IJCAI-21 Tutorial

Machine Reading Comprehension: The Role of Pre-trained Language Models and Beyond

Zhuosheng Zhang
zhangzs@sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhangzs

Hai Zhao
zhaohai@cs.sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhaohai

Schedule

- Part 1: Machine Reading Comprehension (Zhuosheng Zhang)
- Break (5 min)
- Part 2: Pre-trained Language Model (Hai Zhao)
- Break (10 min)
- Part 3: Technical Methods, Discussions, and Frontiers (Zhuosheng Zhang)

Machine Reading Comprehension

Zhuosheng Zhang
zhangzs@sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhangzs

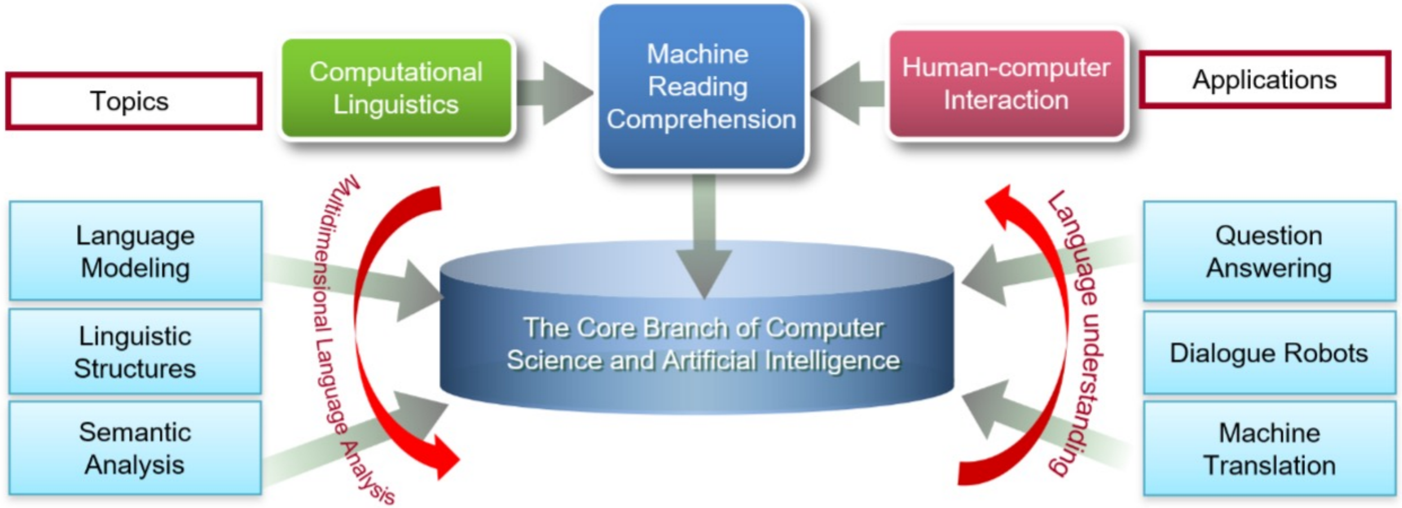
Introductions to MRC

There are two categories of branches in natural language processing (NLP)

- Core/fundamental NLP
  - Language model/representation
  - Linguistic structure parsing/analysis
    - Morphological analysis/word segmentation
    - Syntactic/semantic/discourse parsing
    - …
- Application NLP
  - Machine Reading Comprehension (MRC)
  - Text Entailment (TE) or Natural Language Inference (NLI)
    - SNLI, GLUE
  - QA/Dialogue
  - Machine translation
  - …

Introductions to MRC

- Aim: teach machines to read and comprehend human languages
- Form: find the accurate Answer for a Question according to a given Passage (document).
- Types
  - Cloze-style
  - Multi-choice
  - Span extraction
  - Free-form



- Before 2015
  - MCTest
  - ProcessBank

From shared task to leaderboard

- After 2015
  - CNN/Daily Mail
  - Children Book Test
  - WikiReading
  - LAMBADA
  - SQuAD
  - Who did What
  - NewsQA
  - MS MARCO
  - TriviaQA
  - CoQA
  - QuAC
  - ……

