    - Reading comprehension
    - SQuAD dataset
    - **QA models (spamBERT and better pre-training obj.)**
    - Is reading comprehension solved?

Can we design better pre-training objectives?

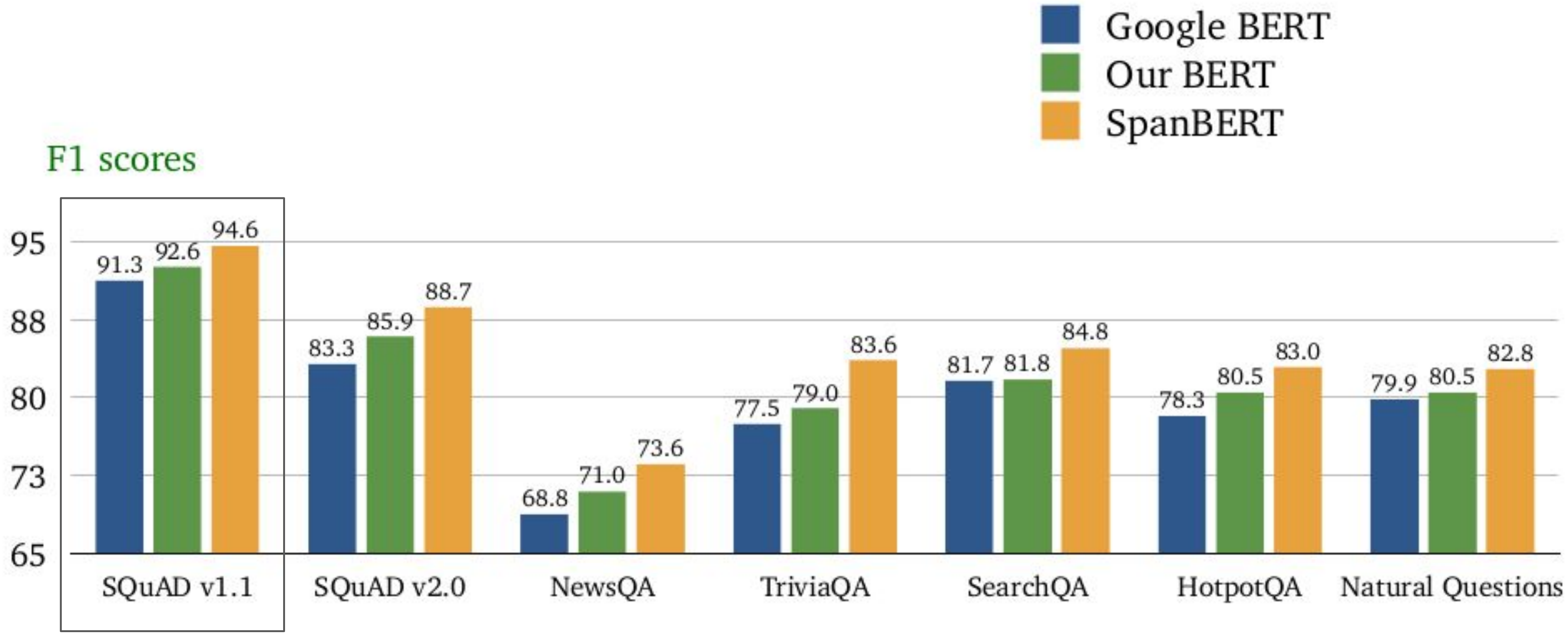
- Of course yes!
- Two ideas by **SpanBERT**:
  - masking contiguous spans of words instead of 15% random words
  - using the two end points of span to predict all the masked words in between = compressing the information of a span into its two endpoints

$$\begin{aligned}\mathcal{L}(\text{football}) &= \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})\\ &= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)\end{aligned}$$



(Joshi & Chen et al., 2020): **SpanBERT:** Improving Pre-training by Representing and Predicting Spans

