Natural Language Processing for the real world

Slav Petrov
on behalf of the
Language Team @ Google Research

Google Research

LXMLS 2011 LISBON MACHINE LEARNING SCHOOL

learning for the web

Home Schedule Speakers Application Practical Information Organizers

Welcome to the 1st Lisbon Machine Learning School!

LxMLS 2011 is over now. Keep posted for LxMLS 2012!
Could not make it to LxMLS? Check the lecture VIDEOS.

LxMLS 2011 took place July 20-25 (*) at Instituto Superior Técnico, a leading Engineering and Science school in Portugal. It was organized jointly by IST, the Instituto de Telecomunicações and the Spoken Language Systems Lab - L2F of INESC-ID.

In its debut year, the topic of the school was ***Learning for the Web***.

The school covers a range of machine learning (ML) Topics, from theory to practice, that are important in solving natural language processing (NLP) problems that arise in the analysis and use of Web data.

July 20-25

INSTITUTO SUPERIOR TÉCNICO
http://www.ist.utl.pt

Recommend 135

NEWS

http://lxmls.it.pt/2011/Home.html

SATURDAY, JULY 23RD

09:00 - 12:30 Morning Lecture (with 30 min coffee break) [SLIDES] [VIDEO]

LECTURE 3: LEARNING STRUCTURED PREDICTORS (XAVIER CARRERAS)

- From HMMs to CRFs: discriminative learning and features
- Structured perceptron, structured SVMs and max-margin Markov networks
- Training and optimization
- Iterative scaling, L-BFGS, perceptron, MIRA, stochastic and batch gradient descent

12:30 - 14:00 Lunch

14:00 - 17:00 Afternoon Labs [PDF]

17:00 - 17:30 Coffee Break

18:00 - 19:30 Evening Talk

PRACTICAL TALK: LANGUAGE IN THE WILD: LEARNING FROM THE WEB TO UNDERSTAND THE WEB (SLAV PETROV)

20:30 Summer School Banquet (at Mercado da Ribeira)

SUNDAY, JULY 24TH

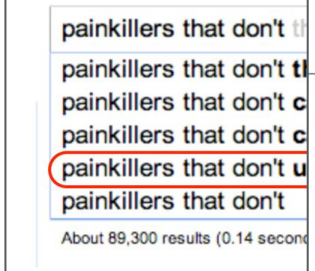
09:00 - 12:30 Morning Lecture (with 30 min coffee break) [SLIDES] [VIDEO]

LECTURE 4: SYNTAX AND PARSING (SLAV PETROV)

http://lxmls.it.pt/2011/Schedule.html

2011

An "easy" query...



... that fails

class negation synonyms

painkillers that don't upset stomach

Search

About 29,700 results (0.15 seconds) Advanced search

Sponsored link

Symptoms of Crohn's
www.crohnsonline.com Do You Know the Signs of Crohn's? Read About Symptoms and More Now.

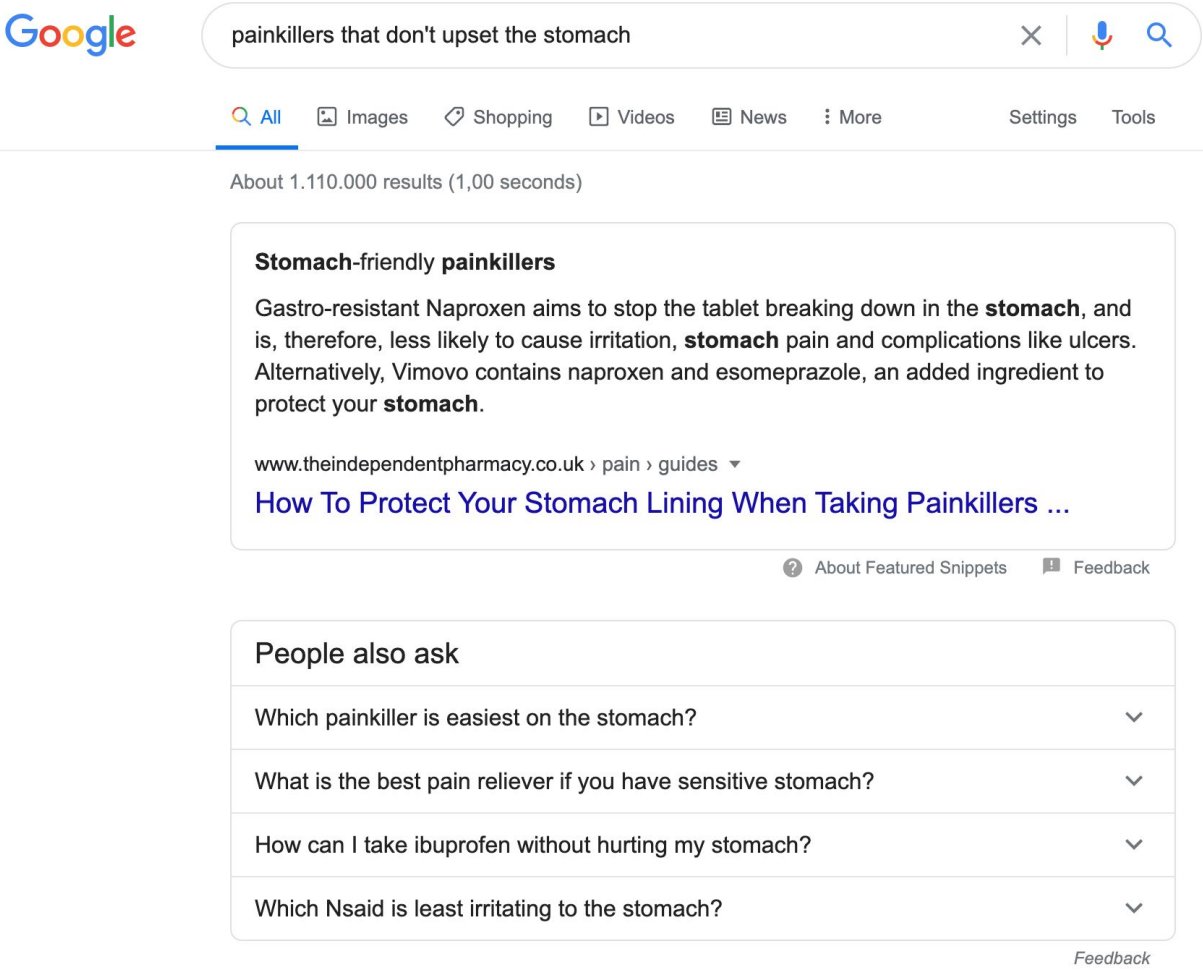
Best Pain Reliever, Over the Counter Pain Relievers, Prescription ...
For **stomach upset**, eating or drinking milk before you take a painkiller can help. ... **Don't** expect **painkillers** to relieve all your pain, but the best pain ...
www.healthblurbs.com/best-pain-reliever-over-the-counter-pain-relievers-prescription-analgesic-**painkillers**/ - Cached - Similar

Why do **painkillers** make your **stomach** hurt? - Yahoo! Answers
Aug 30, 2008 ... **stomach upset**, reflux/indigestion, or if severe ulcers ... I often take **painkillers** FOR **stomach** aches lol. **Don't** do this again. 2 years ago ...
answers.yahoo.com › Health › Other - Health - Cached - Similar