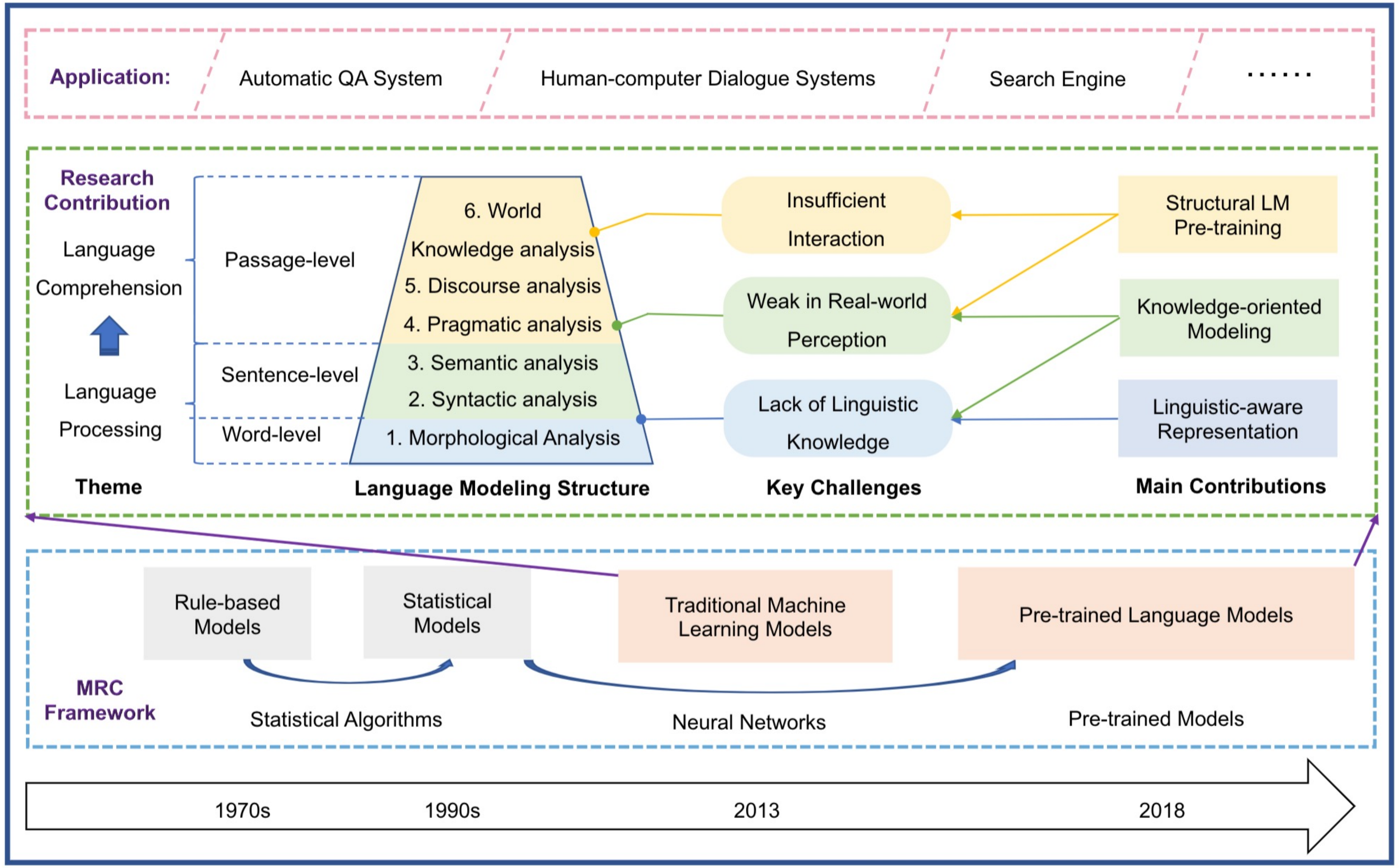
Introductions to MRC

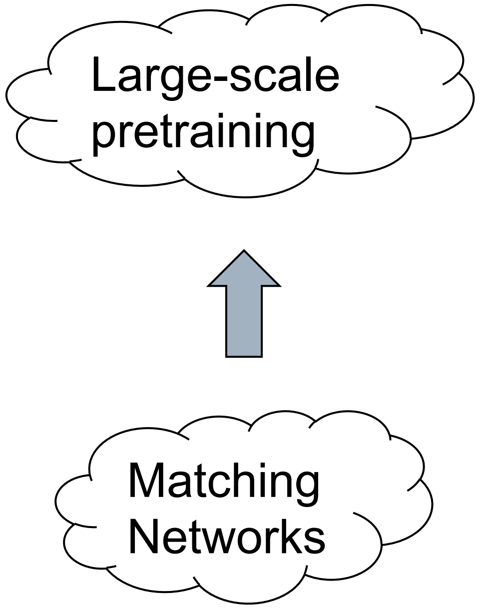
Cloze-style	from CNN (Hermann et al. 2015)
Context	*(@entity0) – a bus carrying members of a @entity5 unit overturned at an @entity7 military base sunday , leaving 23 @entity8 injured , four of them critically , the military said in a news release . a bus overturned sunday in @entity7 , injuring 23 @entity8 , the military said . the passengers , members of @entity13 , @entity14 , @entity15 , had been taking part in a training exercise at @entity19 , an @entity21 post outside @entity22 , @entity7 . they were departing the range at 9:20 a.m. when the accident occurred . the unit is made up of reservists from @entity27 , @entity28 , and @entity29 , @entity7 . the injured were from @entity30 and @entity31 out of @entity29 , a @entity32 suburb . by mid-afternoon , 11 of the injured had been released to their unit from the hospital . pictures of the wreck were provided to the news media by the military . @entity22 is about 175 miles south of @entity32 . e-mail to a friend*
Question	*bus carrying @entity5 unit overturned at ______ military base*
Answer	*@entity7*
Multi-choice	**from RACE (Lai et al. 2017)**
Context	*Runners in a relay race pass a stick in one direction. However, merchants passed silk, gold, fruit, and glass along the Silk Road in more than one direction. They earned their living by traveling the famous Silk Road. The Silk Road was not a simple trading network. It passed through thousands of cities and towns. It started from eastern China, across Central Asia and the Middle East, and ended in the Mediterranean Sea. It was used from about 200 B, C, to about A, D, 1300, when sea travel offered new routes, It was sometimes called the world ' s longest highway. However, the Silk Road was made up of many routes, not one smooth path. They passed through what are now 18 countries. The routes crossed mountains and deserts and had many dangers of hot sun, deep snow, and even battles. Only experienced traders could return safely.*
Question	*The Silk Road became less important because ______.*
