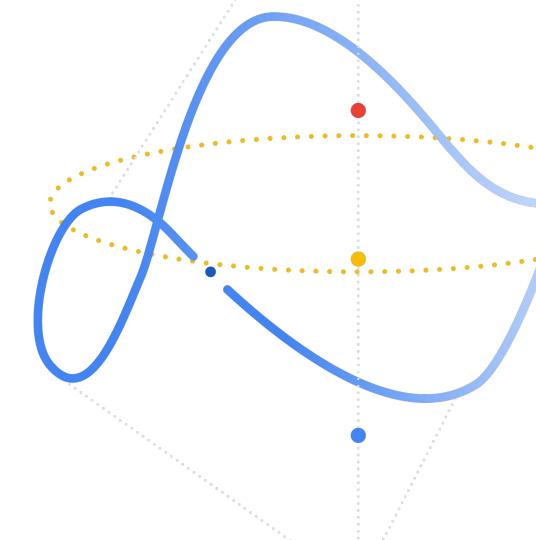




NLP & IR How Deep Learning has Bridged the Gap

Ryan McDonald

Material from a number of Google research projects

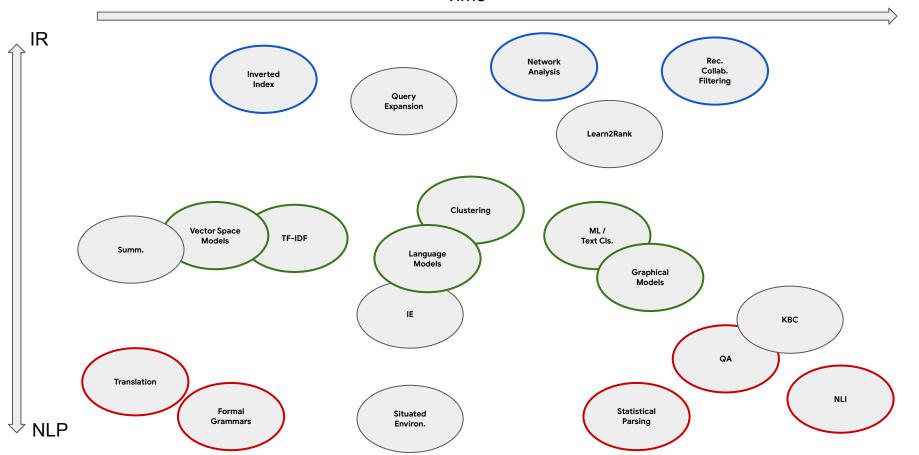








Time

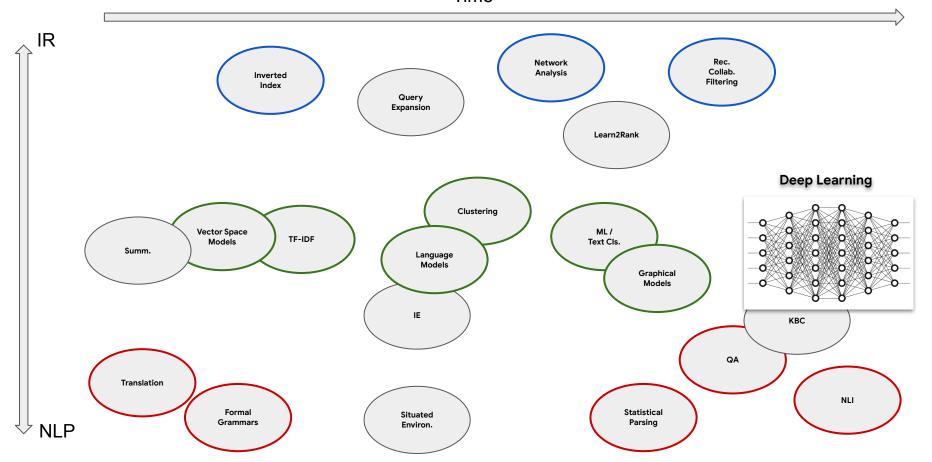








Time





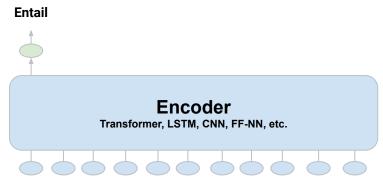
f(text, text) f(text, entity) f(text, image) f(text, object)



f(text, text)

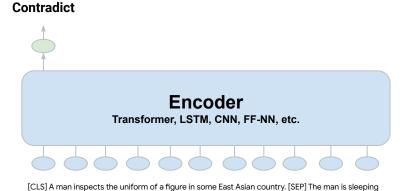






[CLS] A boy and his mother and father are at the beach [SEP] A family is doing something outside

Entail Encoder Encoder Transformer, LSTM, Transformer, LSTM, CNN, FF-NN, etc. CNN, FF-NN, etc. A boy and his mother and father are at the beach A family is doing something outside



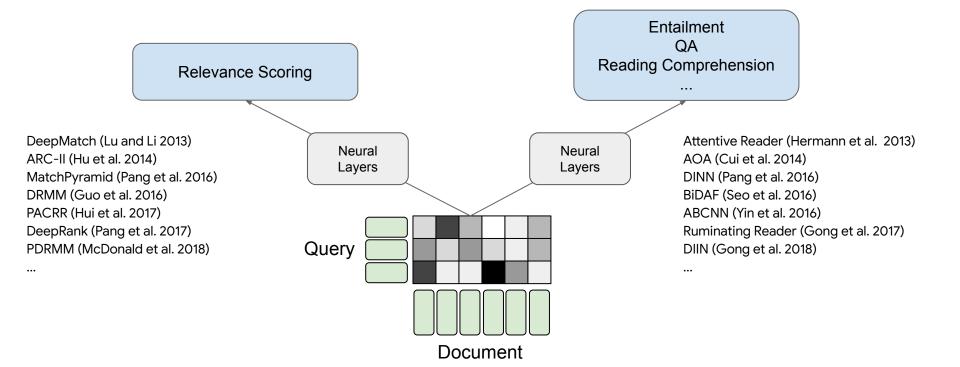
Encoder **Encoder** Transformer, LSTM, Transformer, LSTM, CNN, FF-NN, etc. CNN, FF-NN, etc. A man inspects the uniform of a figure in some East Asian country. The man is sleeping