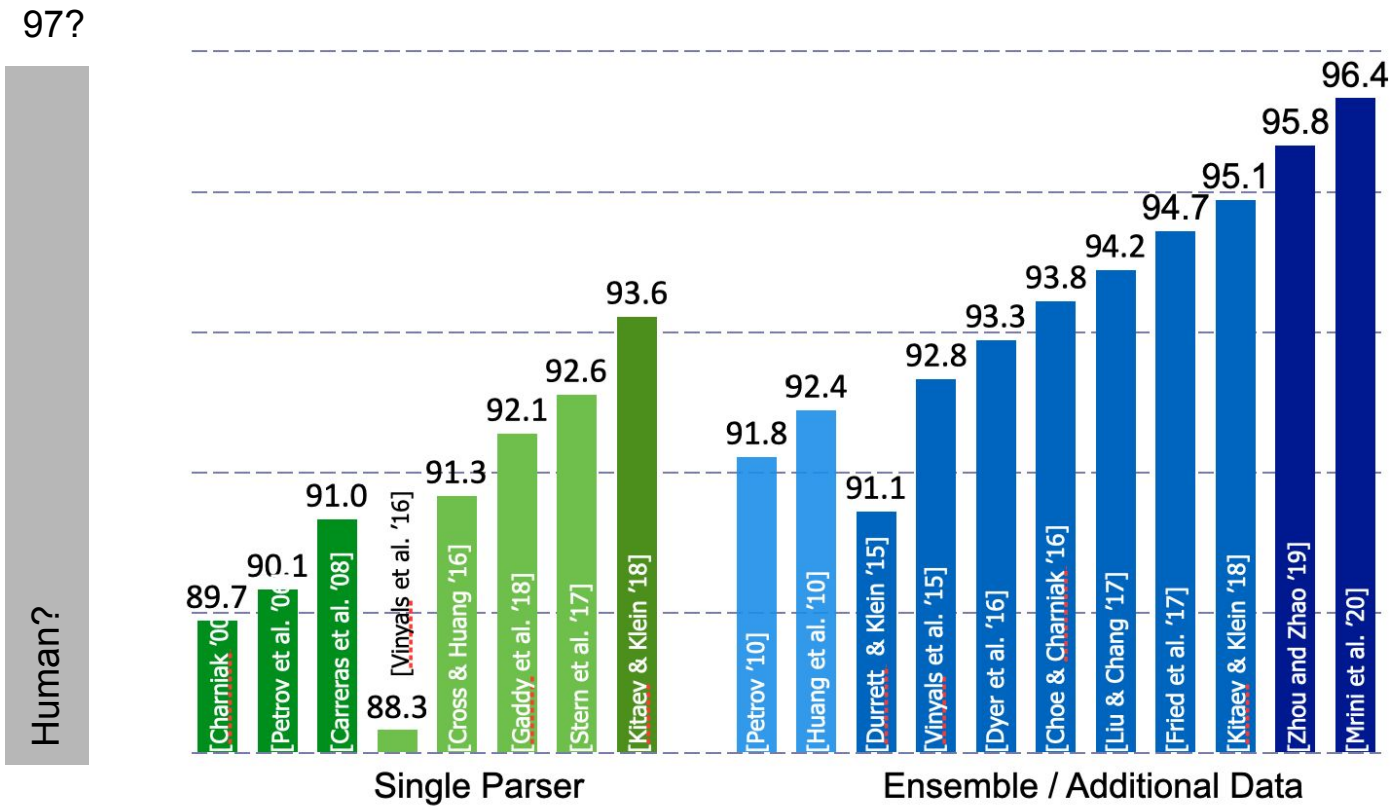
How are **painkillers** bad for you? - Apr 13, 2010
What effect do **painkillers** have on you? - Sep 30, 2009
What's the difference between OTC **painkillers**? (Advil, Aleeve ... - Nov 19, 2008
Is it bad to take **painkillers** on an empty **stomach** ? - Aug 20, 2008
More results from answers.yahoo.com »

Phys Ed: Does Ibuprofen Help or Hurt During Exercise? - NYTimes.com
Sep 1, 2009 ... When are ibuprofen and other anti-inflammatory **painkillers** justified? ... Can't take NSAIDS due to **upset stomach**. — bonemri ... Perhaps part of the problem is that you **don't** always have the option to rest your injury if ...
well.blogs.nytimes.com/.../phys-ed-does-ibuprofen-help-or-hurt-during-exercise/ - Cached - Similar

Today


Google
painkillers that don't upset the stomach
All
Images
Shopping
Videos
News
More
Settings
Tools
About 1.110.000 results (1,00 seconds)
Stomach-friendly painkillers
Gastro-resistant Naproxen aims to stop the tablet breaking down in the stomach, and is, therefore, less likely to cause irritation, stomach pain and complications like ulcers. Alternatively, Vimovo contains naproxen and esomeprazole, an added ingredient to protect your stomach.
www.theindependentpharmacy.co.uk › pain › guides
How To Protect Your Stomach Lining When Taking Painkillers ...
About Featured Snippets
Feedback
People also ask
Which painkiller is easiest on the stomach?
What is the best pain reliever if you have sensitive stomach?
How can I take ibuprofen without hurting my stomach?
Which Nsaid is least irritating to the stomach?
Feedback

English Constituency Parsing Results



SQuAD2.0

The Stanford Question Answering Dataset

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694v2	90.578	92.978
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799
5 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777
6 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694v2	90.115	92.580
7 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425

https://rajpurkar.github.io/SQuAD-explorer/

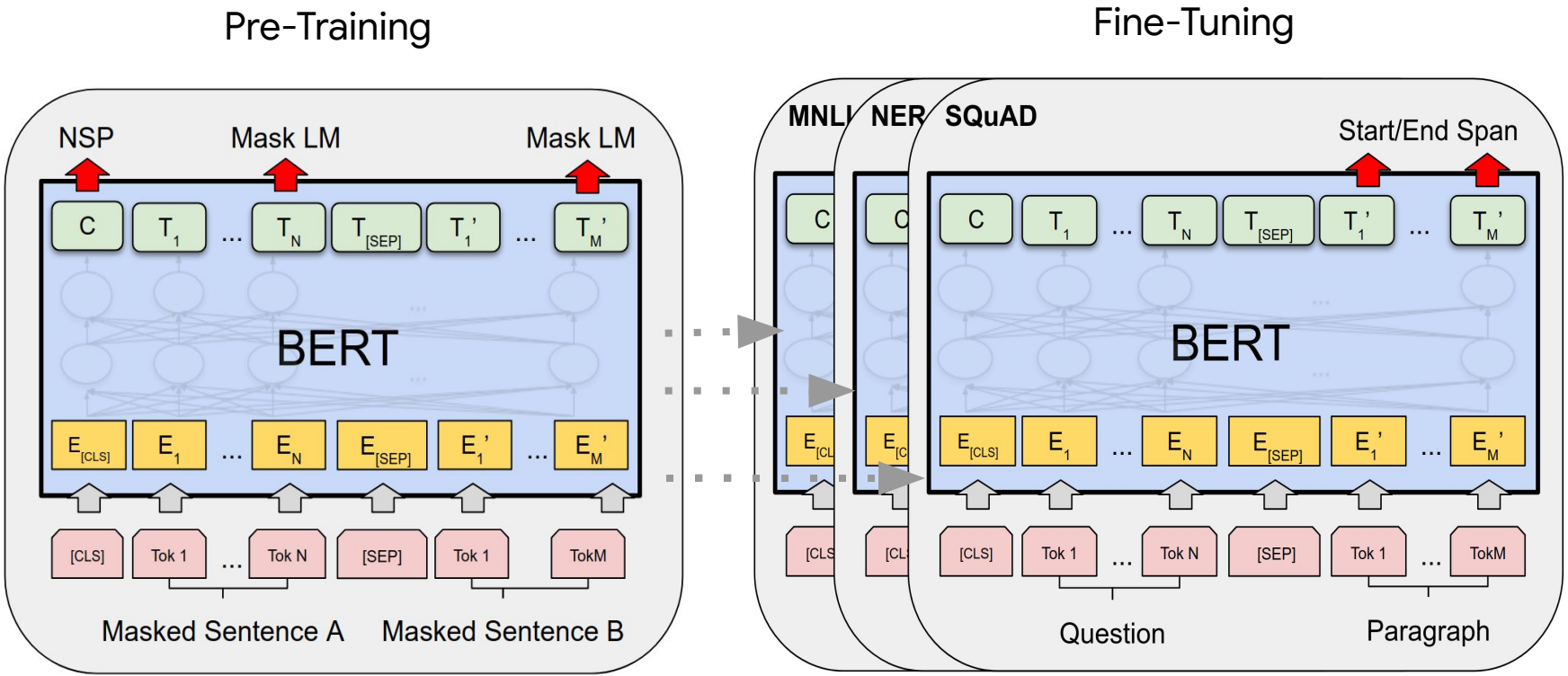
	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
+	1	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
	2	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	91.0	96.6	90.9	94.5	51.7
+	3	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.9	90.7	96.4	90.2	94.5	49.1
	4	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
	5	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
+	6	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5	51.6
+	7	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
+	8	Huawei Noah's Ark Lab	NEZHA-Large		89.1	69.9	97.2	93.2/91.0	92.2/91.6	74.2/90.6	91.0	90.7	95.7	88.5	93.2	45.0
+	9	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
	10	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
	11	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
+	12	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
	13	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