Answer	*A.it was made up of different routes* *B.silk trading became less popular* ***C.sea travel provided easier routes*** *D.people needed fewer foreign goods*

Span Extraction	from SQuAD (Rajpurkar et al. 2016)
Context	*Robotics is an interdisciplinary branch of engineering and science that includes mechanical engineering, electrical engineering, computer science, and others. Robotics deals with the design, construction, operation, and use of robots, as well as computer systems for their control, sensory feedback, and information processing. These technologies are used to develop machines that can substitute for humans. Robots can be used in any situation and for any purpose, but today many are used in dangerous environments (including bomb detection and de-activation), manufacturing processes, or where humans cannot survive. Robots can take on any form, but some are made to resemble humans in appearance. This is said to help in the acceptance of a robot in certain replicative behaviors usually performed by people. Such robots attempt to replicate walking, lifting, speech, cognition, and basically anything a human can do.*
Question	*What do robots that resemble humans attempt to do?*
Answer	*replicate walking, lifting, speech, cognition*
Free-form	**from DROP (Dua et al. 2019)**
Context	*The Miami Dolphins came off of a 0-3 start and tried to rebound against the Buffalo Bills. After a scoreless first quarter the Dolphins rallied quick with a 23-yard interception return for a touchdown by rookie Vontae Davis and a 1-yard touchdown run by Ronnie Brown along with a 33-yard field goal by Dan Carpenter making the halftime score 17-3. Miami would continue with a Chad Henne touchdown pass to Brian Hartline and a 1-yard touchdown run by Ricky Williams. Trent Edwards would hit Josh Reed for a 3-yard touchdown but Miami ended the game with a 1-yard touchdown run by Ronnie Brown. The Dolphins won the game 38-10 as the team improved to 1-3. Chad Henne made his first NFL start and threw for 115 yards and a touchdown.*
Question	*How many more points did the Dolphins score compare to the Bills by the game's end?*
Answer	*28*