Contradict

Cross-attention (AKA one-tower; AKA interaction-based)





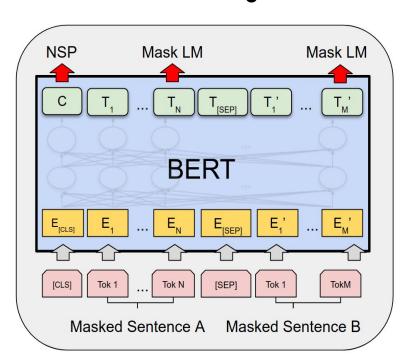


BERT: Transformers + Pre-training + Fine-Tuning





Pre-Training



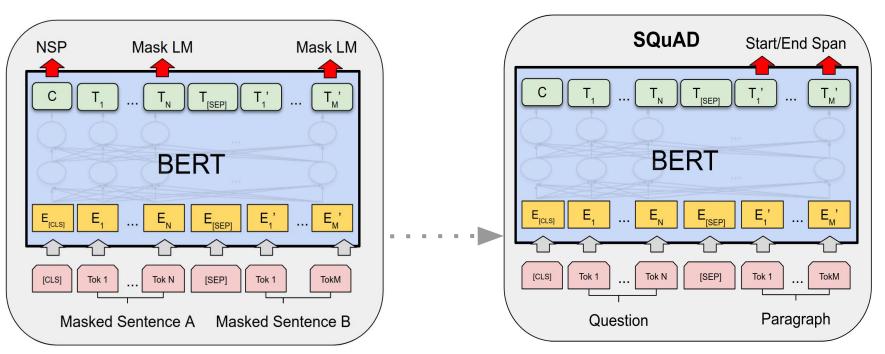
BERT: Transformers + Pre-training + Fine-Tuning







Fine-Tuning



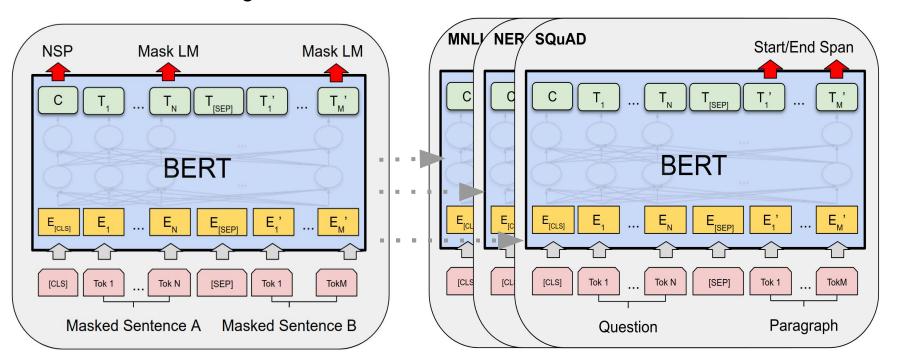
BERT: Transformers + Pre-training + Fine-Tuning





Pre-Training

Fine-Tuning



Transformers + Pre-training -- new dawn of NLP





System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	_
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERTBASE	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

Devlin et al. 2019 Yang et al. 2019 Lan et al. 2019

Transformers + Pre-training -- new dawn of NLP





System	MNLI-(m/mm)	QQP	QNL
	392k	363k	1081
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8
OpenAI GPT	82.1/81.4	70.3	87.4
BERT _{BASE}	84.6/83.4	71.2	90.5
BERT _{LARGE}	86.7/85.9	72.1	92.7

Dataset	XLNet-Large	XLNet-Large	BERT-Large
	(as in paper)	-wikibooks	-wikibooks
			best of 3 variants
SQuAD1.1 EM	89.0	88.2	86.7 (II)
SQuAD1.1 F1	94.5	94.0	92.8 (II)
SQuAD2.0 EM	86.1	85.1	82.8 (II)
SQuAD2.0 F1	88.8	87.8	85.5 (II)
RACE	81.8	77.4	75.1 (II)
MNLI	89.8	88.4	87.3 (II)
QNLI	93.9	93.9	93.0 (II)
QQP	91.8	91.8	91.4 (II)
RTE	83.8	81.2	74.0 (III)
SST-2	95.6	94.4	94.0 (II)
MRPC	89.2	90.0	88.7 (III)
CoLA	63.6	65.2	63.7 (II)
STS-B	91.8	91.1	90.2 (III)