https://gluebenchmark.com/

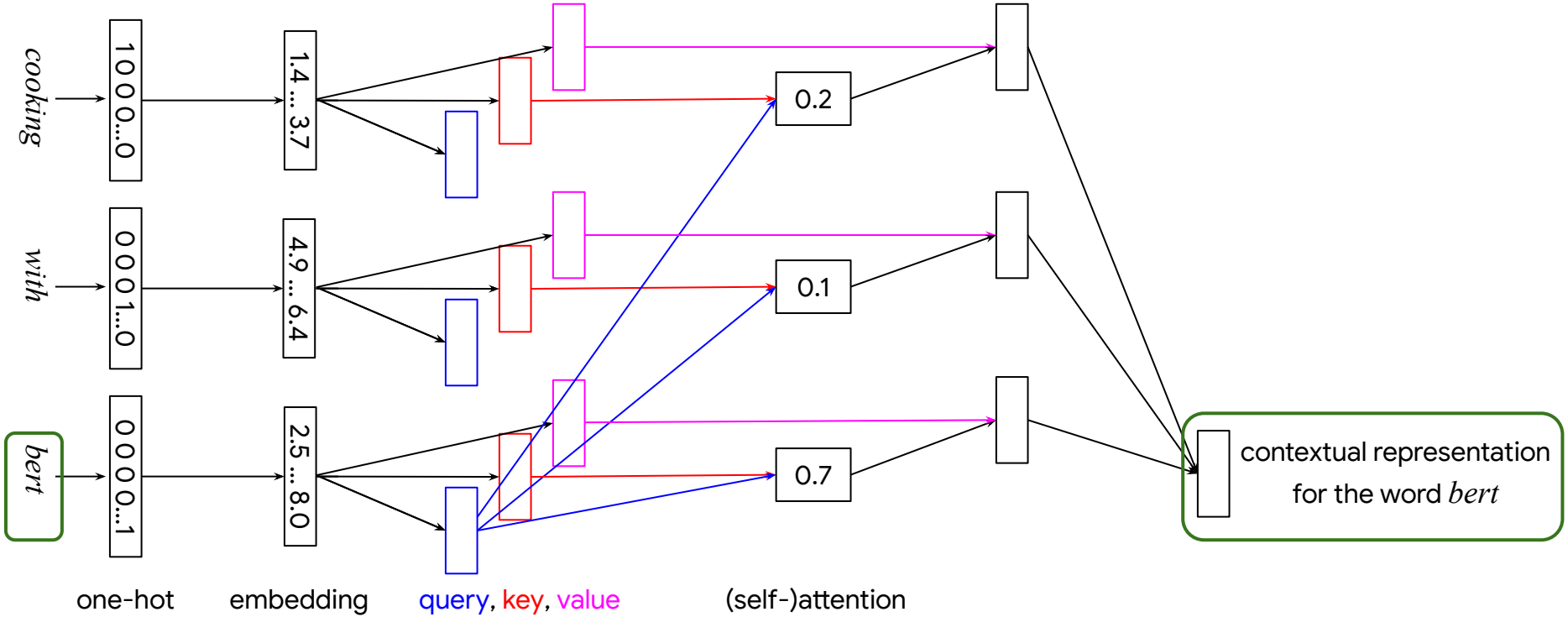
	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	2	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
	3	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
	4	Tencent Jarvis Lab	RoBERTa (ensemble)		85.4	86.4	92.5/95.6	90.8	83.9/52.0	90.6/90.2	87.9	74.1	91.8	57.6	89.3/75.6
+	5	Huawei Noah's Ark Lab	NEZHA-Large		84.8	86.8	94.4/96.0	91.2	82.9/48.8	87.4/86.7	88.5	73.1	90.4	58.0	87.1/74.4
	6	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	7	Peng Li	ALBERT XXLarge		84.4	87.4	90.0/93.6	91.8	84.6/54.7	85.3/83.7	87.2	73.6	89.0	75.6	98.3/99.2
+	8	Infosys : DAWN : AI Research	RoBERTa-iCETS		77.4	84.7	88.2/91.6	85.8	78.4/37.5	82.9/82.4	83.8	69.1	65.1	35.2	93.8/68.8
	9	IBM Research AI	BERT-mtl		73.5	84.8	89.6/94.0	73.8	73.2/30.5	74.6/74.0	84.1	66.2	61.0	29.6	97.8/57.3
	10	Ben Mann	GPT-3 few-shot - OpenAI		71.8	76.4	52.0/75.6	92.0	75.4/30.5	91.1/90.2	69.0	49.4	80.1	21.1	90.4/55.3
	11	SuperGLUE Baselines	BERT++		71.5	79.0	84.8/90.4	73.8	70.0/24.1	72.0/71.3	79.0	69.6	64.4	38.0	99.4/51.4
			BERT		69.0	77.4	75.7/83.6	70.6	70.0/24.1	72.0/71.3	71.7	69.6	64.4	23.0	97.8/51.7

https://super.gluebenchmark.com/

So, what happened? BERT happened.



From One-Hot Vectors to Word Embeddings & Self-Attention



Google Research

The Annotated Transformer, The Illustrated Transformer, The Illustrated BERT

BERT in practice

Sundar Pichai,
CEO of Google and
Alphabet

First, powered by our long-term investment in AI, we dramatically improved our understanding of the questions people ask Google Search. ***It's the biggest leap forward for Search in the past five years. It's all possible because of a new type of neural-network based technique for natural language processing called BERT****, which recognizes subtle patterns in language and provides more relevant results.*

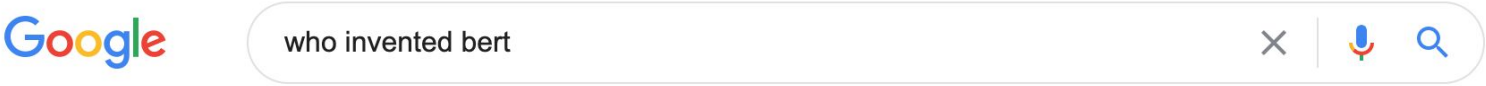
[Q3 2019 Earnings Call]

Pandu Nayak,
Google Fellow and
Vice President,
Search

In fact, when it comes to ranking results, BERT will help Search better understand one in 10 searches in the U.S. in English, and we'll bring this to more languages and locales over time.

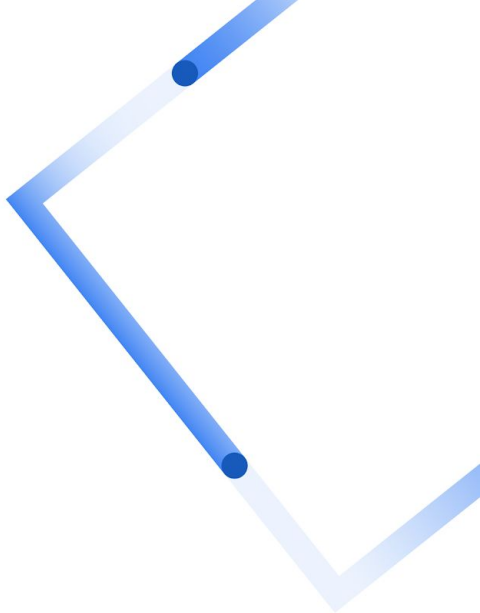
Google Research

[The Keyword, Oct 2019]

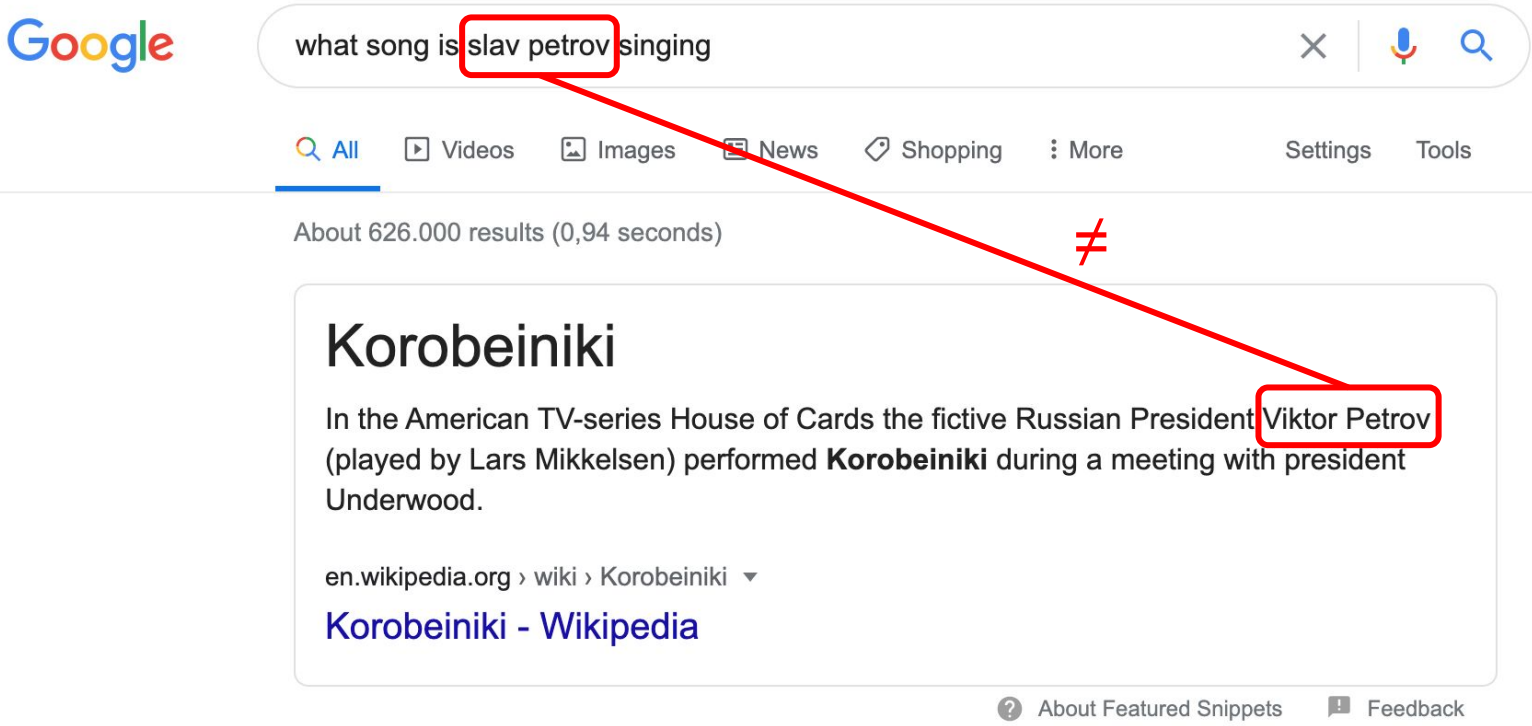


Google Research

So, are we done yet?