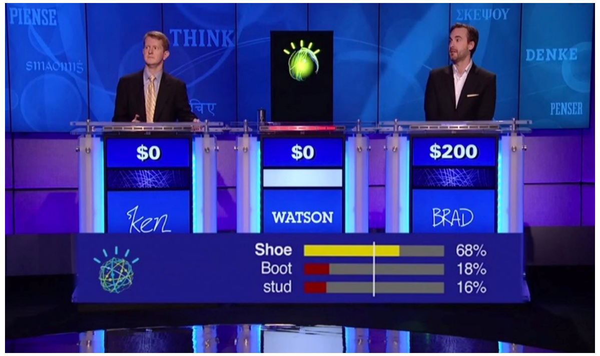
The Boom of MRC researches

- The burst of deep neural networks, especially attention-based models
- The evolution of pre-trained language models (large-scale pre-training and task-specific



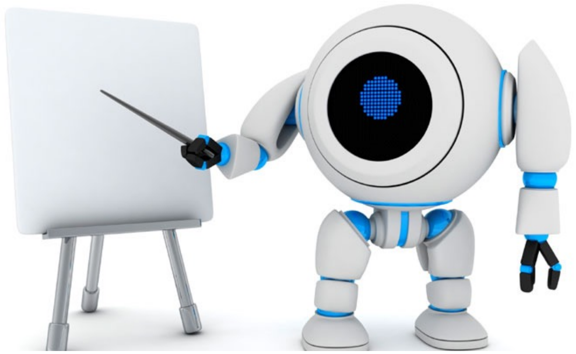


Applications



Question Answering


Dialogue System


Intelligent Teacher

Fake News Identifier


Legal Advisor


Medical Diagnosis

Classic NLP Meets MRC

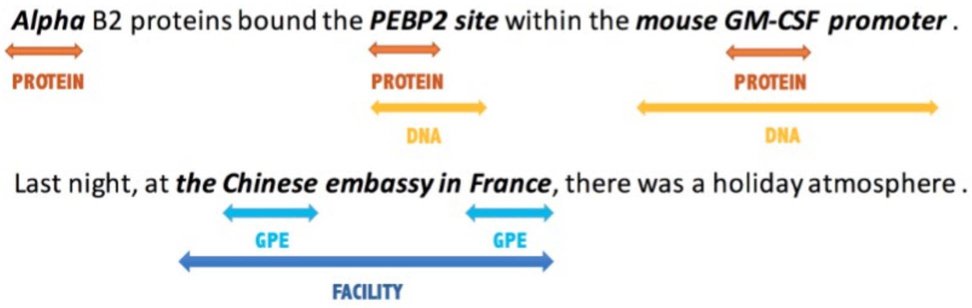
MRC has great inspirations to the NLP tasks.

- strong capacity of MRC-style models, e.g., similar training pattern with pre-training of PrLMs
- unifying different tasks as MRC formation, and taking advantage of multi-tasking to share knowledge.

Most NLP tasks can benefit from the new task formation as MRC, including question answering, machine translation, summarization, natural language inference, sentiment analysis, relation extraction, dialogue, etc.

Example: nested named entity recognition

Questoin: Find XXX in the text.