Devlin et al. 2019 Yang et al. 2019 Lan et al. 2019

Transformers + Pre-training -- new dawn of NLP





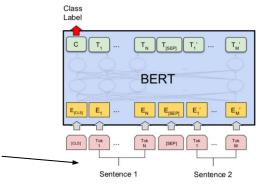
				Data	set		et-Large paper)	XLNet-l		BERT- -wikil		
System	MNLI-(m/mm)	QQP	QNL -			107.0				best of 3	variants	_ :rag
by stem	392k	363k	1081	SQuAD1	.1 EM	8	9.0	88.	2	86.7	(II)	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	SQuAD	1.1 F1	9	4.5	94.0	0	92.8	(II)	4.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	SQuAD2	2.0 EM	8	6.1	85.	1	82.8	(II)	1.0
OpenAI GPT	82 1/81 4	70.3	87 4	SQuAD:	2.0 F1	8	8.8	87.	8	85.5	(II)	5.1
BERT _{BASE}	Models	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg	9.6
$BERT_{LARGE}$	Single-task sing BERT-large	le models o 86.6	92.3	91.3	70.4	93.2	88.0	60.6	90.0	_	_	2.1
	XLNet-large RoBERTa-large ALBERT (1M)	89.8 90.2 90.4	93.9 94.7 95.2	91.8 92.2 92.0	83.8 86.6 88.1	95.6 96.4 96.8	89.2 90.9 90.2	63.6 68.0 68.7	91.8 92.4 92.7	-	-	
Devlin et al. 2019	ALBERT (1.5M	90.8	95.3	92.2	89.2	96.9	90.9	71.4	93.0	-		
Yang et al. 2019	Ensembles on te ALICE	88.2	aerboara 95.7	90.7	83.5	95.2	92.6	69.2	91.1	80.8	87.0	
Lan et al. 2019	MT-DNN	87.9	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6	
	XLNet RoBERTa	90.2 90.8	98.6 98.9	90.3 90.2	86.3 88.2	96.8 96.7	93.0 92.3	67.8 67.8	91.6 92.2	90.4 89.0	88.4 88.5	
	Adv-RoBERTa ALBERT	91.1 91.3	98.8 99.2	90.3 90.5	88.7 89.2	96.8 97.1	93.1 93.4	68.0 69.1	92.4 92.5	89.0 91.8	88.8 89.4	-

BERT: Fine-tuning Paradigms

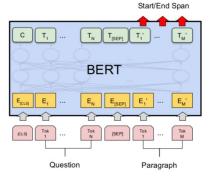




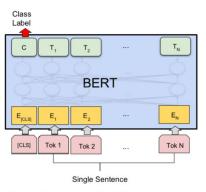
This can be a relevance scoring model



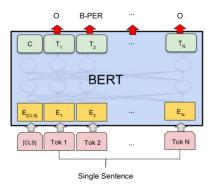
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(c) Question Answering Tasks: SQuAD v1.1