Google Research



Google Research

So, are we done yet?

No, not at all. Fortunately :)

Agenda

01 Intro

02 Datasets

03 Evaluation

04 Models for Reasoning

05 Conclusions

Google Research

Reading Comprehension vs. Information-Seeking QA

Passage

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".

Annotator writes question

Question

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

Answer

graupel

High lexical overlap.

Many datasets solved.

Question: What ship did Han Solo pilot?

Annotator finds answer in article

Article

The **Millennium Falcon** is a fictional starship in the Star Wars franchise. The modified YT-1300 Corellian light freighter is primarily commanded by Corellian smuggler Han Solo (Harrison Ford) and his Wookiee first mate, Chewbacca (Peter Mayhew). Designed by the Corellian Engineering Corporation (CEC), the highly modified YT-1300 is durable, modular, and is stated as being the second-fastest vessel in the Star Wars canon.

Subtle *relationship* between Q&A.

Many natural no-answer Q's.

Google Natural Questions

Long Answer Leaderboard

Rank	Model	Participant	Affiliation	Attempt Date	F1	Precision	Recall	R@P = 90	R@P = 75	R@P = 50
1	ETC-large	AnonymousOwl	Anonymous	5/24/20	0.7778	0.77476	0.78087	0.52204	0.79864	0.86598
2	ReflectionNet-ensemble	Wide_Field	Anonymous	2/9/20	0.77185	0.76791	0.77583	0.53345	0.78526	0.85238
3	RikiNet_v2	DREAM_Losin	anonymous	11/29/19	0.76094	0.78106	0.74183	0.4014	0.76991	0.85677
4	RikiNet-ensemble	DREAM_Losin	anonymous	11/14/19	0.75609	0.75305	0.75916	0.40535	0.76047	0.8526
5	MagNet	Wide_Field	Anonymous	11/19/19	0.75471	0.75796	0.75148	0.51086	0.75609	0.8186

Google Research

https://ai.google.com/research/NaturalQuestions

TyDi QA

Typologically **Di**verse **Q**uestion **A**nswering

A benchmark for information-seeking question answering in typologically diverse languages

Passage Answer Task

Rank	Model	Participant	Affiliation	Attempt Date	F1	Precision	Recall
1	Anonymous_submission	Anonymous_submission	Anonymous	6/4/2020	77.65	77.43	78.00
2	tydiqa-baseline	tydiqa-team	Google Research	2/15/2020	64.40	62.32	67.13

Google Research

https://ai.google.com/research/tydiqa

XTREME

(X) Cross-Lingual Transfer Evaluation of Multilingual Encoders

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	Anonymous1	Anonymous1	Anonymous1	Jun 17, 2020	70.7	84.0	53.4	68.9	77.3
2	XLM-R (large)	XTREME Team	Alphabet, CMU	-	68.2	82.8	69.0	62.3	61.6
3	mBERT	XTREME Team	Alphabet, CMU	-	59.6	73.7	66.3	53.8	47.7
4	MMTE	XTREME Team	Alphabet, CMU	-	59.3	74.3	65.3	52.3	48.9

Google Research

https://sites.research.google/xtreme

ToTTo: A Controlled Table-To-Text Generation Dataset

Table Title: Cristhian Stuani
Section Title: International goals
Table Description: As of 25 March 2019 (Uruguay score listed first, score column indicates score after each Stuani goal)

No.	Date	Venue	Opponent	Score	Result	Competition
1.	10 September 2013	Estadio Centenario, Montevideo, Uruguay	Colombia	2-0	2-0	2014 FIFA World Cup qualification
2.	13 November 2013	Amman International Stadium, Amman, Jordan	Jordan	2-0	5-0	2014 FIFA World Cup qualification
3.	31 May 2014	Estadio Centenario, Montevideo, Uruguay	Northern Ireland	1-0	1-0	Friendly
4.	5 June 2014		Slovenia	2-0	2-0	

Original Text: On 13 November 2013, he netted the Charruas' second in their 5 – 0 win in Jordan for the playoffs first leg, finishing Nicolas Lodeiro's cross at close range.
Text after Deletion: On 13 November 2013, he netted the second in their 5 – 0 win in Jordan.
Text after Decontextualization: On 13 November 2013, Cristhian Stuani netted the second in 5 – 0 win in Jordan.
Final Text: On 13 November 2013 Cristhian Stuani netted the second in a 5 – 0 win in Jordan.

Google Research

https://github.com/google-research-datasets/ToTTo

Google Research Datasets

Datasets released by Google Research

Mountain View, CA http://research.google.com

Repositories 39 Packages People 7 Teams Projects Insights Settings

Pinned repositories

Customize pinned repositories

natural-questions

Natural Questions (NQ) contains real user questions issued to Google search, and answers found from Wikipedia by annotators. NQ is designed for the training and evaluation of automatic question ans...

Python 522 101

conceptual-captions

Conceptual Captions is a dataset containing (image-URL, caption) pairs designed for the training and evaluation of machine learned image captioning systems.

Shell 213 13

ToTTo

ToTTo is an open-domain English table-to-text dataset with over 120,000 training examples that proposes a controlled generation task: given a Wikipedia table and a set of highlighted table cells, p...

95 4

dakshina

The Dakshina dataset is a collection of text in both Latin and native scripts for 12 South Asian languages. For each language, the dataset includes a large collection of native script Wikipedia tex...

93 8

tydiqa

TyDi QA contains 200k human-annotated question-answer pairs in 11 Typologically Diverse languages, written without seeing the answer and without the use of translation, and is designed for the trai...

Python 91 18

gap-coreference

GAP is a gender-balanced dataset containing 8,908 coreference-labeled pairs of (ambiguous pronoun, antecedent name), sampled from Wikipedia for the evaluation of coreference resolution in practica...

Python 169 68

https://github.com/google-research-datasets

Agenda

01 Intro

02 Datasets

03 Evaluation

04 Models for Reasoning

05 Conclusions

Google Research

Ground Truth is a Myth

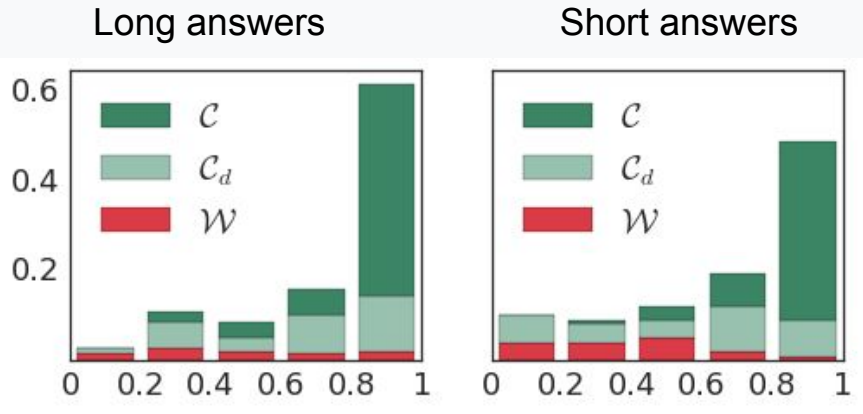
Question: *who was the first person to see earth from space*

Yuri Alekseyevich Gagarin[a] (9 March 1934 – 27 March 1968) was a Soviet Air Forces pilot and cosmonaut who became the first human to journey into outer space, achieving a major milestone in the Space Race; his capsule, Vostok 1, completed one orbit of Earth on 12 April 1961. Gagarin became an international celebrity and was awarded many medals and titles, including Hero of the Soviet Union, his nation's highest honour.

Google Research

Kwiatkowski et al., “Natural Questions: A Benchmark for Question Answering Research,” TACL ‘19

25-way Annotation of Answers



X-axis - proportion of annotations that are non-null for question.
Y-axis - expectation that a non-null annotation's question is in this bucket.

Also broken down, conditioned on annotation being: Correct ($\mathcal{C}$); Correct but debatable ($\mathcal{C}_d$); or Wrong ($\mathcal{W}$).