SpanBERT



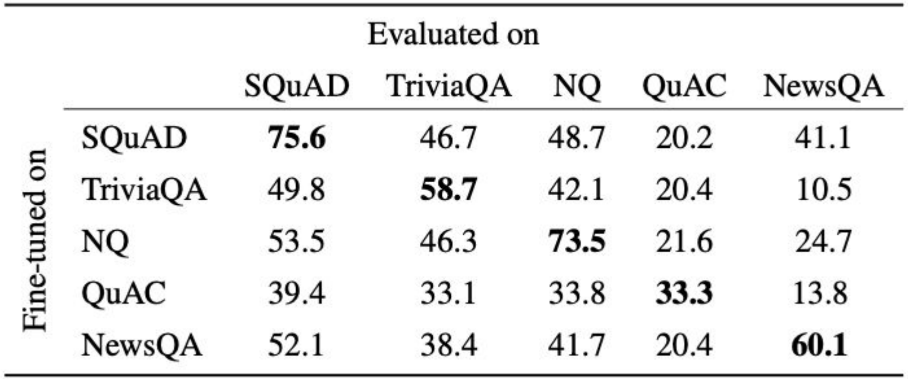
(Joshi & Chen et al., 2020): **SpanBERT:** Improving Pre-training by Representing and Predicting Spans

Course Outline

- **Question Answering systems (Q.A.)**
    - Introduction
    - QA practical applications
    - Reading comprehension
    - SQuAD dataset
    - QA models
    - **Is reading comprehension solved?**

Is reading comprehension solved?

- We have already surpassed human performance on SQuAD. Does it mean that reading comprehension is already solved? **Of course not!**
- The current systems still perform poorly on adversarial examples or examples from out-of-domain distributions

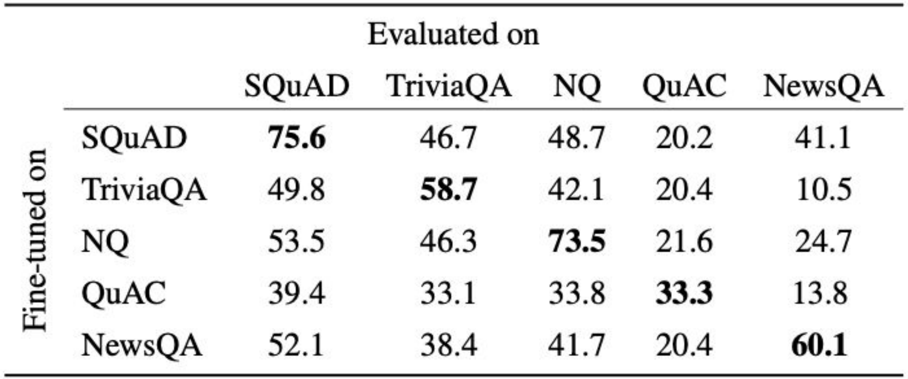


		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
Fine-tuned on	SQuAD	**75.6**	46.7	48.7	20.2	41.1
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5
	NQ	53.5	46.3	**73.5**	21.6	24.7
	QuAC	39.4	33.1	33.8	**33.3**	13.8
	NewsQA	52.1	38.4	41.7	20.4	**60.1**

(Sen and Saffari, 2020): What do Models Learn from Question Answering Datasets?
(Jia and Liang, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems

Is reading comprehension solved?

- Why not train in **SQuAD + QuAC + TriviaQA…** ?
  - Results in QuAC test: **Train QuAC > Train QuAC+SQuAD**
  - Results in SQuAD test: **Train SQuAD ~= Train QuAC+SQuAD**