Better performance

Related paper:

[1] MCCANN, Bryan, et al. The natural language decathlon: Multitask learning as question answering. *arXiv:1806.08730*, 2018.

[2] KESKAR, Nitish Shirish, et al. Unifying Question Answering, Text Classification, and Regression via Span Extraction. *arXiv:1904.09286*, 2019.

[3] KESKAR, Nitish Shirish, et al. Unifying Question Answering, Text Classification, and Regression via Span Extraction. *arXiv:1904.09286*, 2019.

[4] LI, Xiaoya, et al. Entity-Relation Extraction as Multi-Turn Question Answering. ACL 2019. p. 1340-1350.

[5] LI, Xiaoya, et al. A Unified MRC Framework for Named Entity Recognition. ACL 2020.

MRC Goes Beyond QA

MRC is a generic concept to **probe for language understanding capabilities**

-> difficulty to measure directly.

QA is a fairly simple and effective **format**.

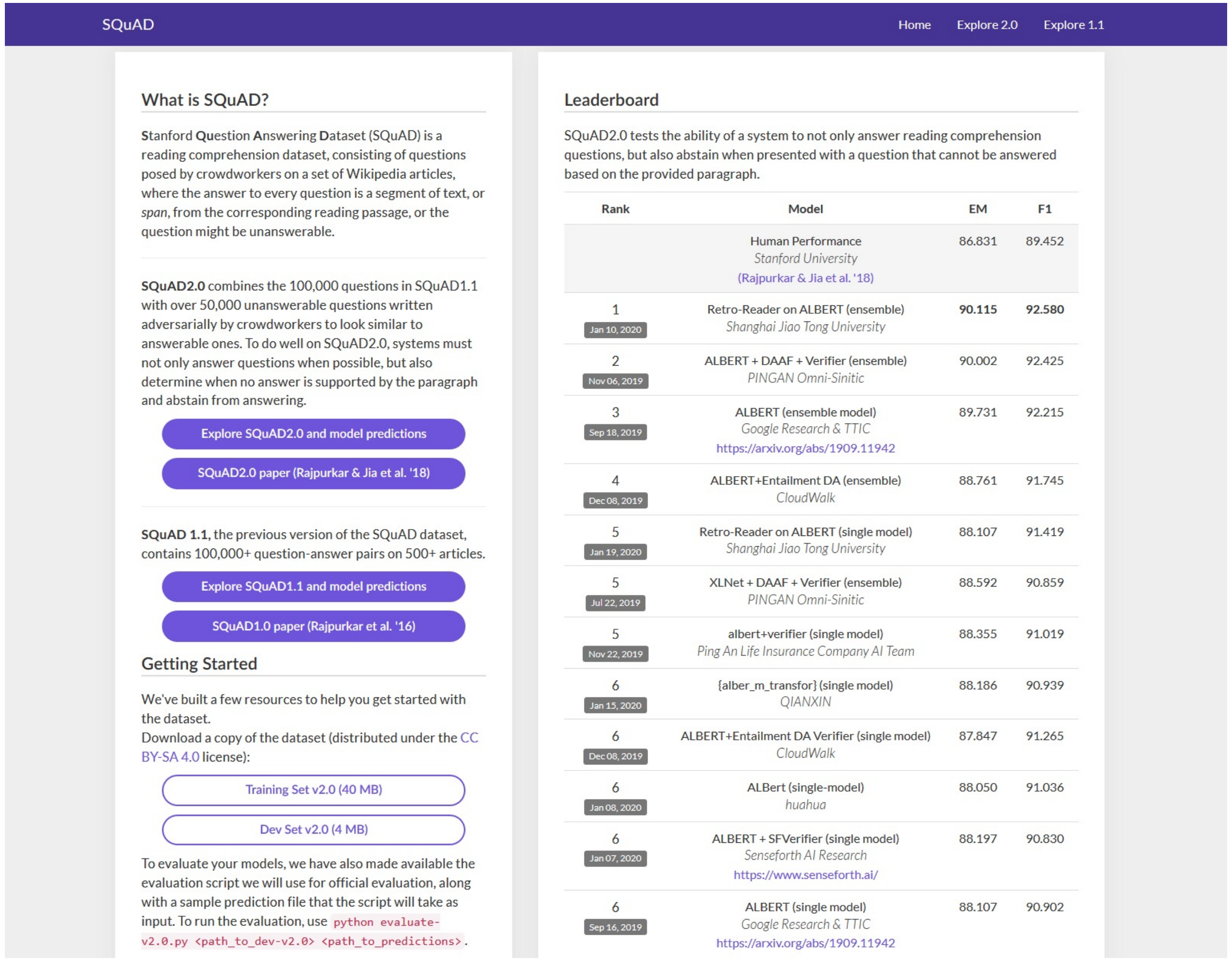
Reading comprehension is an old term to measure the knowledge accrued through reading.