(b) Single Sentence Classification Tasks: SST-2, CoLA



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

BERT 4 Document Relevance Scoring

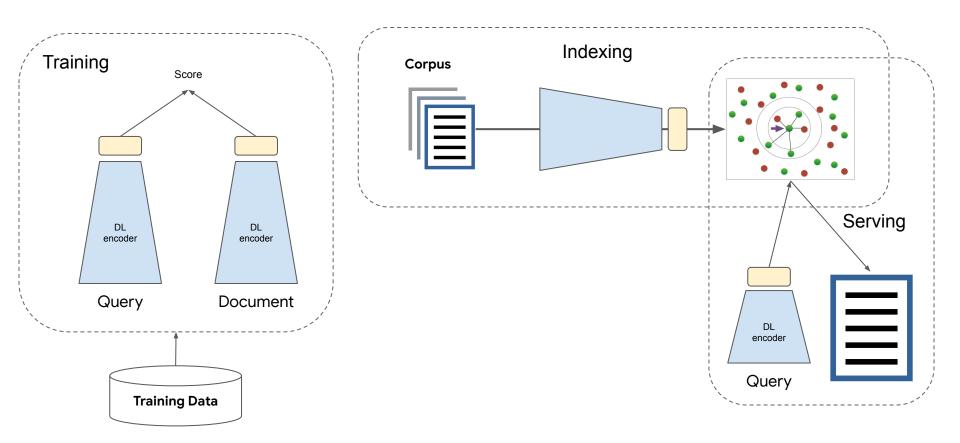




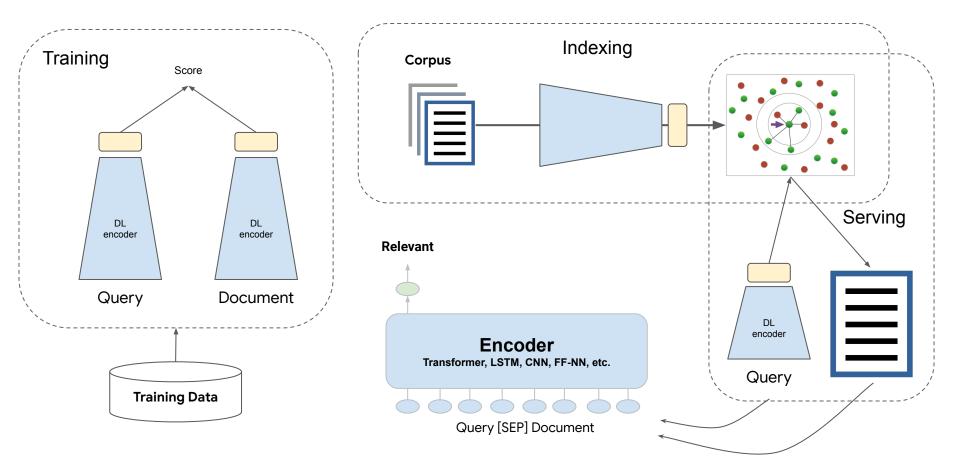
CEDR: Macaveney et al. 2019

		Robi	ust04	WebTrac	k 2012–14
Ranker	Input Representation	P@20	nDCG@20	nDCG@20	ERR@20
BM25	n/a	0.3123	0.4140	0.1970	0.1472
SDM [13]	n/a	0.3749	0.4353	-	-
TREC-Best	n/a	0.4386	0.5030	0.2855	0.2530
ConvKNRM	GloVe	0.3349	0.3806	[B] 0.2547	[B] 0.1833
Vanilla BERT	BERT (fine-tuned)	[BC] 0.4042	[BC] 0.4541	[BC] 0.2895	[BC] 0.2218
PACRR	GloVe	0.3535	[C] 0.4043	0.2101	0.1608
PACRR	ELMo	[C] 0.3554	[C] 0.4101	[BG] 0.2324	[BG] 0.1885
PACRR	BERT	[C] 0.3650	[C] 0.4200	0.2225	0.1817
PACRR	BERT (fine-tuned)	[BCVG] 0.4492	[BCVG] 0.5135	[BCG] 0.3080	[BCG] 0.2334
CEDR-PACRR	BERT (fine-tuned)	[BCVG] 0.4559	[BCVG] 0.5150	[BCVGN] 0.3373	[BCVGN] 0.2656
KNRM	GloVe	0.3408	0.3871	[B] 0.2448	0.1755
KNRM	ELMo	[C] 0.3517	[CG] 0.4089	0.2227	0.1689
KNRM	BERT	[BCG] 0.3817	[CG] 0.4318	[B] 0.2525	[B] 0.1944
KNRM	BERT (fine-tuned)	[BCG] 0.4221	[BCVG] 0.4858	[BCVG] 0.3287	[BCVG] 0.2557
CEDR-KNRM	BERT (fine-tuned)	[BCVGN] 0.4667	[BCVGN] 0.5381	[BCVG] 0.3469	[BCVG] 0.2772
DRMM	GloVe	0.2892	0.3040	0.2215	0.1603
DRMM	ELMo	0.2867	0.3137	[B] 0.2271	0.1762
DRMM	BERT	0.2878	0.3194	[BG] 0.2459	[BG] 0.1977
DRMM	BERT (fine-tuned)	[CG] 0.3641	[CG] 0.4135	[BG] 0.2598	[B] 0.1856
CEDR-DRMM	BERT (fine-tuned)	[BCVGN] 0.4587	[BCVGN] 0.5259	[BCVGN] 0.3497	[BCVGN] 0.2621

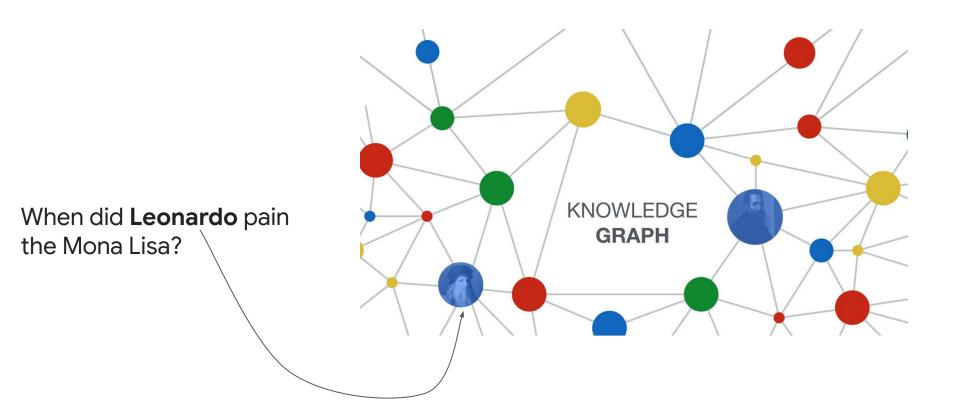
Dual Encoder Retrieval (AKA two-tower; AKA relevance-based)



Dual Encoder Retrieval (AKA two-tower; AKA relevance-based)



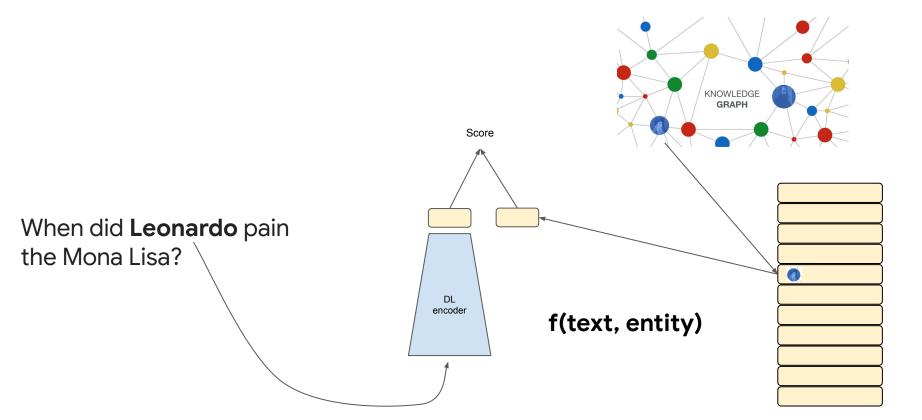
Classification with massive output spaces



Embedding Objects vs. Descriptions

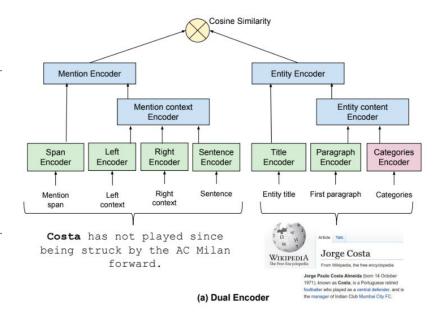


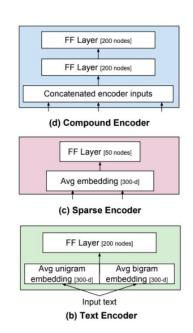




Embedding Objects vs. Descriptions

		R@1	
Dataset	AT-Prior	DE-RN	DE-MN
CoNLL	65.71	40.87	77.93
ACE2004	80.93	55.64	87.55
AQUAINT	82.64	54.55	86.78
MSNBC	65.62	42.11	75.30
WikiSample	79.01	59.17	84.06
TACKBP 2009	69.40	51.34	78.60
TACKBP 2010	72.45	48.63	87.35
TACKBP 2011	55.52	35.85	73.04
TACKBP 2012	26.45	21.85	49.91
Wikinews	86.66	66.25	91.56
Average	68.44	47.63	79.21





Gillick et al. 2019

f(text, text)