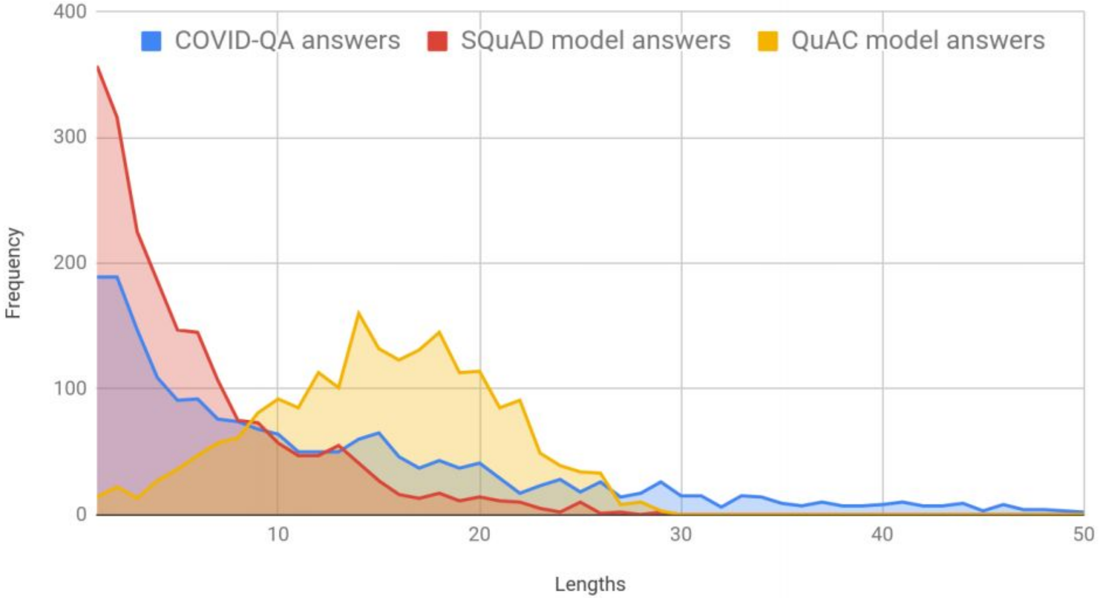


		Evaluated on				
		SQuAD	TriviaQA	NQ	QuAC	NewsQA
Fine-tuned on	SQuAD	**75.6**	46.7	48.7	20.2	41.1
	TriviaQA	49.8	**58.7**	42.1	20.4	10.5
	NQ	53.5	46.3	**73.5**	21.6	24.7
	QuAC	39.4	33.1	33.8	**33.3**	13.8
	NewsQA	52.1	38.4	41.7	20.4	**60.1**

(Sen and Saffari, 2020): What do Models Learn from Question Answering Datasets?
(Jia and Liang, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems

Is reading comprehension solved?

- Why not train in **SQuAD + QuAC + TriviaQA…** ?
    - Automatic Evaluation vs. User Preference in Neural Textual Question Answering over COVID-19 Scientific Literature



Is reading comprehension solved?

- List of Q.A. datasets: http://nlpprogress.com/english/question_answering.html
- We will focus on SQuAD: https://rajpurkar.github.io/SQuAD-explorer/
  - SQuAD1.1 benchmark: https://paperswithcode.com/sota/question-answering-on-squad11
  - SQuAD2.0 bechmark: https://paperswithcode.com/sota/question-answering-on-squad20

Thanks!

ander.barrena@ehu.eus
@4nderB

Lab Time! http://ixa2.si.ehu.eus/~jibloleo/nlpapp2

- **[6. Fine-Tune DistilBERT for Question Answering]**
  - Edited from Huggingface notebooks
    - You will find notebooks for sequence labeling, language modeling....
  - Preprocesing for BERT models (DistilBERT)
    - 90% of the code...
  - Fine-tune your own model (skip running this part it takes 3h, but check the code)
    - We will load a fine-tuned model
    - Check runtime (GPU)
  - Test the model and check the results in the SQuAD1.1 leaderboard
    - DistilBERT Vs BERT
    - DistilBERT Vs BiDAF
  - Ask some questions to your model!

- **Next lab we will show how to build a Q.A. system**
  - Monolingual and Cross-Lingual models
  - Test your own questions