Google Research

Kwiatkowski et al., “Natural Questions: A Benchmark for Question Answering Research,” TACL ‘19

The “Entailment” Task - A Brief History

“The basic aim of semantics is to characterize the notion of a true sentence (under a given interpretation) and of entailment”

Montague (1970)

p entails *h* if, in every possible world in which *p* is true, *h* is also true.

Formal Logic/Semantics

p entails *h* if *“typically, a human reading p would infer that h is most likely true assuming common human understanding of language as well as common background knowledge.”*

“... the task definition and evaluation methodologies are clearly not mature yet. We expect them to change over time....”

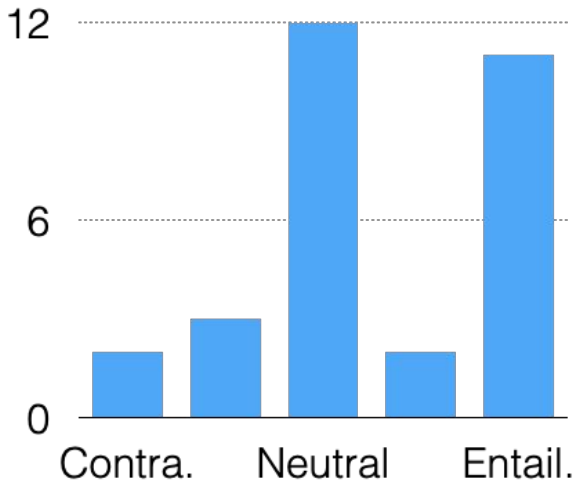
Dagan (2006)

Pavlick & Kwiatkowski, “Inherent Disagreements in Human Textual Inferences,” TACL ’19

50-way Annotation of Entailment

Premise *A young woman stands by a barbecue.*

Hypothesis *The young female is near a machine.*



Pavlick & Kwiatkowski, “Inherent Disagreements in Human Textual Inferences,” TACL ’19

Take-aways on human annotations

- As a field, we've opted **not** to be **prescriptivist**. This reduces work on developing guidelines, but requires **increased work in defining eval metrics**.
- How do we ask models to "do what humans do" if **humans don't all do the same thing**?
- Short-term way forward? Explicitly **reward models for predicting the full distribution** of human judgments.
- Long-term way forward? Models that can rationalize their judgements, and/or ask for clarification when relevant.

Google Research

Pavlick & Kwiatkowski, "Inherent Disagreements in Human Textual Inferences," TACL '19

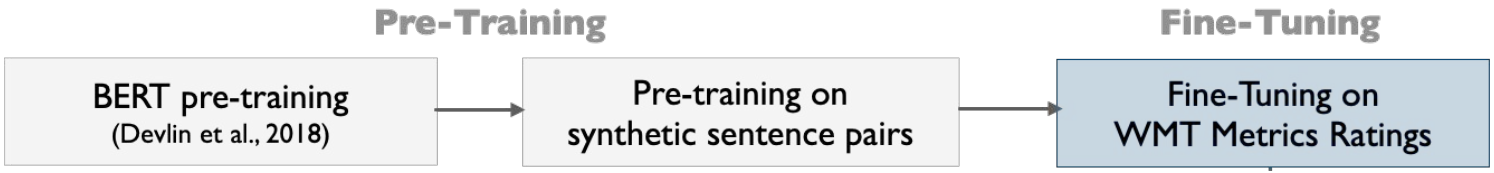
What about Evaluation for Text Generation?

- **Human Evals**: considered reliable (if done properly), but slow and expensive
- **Automatic Metrics**: cheap and fast, but not very accurate (e.g. BLEU, ROUGE)
- **Hybrid metrics**: could provide the best of both worlds but are often not so flexible (BERTscore (Zhang et al., 2019), YiSi (Chi-kiu Lo, 2019), Sentence Mover's Similarity (Clark et al., 2019)), or nor so robust (BEER (Stanojević et al., 2014), RUSE (Shimanaka et al., 2018), ESIM (Mathur et al., 2019))
- **Our Proposal: BLEURT**

Google Research

Sellam et al., "BLEURT: Learning Robust Metrics for Text Generation," ACL '20

BLEURT - Pre-Training for Robustness



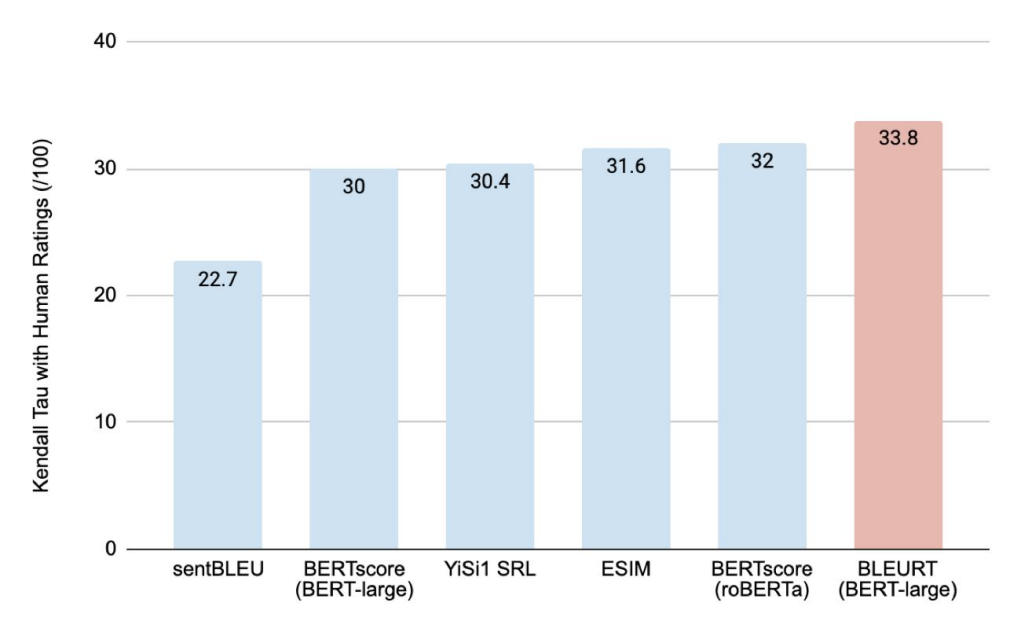
Task Type	Pre-training Signals	Loss Type
BLEU	τ_{BLEU}	Regression
ROUGE	$\tau_{\text{ROUGE}} = (\tau_{\text{ROUGE-P}}, \tau_{\text{ROUGE-R}}, \tau_{\text{ROUGE-F}})$	Regression
BERTscore	$\tau_{\text{BERTscore}} = (\tau_{\text{BERTscore-P}}, \tau_{\text{BERTscore-R}}, \tau_{\text{BERTscore-F}})$	Regression
Backtrans. likelihood	$\tau_{\text{en-fr},z\|\tilde{z}}, \tau_{\text{en-fr},\tilde{z}\|z}, \tau_{\text{en-de},z\|\tilde{z}}, \tau_{\text{en-de},\tilde{z}\|z}$	Regression
Entailment	$\tau_{\text{entail}} = (\tau_{\text{Entail}}, \tau_{\text{Contradict}}, \tau_{\text{Neutral}})$	Multiclass
Backtrans. flag	$\tau_{\text{backtran\_flag}}$	Multiclass

Google Research

Sellam et al., “BLEURT: Learning Robust Metrics for Text Generation,” ACL ’20

How accurate is BLEURT?