MRC goes beyond the traditional QA, such as factoid QA or knowledge base QA

- reference to open texts
- avoiding efforts on retrieving facts from a structured manual-crafted knowledge corpus.

Sources

- Leaderboards
  - SQuAD v1.1/2.0
  - RACE
  - CoQA
  - QuAC
  - DREAM
  - MuTual
  - ShARC
  - …
- Venues
  - AI/ML: NeurIPS, IJCAI, AAAI, etc.
  - NLP/CL: ACL, EMNLP, COLING, etc.
- Surveys
  - Chen et al, 2018. Neural Reading Comprehension and Beyond
  - Zhang et al, 2020. Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond



SQuAD   Home   Explore 2.0   Explore 1.1

What is SQuAD?

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

Getting Started

We've built a few resources to help you get started with the dataset.

Download a copy of the dataset (distributed under the CC BY-SA 4.0 license):

Training Set v2.0 (40 MB)

Dev Set v2.0 (4 MB)

To evaluate your models, we have also made available the evaluation script we will use for official evaluation, along with a sample prediction file that the script will take as input. To run the evaluation, use `python evaluate-v2.0.py <path_to_dev-v2.0> <path_to_predictions>`.

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University*	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Dec 08, 2019	ALBERT+Entailment DA (ensemble) *CloudWalk*	88.761	91.745
5 Jan 19, 2020	Retro-Reader on ALBERT (single model) *Shanghai Jiao Tong University*	88.107	91.419
5 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
5 Nov 22, 2019	albert+verifier (single model) *Ping An Life Insurance Company AI Team*	88.355	91.019
6 Jan 15, 2020	{alber_m_transfor} (single model) *QIANXIN*	88.186	90.939
6 Dec 08, 2019	ALBERT+Entailment DA Verifier (single model) *CloudWalk*	87.847	91.265
6 Jan 08, 2020	ALBert (single-model) *huahua*	88.050	91.036
6 Jan 07, 2020	ALBERT + SFVerifier (single model) *Senseforth AI Research* https://www.senseforth.ai/	88.197	90.830
6 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902

Part 2

Pre-trained Language Models

Hai Zhao
zhaohai@cs.sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhaohai

Pre-trained Language Models (PrLMs)