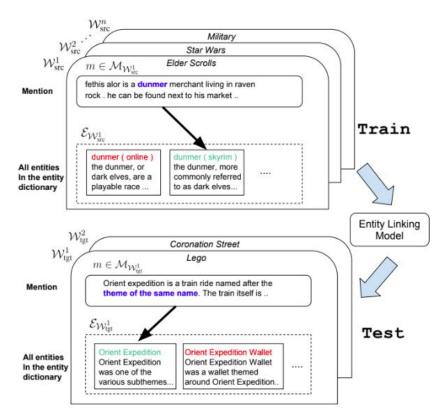
to indicate this is the focus mention

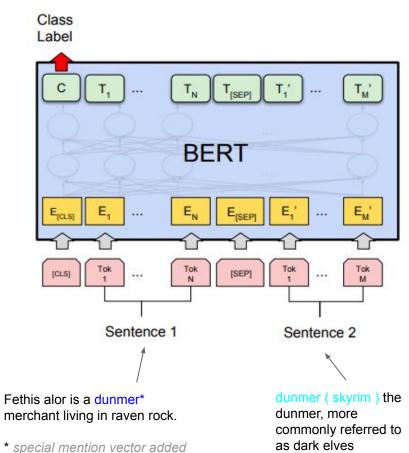




Zero-Shot Entity Linking

Lee et al. 2019



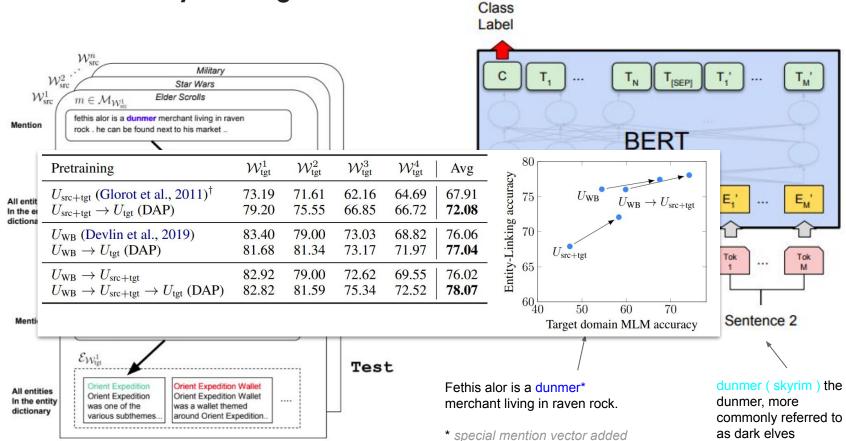






Zero-Shot Entity Linking

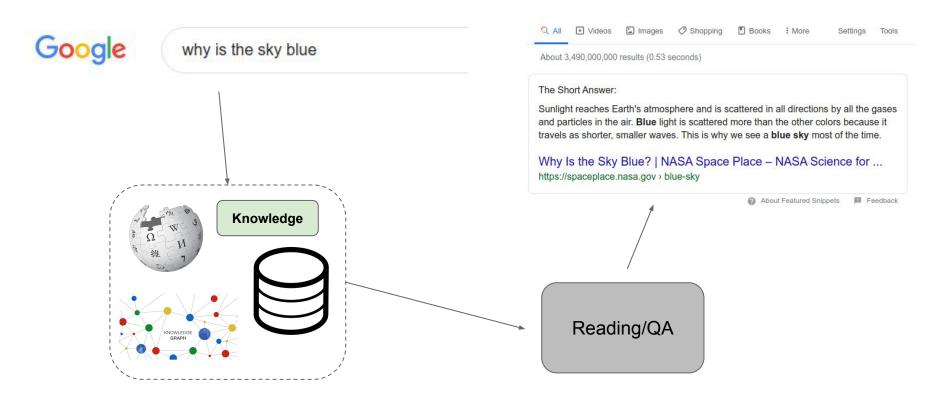
Lee et al. 2019



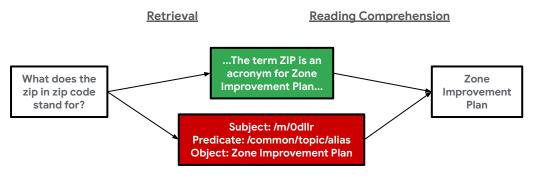
to indicate this is the focus mention

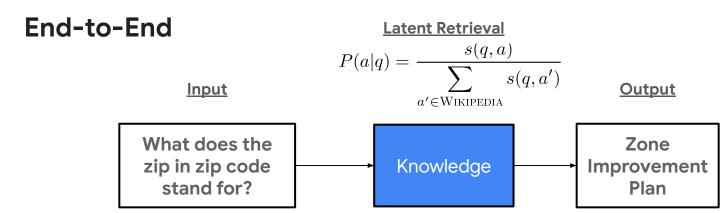


Knowledge Retrieval → Comprehension

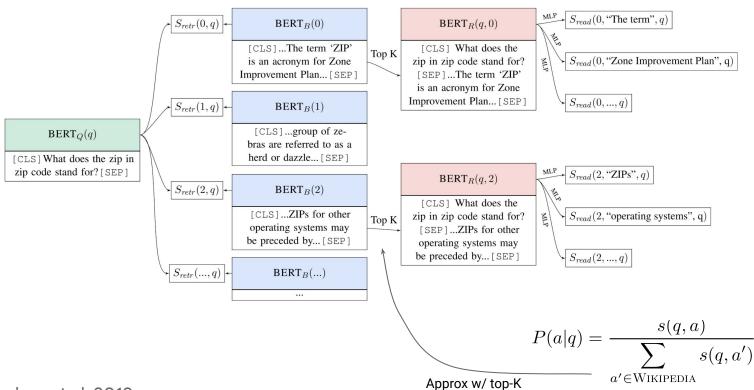


Pipelined Approach





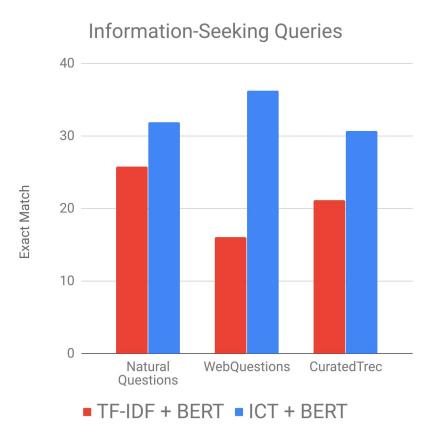
ORQA Overview





Results

End-to-end learning is crucial for information-seeking queries!