Agreement with Human Ratings on WMT Metrics Shared Task ’19



Google Research

Sellam et al., “BLEURT: Learning Robust Metrics for Text Generation,” ACL ’20

Agenda

01 Intro

02 Datasets

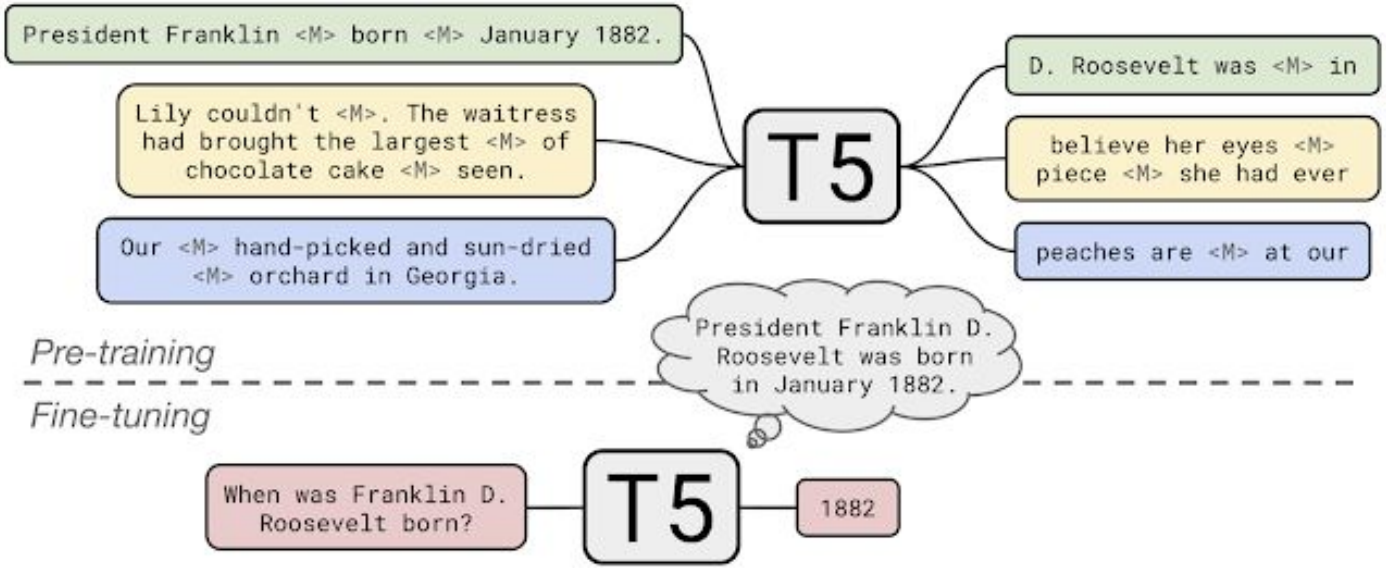
03 Evaluation

04 Models for Reasoning

05 Conclusions

Google Research

Are giant language models enough?

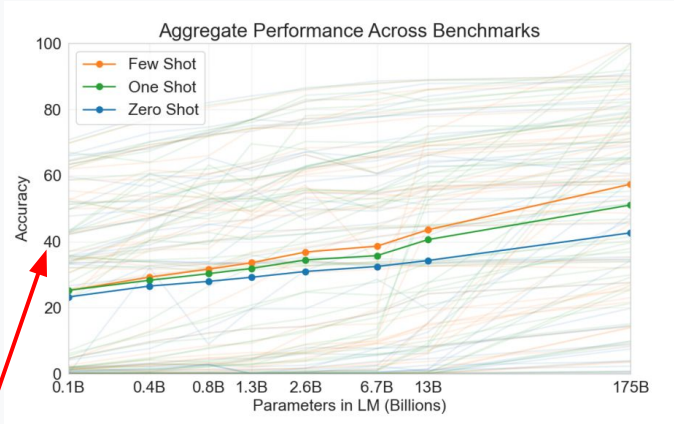


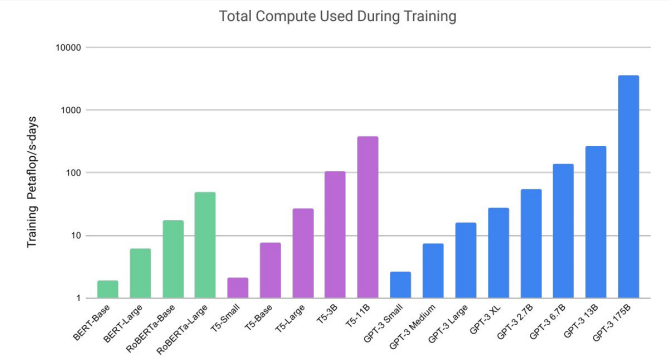
Google Research

Raffael et. al, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Arxiv '19

My biased view

- **Multi-Task** and **Few-Shot Learning** are the holy grail
- But I don't believe that it is sufficient to simply make language modes bigger
- That's a lot of compute for pretty poor



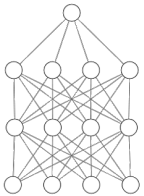


Google Research

Brown et. al, "Language Models are Few-Shot Learners" Arxiv '20

What are pre-trained models learning?

__ = ‘had’



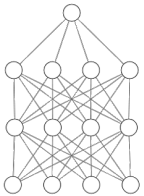
He said he __ eaten the asparagus

$$\sum_z p(y \mid x, z)\, p(z \mid x) = p(y \mid x)$$

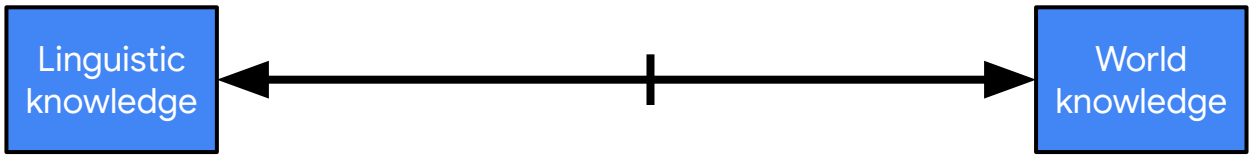
Knowledge-Augmented Encoder

Neural Retriever

__ = ‘Euro’



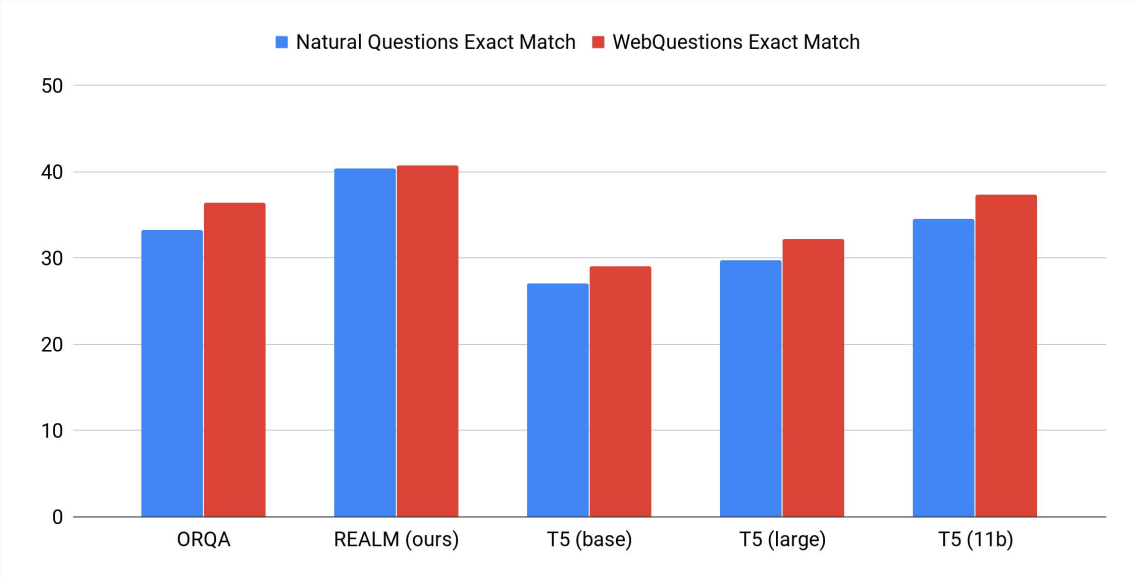
We paid 20 __ at the Louvre gift shop



Google Research

Guu et al., “REALM: Retrieval-Augmented Language Model Pre-training” ICML ’20

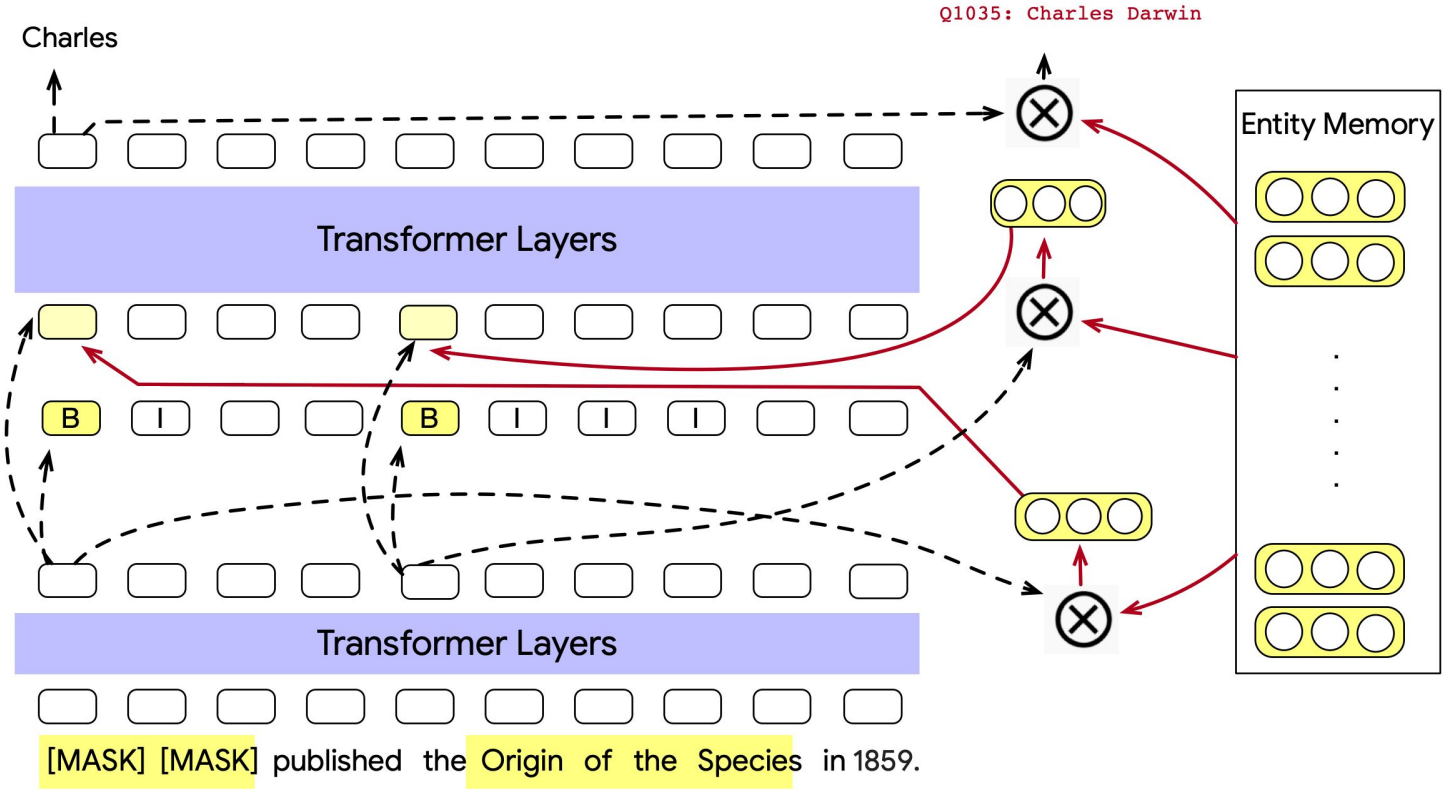
Open Domain Question Answering



Google Research

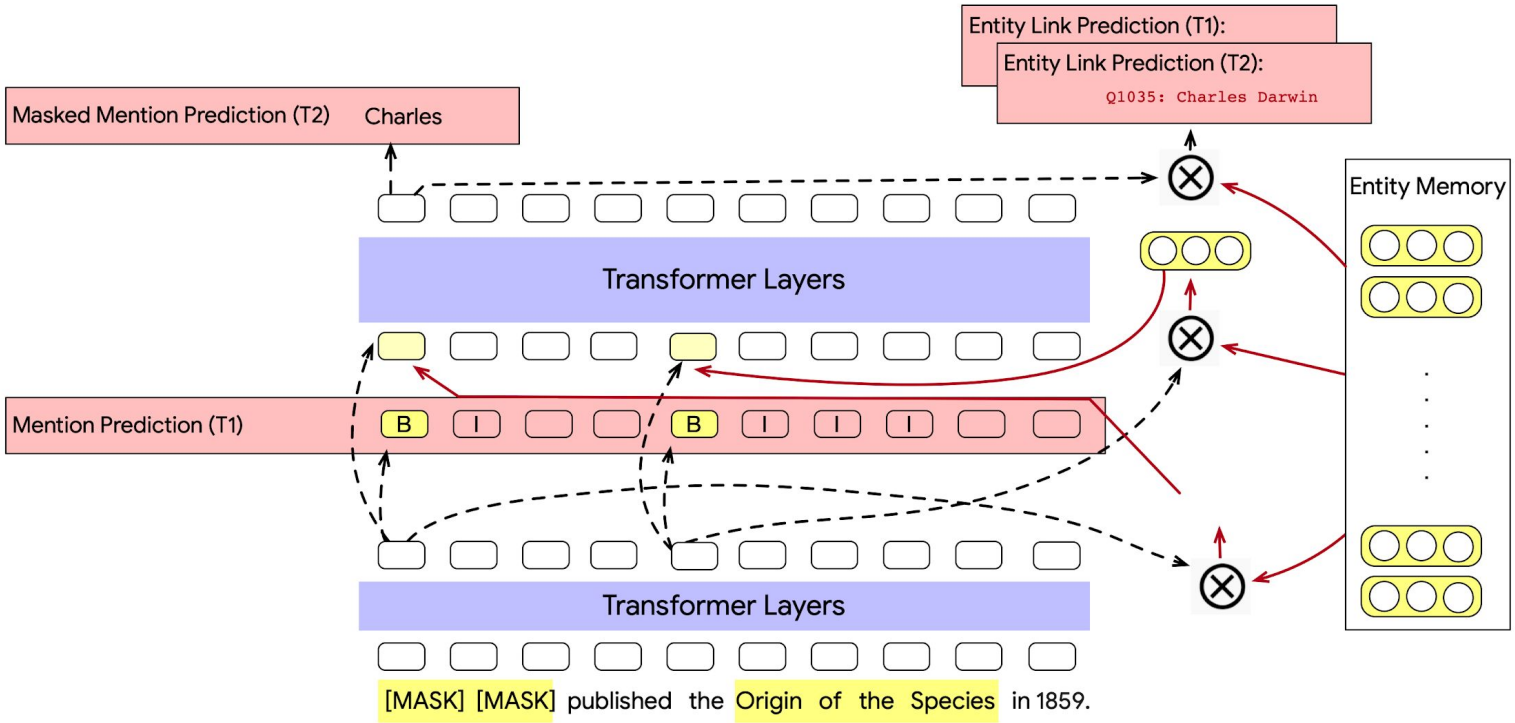
Guu et al., “REALM: Retrieval-Augmented Language Model Pre-training” ICML ’20

Entities as Experts model architecture



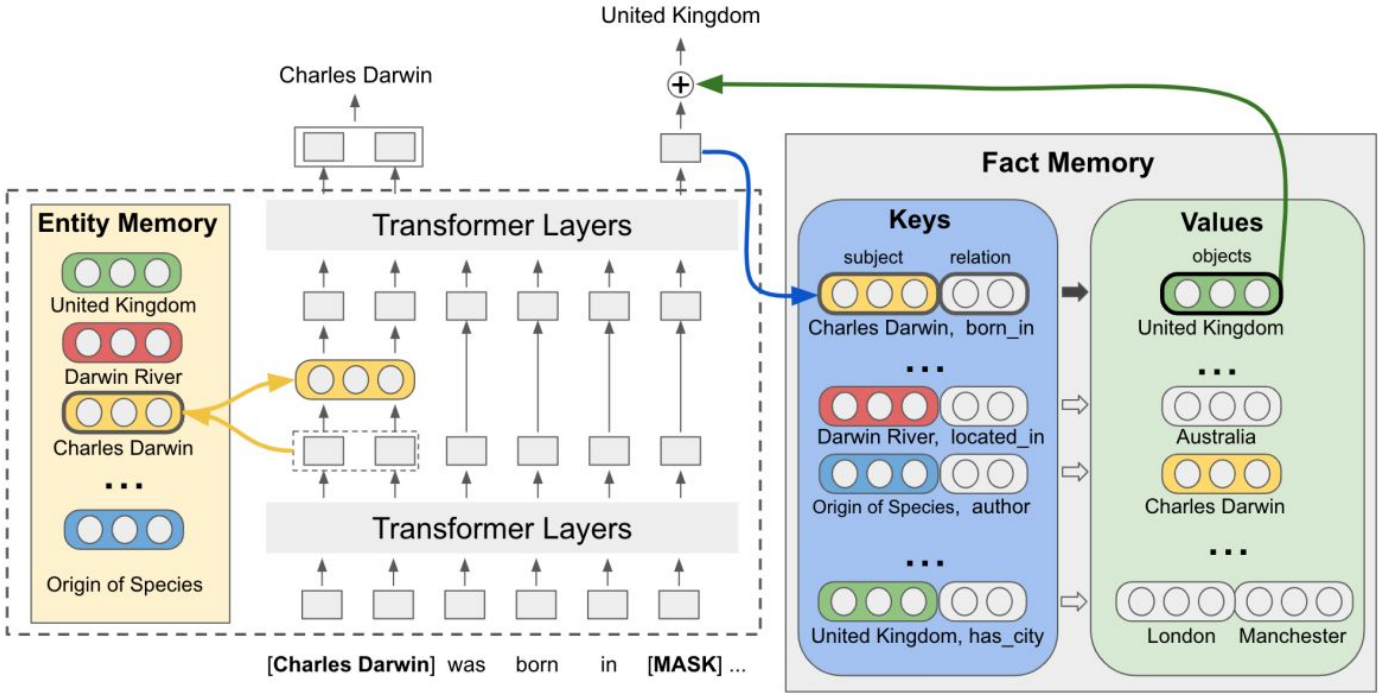
Févry et al., "Entities as Experts: Sparse Memory Access with Entity Supervision," Arxiv '20