Sequential Question Answering

Building	City	Floors
First Canadian Place	Toronto	72
Commerce Court West	Toronto	57
Tour de la Bourse	Montreal	47
Place Ville-Marie	Montreal	44

What are the buildings in Toronto?
First Canadian Place, Commerce Court
West

Sequential Question Answering

Building	City	Floors
First Canadian Place	Toronto	72
Commerce Court West	Toronto	57
Tour de la Bourse	Montreal	47
Place Ville-Marie	Montreal	44

What are the buildings in Toronto?

First Canadian Place, Commerce Court

West

Of those, which buildings have more
than 60 floors?

First Canadian Place

Sequential Question Answering

Building	City	Floors
First Canadian Place	Toronto	72
Commerce Court West	Toronto	57
Tour de la Bourse	Montreal	47
Place Ville-Marie	Montreal	44

What are the buildings in Toronto?

First Canadian Place, Commerce Court
West

Of those, which buildings have more
than 60 floors?

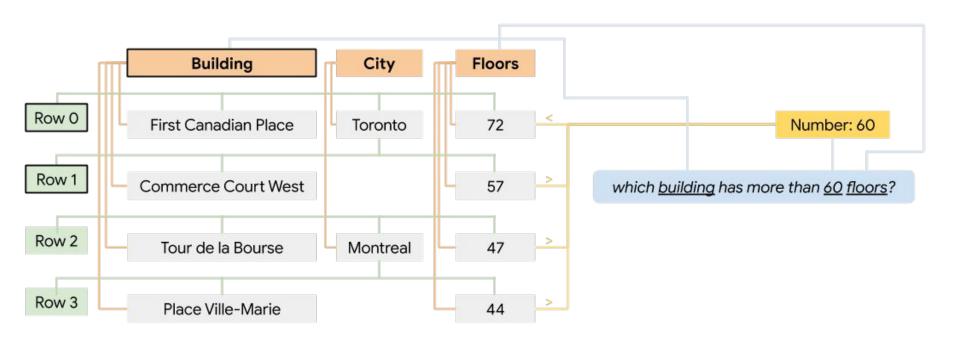
First Canadian Place

How many floors does it have?

72

Graph Transformer

Müller et al. 2019



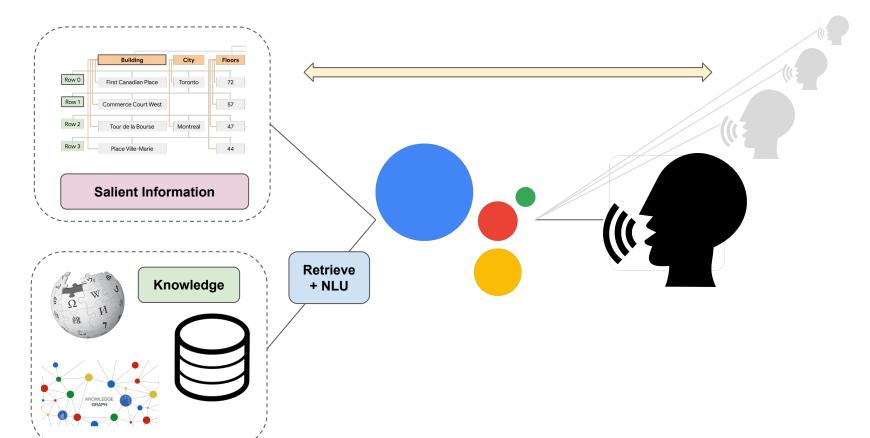
Results on SQA

Model	ALL	Q1	$\mathbf{Q2}$	Q3
Iyyer et al. (2017) †	44.7	70.4	41.1	23.6
Sun et al. (2018) † *	45.6	70.3	42.6	24.8
Müller et al. (2019) † *	55.1	67.2	52.7	46.8

Retrieval + NLP







Latent Retrieval for Weakly Supervised Onen Domain Question Answering

Kenton Lee Ming-Wei Chang Kristina Toutanova Google Research

Scortle, WA [kentonl,mingweichang,kristout]@google.com

Abstract

Recent work on onen domain oversion answing (QA) assumes strong supervision of the supporting evidence and/or assumes a black-box information ratrieval (IR) system to reare suboptimal, since gold evidence is not al-ways available, and QA is fundamentally difit is possible to jointly learn the retriever and a latest variable. Since this is impracti cal to learn from scratch, we pre-train the re-

the answer, a traditional IR system such as BM25 is sufficient. On datasets where a BM25 by up to 19 points in exact match.