Entities as Experts model architecture



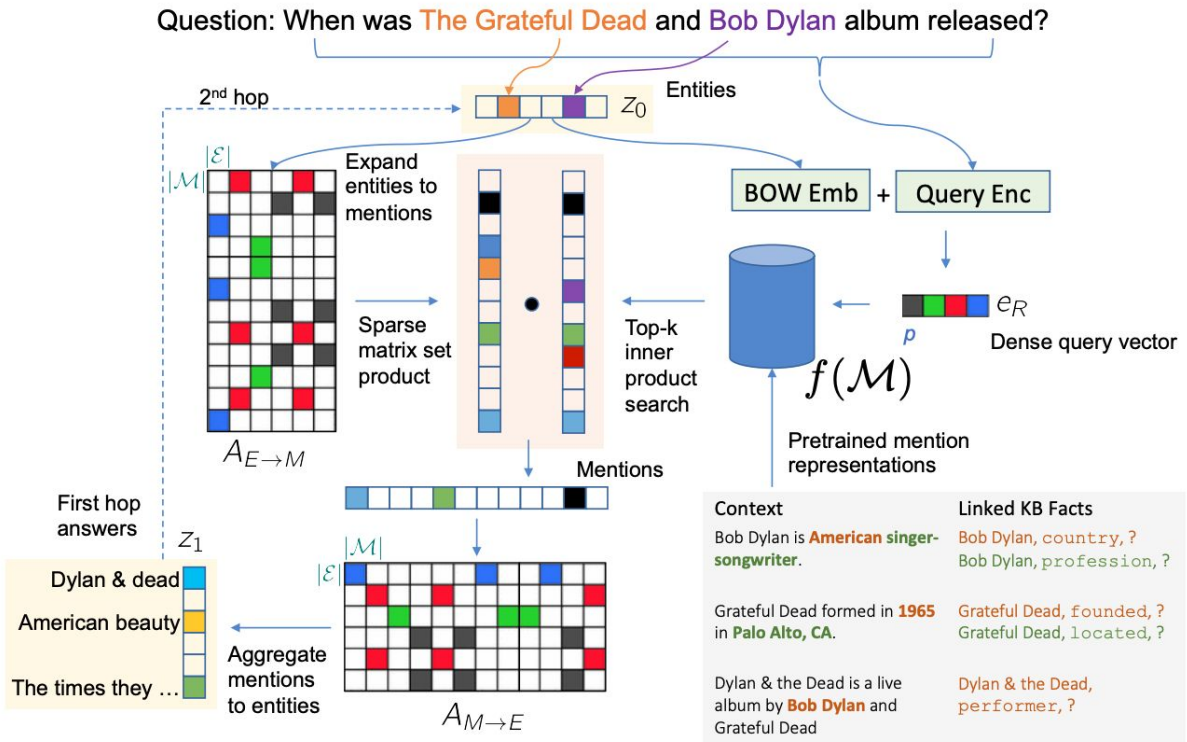
Févry et al., "Entities as Experts: Sparse Memory Access with Entity Supervision," Arxiv '20

From Entities to Facts as Experts



Verga et al., "Facts as Experts: Adaptable and Interpretable Neural Memory over Symbolic Knowledge," Arxiv '20

DrKIT: multi-hop question answering



Dhingra et al., "Differentiable Reasoning over a Virtual Knowledge Base," ICLR '20

What about simple math?

Context:

Still trying to get their first win of the year, the Dolphins went home for a Week 13 AFC East rematch with the throwback-clad New York Jets. In the first quarter, Miami trailed early as Jets RB Leon Washington got an 18-yard TD run. The Dolphins ended the period with kicker Jay Feely getting a **53-yard field goal.** In the second quarter, Miami drew closer as Feely kicked a **44-yard field goal**, yet New York replied with kicker Mike Nugent getting a 29-yard field goal. Afterwards, the Dolphins took the lead as CB Michael Lehan returned a fumble 43 yards for a touchdown. However, the Jets regained the lead with QB Kellen Clemens completing a 19-yard TD pass to WR Brad Smith, along with Nugent kicking a 40-yarder and a 35-yard field goal. In the third quarter, New York increased their lead with Nugent kicking a 35-yard field goal for the only score of the period. In the fourth quarter, the Jets sealed the win with RB Thomas Jones getting a 1-yard TD run, Nugent

Question:

How many yards do the first two field goals converted add up to?

Answer:

(53 + 44) ⇒ 97

Google Research

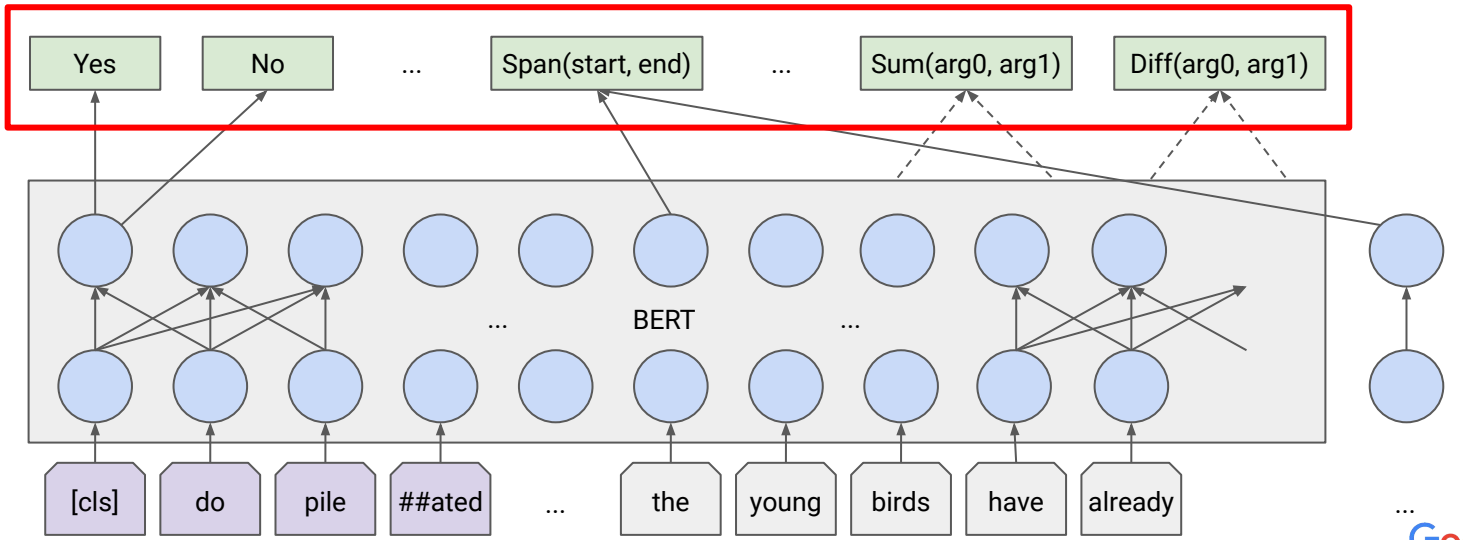
Andor et al., "Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension," ACL '20

BERT-Based Structured Prediction

We **merge**

- **classification**
- **span selection**

Jointly predict *operations* and their *arguments*



Google Research

Andor et al., "Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension," ACL '20

Ops With Examples

	Derivations	Example Question	Answer Derivation
Literals	YES, NO, UNKNOWN, 0, 1 ..., 9	How many field goals did Stover kick?	4
Numerical	`Diff100` : $n_0 \rightarrow 100 - n_1$	How many percent of the national population does not live in Bangkok?	$100 - 12.6 = 87.4$
	`Sum` : $n_0, n_1 \rightarrow n_0 + n_1$ as well as: `Diff`, `Mul`, `Div`	How many from the census were in Ungheni and Cahul?	$32,828 + 28,763 = 61591$
Text spans	`Span` : $i, j \rightarrow s$	Does Bangkok have more Japanese or Chinese nationals?	"Japanese"
Compositions	`Merge` : $s_0, s_1 \rightarrow \{s_0, s_1\}$	What languages are spoken by more than 1%, but fewer than 2% of Richmond's residents?	"Hmong-Mien languages", "Laotian"
	`Sum3` : $n_0, n_1, n_2 \rightarrow (n_0 + n_1) + n_2$	How many residents, in terms of percentage, speak either English, Spanish, or Tagalog?	$\texttt{Sum}(64.56, 23.13) + 2.11 = 89.8$

Google Research

Andor et al., "Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension," ACL '20

Conclusions

- We have made a lot of progress in the last years, but we have also cut many corners:
  - Many of our datasets are artificially easy
  - Our evaluation metrics are simplistic

- To make progress we need to:
  - Embrace the ambiguity inherent in human judgments
  - Design models that are more efficient learners
  - Invest more in trustworthiness

Google Research

Thank You

Slav Petrov
on behalf of the
Language Team @ Google Research

slav@google.com

Google Research