Systems, there has been a revisal of inserts in open domain question answering (QA), where the open corpus must be considered rather than being given as input. This persons a from scratch. IR systems offer a reasonable but ruther that being given as ispet. This precents a Common approaches a size of the precent approaches the control and the properties a size of the bears of the leaves of t

et al., 2017), SearchQA (Dunn et al., 2017), and Oussay (Dhiners et al., 2017), the dependency of

These approaches rely on the IR system to mar rious ambiguity. However, QA is fundamentally different from IR (Singh, 2012). Whereas IR is concerned with lexical and semantic matching questions are by definition under-specified and re Instead of being subject to the recall ceiling from blackbox IR systems, we should directly learn to retrieve using question-answering data. In this work, we introduce the first Open

ORQA learns to retrieve evidence from an oper corpus, and is supervised only by question answer string pairs. While recent work on improving evidence retrieval has made significant Due to recent advances in reading comprehension they still only rerank a closed evidence set. The evidence must be retrieved from an open corpus, a latent variable that would be impractical to train

Petriesal Osertice Aspecting system (OPOA)

they also assume a reading correportensism model evidence. Given a pseudo-question, LT requires trained on aquestion-neuroever-ordineer tiles, seeks sheeling tiles, seeks seeks

Thanks!

where mentions must be linked to unseen en-tities without in-domain labeled data. The goal is to crable robust transfer to highly spestanding to resolve the new entities. First, we show that strong reading comprehension models pre-trained on large unlabeled data can be pre-training strategy, which we term domain-adaptive pre-training (DAP), to address the domain shift problem associated with linking unseen cutties in a new domain. We present for this task and show that DAP improves over strong pre-training baselines, including BERT. The data and code are available at https:

Abstract

1 Introduction

Entity linking systems have achieved high perambiguated mentions of entities in a target en-tity dictionary is available for training. Such amequiant anomas or entire is a ringer to the policiesty of another for milling. Soil with the policy of the high coverage this tolds, streamed dats, and lisking frequency suitables. For example, this and Wirse (2008) show that by only using the prior probability gathered must popular assistance on Wijselpad training anticles, one can achieve 90% accuracy on the ack of probability gathered to the policy of the contraction of the policy of the policy of the policy of the policy according to the side of probability gathered to tumore possible of entire. When the policy tumore possible of the tile Neuron without tumore possible entire. When the policy tumore possible entire to the policy tumore possible entire the policy tumore possible entir

*Work completed while interning at Google 'neshed stands for nero-shot entity linking.

We present the perso-silor entiry finking task, cialized domains, and so no metalane or more tables are assumed. In this setting, entities the set assumed that the metal that the metalane to apply that the set and that the metalane to apply the metalane to apply that the metalane to apply the metalane to apply that the metalane to apply the metalane to app Entity Laterry Model Module Crist specifier is a tion too server after the Module of the same tente. For tear specific. All unities in the united characteristics and the united chara Figure 1: Zern-shot entity linking. Multiple training

Zero-Shot Entity Linking by Reading Entity Descriptions

Lajanugen Logeswaran¹* Ming-Wei Chang¹ Kenton Lee¹ Kristina Toutanova¹ Jacob Devlin¹ Honolak Lee^{1,1}

[†]University of Michigan, [‡]Google Research

specialized entity dictionaries such as legal cases. Whispedia test articles.

While most prior works focus on linking to general entity databases, it is often desimble to lisk to

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova Google Al Language

Piacobdevlin.minmeichang.kentonl.kristoutManogle.com

Abstract

We introduce a new language representa-tion model called BERT, which stands for Bidirectional Eacoder Representations from Transformers. Unlike recent language spec-soration models (Purer et al., 2018s; Rad-ford et al., 2018s; BERT is designed to pre-train deep bidirectional representations from train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a re-sult, the mo-trained BERT model can be finelanguage inference, without substantial task specific architecture modifications.

specific architecture modifications. BIRTI is conceptually simple and empirically proverdid. It obtains new state-of-the-art re-subs on eleven nutural larguage processing tasks, including pushing the GLUE score to 193-96 (739 point abouts improvement), MathNLI accuracy to 8,736 (4,64% absolute improvement), 69,000,011 (appendix an according Test F1 to 93.2 (1.5 point aboute improvement), 61.1 point aboute improvement).

precessing mass (Lin and Le, 2017), refers or al., an operation answering, where it is uncon-2018k; Rafford et al., 2018; Boward and Ruder, 2018). These include sentence-level tasks such as

In this paper, we improve the fine-tuning based natural language inference (Bowman et al., 2015; Williams et al., 2018) and pumphrasing (Delan and Brocken, 2005), which aim to predet the re-and Brocken, 2005), which aim to predet the re-lationships between sentences by analyzing them rectionality constraint by using a "masked lanholistically, as well as token-level tasks such as guage model" (MLM) pre-training objective, inamond entity recognition and question answering, spired by the Close task (Taylor, 1953). The where needes are required to produce fine-grained mostlent at the token level (Tyong Kim Sang and the token from the input, and the objective is to De Moulder, 2003; Rajperkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to down-stream tasks: fonture-based and four-naving. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal trained parameters. The two approaches share the they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the

power of the pre-trained representations, especially for the fine-tuning approaches. The maunidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only at-tend to previous tokens in the self-attention layers of the Transformer (Vaccouni et al., 2017). Such re-1 Introduction strictions are sub-optimal for sentence-level tasks.

Language model pre-training has been shown to and could be very harmful when applying finebe effective for improving many natural language tuning based approaches to token-level tasks such precessing tasks (Dai and Le, 2015; Peters et al., as question answering, where it is crucial to incor-

Learning Dense Representations for Entity Retrieval

Illick@google.com	mayali@google.com	llansing@google.com	aprestalgo
Jason Baldridge			ego Garcia-Ol

Abstract We show that it is feasible to perform entity linking by training a dual encoder (two-tower) the same dense vector souce, where candidat cetities are retrieved by approximate nearest neighbor search. Unlike prior work, this setup does not rely on an alias table followed by a re-ranker, and is thus the first fully learned en-tity retrieval model. We show that our dual scoder, trained using only archor-text link in Wikipedia, outperforms discrete alias ta-ble and BM25 baselines, and is competitive with the best comparable results on the stan-dard TACKBP-2010 dataset. In addition, it Wikinews. On the modeling side, we demon-

negative mining algorithm for this task,

23

1 Introduction

*Equal Contributions †Work done during intermibile with Google

arbinity, had cordif, such as ody including the first post open personal resourcial with a patient can be replaced use. We shee that this configuration can be replaced used to be replaced to the configuration can be replaced and insertions in the overview upon. So in model allows considiate certains to be derively and efficiently interested for a sentence using nature neighbor seath. To over they artificial approach to desirable, we need the sentence of the sentence using nature neighbor seath. To over they artificial approach to desirable, consideration of the sentence of the senten

Larry Lansing' Alessandro Presta'

Costs has not played since being struck by the AC Milon forward... The alias table could be expanded so that last-name

aliases are added for all person entities, but it is impossible to come up with rules covering all scenarios. Consider this harder example:

1. Introduction

Actually pair of advantaging partial sprange in the control pair of advantaging partial, this is that of order to present the control pair of advantaging partial, this is that the control pair of the control pair of the control pair of the control pair of the pair is to the time a surge question as the control pair of the pair is to the time a surge question as the control pair of the pair is to the time a surge question as the control pair of the pair is to the time a surge question as the control pair of the p Origin.

This work includes the following contributions:

Answering Conversational Questions on Structured Data without Logical Forms

Thomas Müller, Francesco Piccinno, Massimo Nicosia, Peter Shaw, Vasemin Altun (thomasmueller, piccinno, massimon, petershaw, altun) @google.com

Abstract

We present a movel approach to answering so We present a nevel approach to answering superiorid apositive based on structured objects such as knowledge hasen or tables without superiorid apositive as an abstraction of the proper surrature. We excede table as graph to a service of the service for the housest are then selected from the curved to the housest are then selected from the excede graph names are then selected from the excede graph names are then selected from the exceded graph names are then selected from the exceeded graph names are then selected from the exceeded graph names are then selected from the exceeded sure them to a transfer mechanism reads of the survey to a queen our abstract the exceeded in seasons exceeded one made to exceede exceeded one made to exceed the exceeded one made to exceede exceeded one made to exceede exceeded to exceede exceeded one made to exceede exceeded to exceeded to exceede exceeded exceeded to exceede exceeded exceeded to exceede exceeded exceeded exceeded exceede

1 Introduction

In recent years, there has been significant progress on conversational question answering posed approach on the Sequential Question An-(OA), where questions can be meaninofully an-swering task (SOA) (fiver et al., 2017) which swered only within the context of a conversation (typer et al., 2017; Choi et al., 2018; Saha et al., 2018). This line of work, as in single QA settings, and context of a conversation the context of the context of a conversation of the same properties of the context performance on all questions, particularly on the follow-up questions 2018). This line of work, as in single QA settings, that require effective monoding of the cornect. 2016). This time of work, is it stripe (pt scrip, fills into two main outgoints, of) the answers are extracted from some text in a reading comprehension setting, (if) the answers are extracted from the setting of the answers are extracted from the behild a QA model for a sequence of questions structured objects, such as knowledge bases or ta-bles. The latter is commendy posed as a semantic passing task, where the goal is to map questions to some logical form which is then executed over the

weak supervision where training data consists of questions and answers alone with the structured resources to recover the logical form representa-tions that would yield the correct answer (Liang

In this paper, we follow this line of research and investigate answering sequential questions with respect to structured objects. In contrast to previous approaches, instead of learning the intermediate logical forms, we propose a novel approach that encodes the structured resources, i.e. tables, along with the questions and answers from the context of the conversation. This approach allow us to handle conversational contest without the definition of detailed operations or a vocabulary dependent on the logical form formalism that are required in the weakly supervised semantic purs

In semantic parsing, there is extensive work on over manually created forical forms in a super- table in relation to as which is the follow an ones over manually created hegical forms in a super-vised learning settly final rad Lings, [95] (E. Ling et al., 2016; Dong and Liapus, 2018). However, centing tabeled that for this tack can be expen-sive and time-constraing. This problem results in time-came that investigate sensing units with that 44 discided days that connect contents.

Matchine the Blanks: Distributional Similarity for Relation Learning

Livio Baldini Soares Nicholas FitzGerald Jeffrey Ling' Tom Kwiatkowski Goorle Research

{liviobs, nfitz, jeffreyling, tomkwiat}8google.com

Abstract

General purpose relations, are a consumer, which was model relative relations, are a core appealed to the consumer and the consumer and models to bodd portugue propose excesses that appeared relations with these under forms, or force to the consumer and the consumer and proposed to the consumer and the consumer and the consumer and the consumer and proposed to the consumer and the con

cantly outperform the previous methods on: mEval 2000 Task 8, KBP37, and TACRED

Reading text to identify and extract relations between entities has been a long standing good in natural language processing (Cardie, 1997). Typ-2014), or distantly supervised relation extractors (Minte et al., 2009) learn a mapping from text to relations in a limited schema. Ferming a sec-ond group, open information extraction removes representing relations using their surface forms (Banko et al., 2007; Fider et al., 2011; Stanovsky et al., 2018), which increases scope but also leads Work done as part of the Google Al residency.

to an associated lack of generality since many surface forms can express the same relation. Finally, the universal schema (Riedel et al., 2013) embraces both the diversity of text, and the concise resentation that has been extended to arbitrary ten teal input (European et al., 2015), and arbitrary entity pairs (Verga and McCallum, 2016). How-ever, like distantly supervised relation extractors. universal schema rely on large knowledge graphs

of Harris' distributional hypothesis (Harris, 1954) to relations, as well as recent advances in learning word representations from observations of their contexts (Mikolov et al., 2013; Peters et al., 2018; Devlin et al., 2018), we reprose a new method of learning relation representations directly from text. First, we study the ability of the Transformer neural network architecture (Vaswani et al., 2017) dentify a method of representation that outperforms previous work in supervised relation estrac tion. Then, we present a method of training this relation representation without any supervision from

[BLANK]'s rendition of "[BLANK]" has been called "one of the great songs" by Time, and is included on Rolling Stone's list of "The 500 Greatest Songs of All Time."

the limitations of a predefined schema by instead

Figure 1: "Maching the blacks" example where both rela-

Following Riedel et al. (2013), we assume aclinked to unique identifiers and we define a refo-