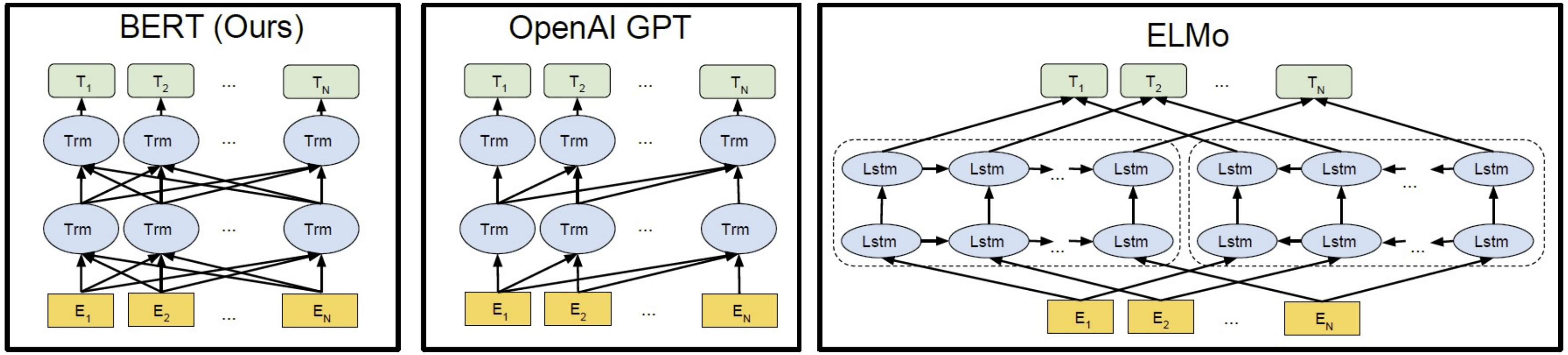
- Pre-trained Model ×
  - Unable to distinguish non- language models
- Pre-trained Language Model √
  - Unable to distinguish non-contextualized language models like Word2Vec and GloVe
- Pre-trained Language Representation √
- Pretrained Contextualized Language Model √ √
- Pre-trained Contextualized Language Representation Model √ √

Working Mode

Essential characteristics different from existing language models

Embedding Form

Contextualized Representations



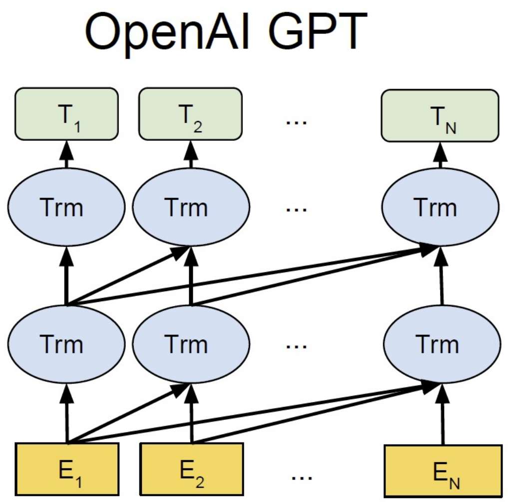
ELMo: Embedding from Language Models

GPT: Generative Pre-Training

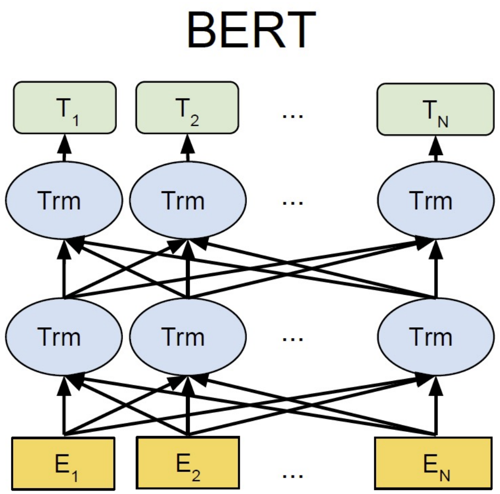
BERT: Pre-training of Deep Bidirectional Transformers

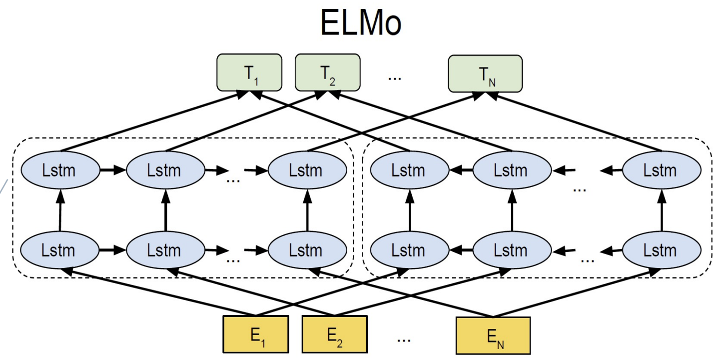
[1] Peters, Matthew E., et al. Deep contextualized word representations. NAACL-HLT. 2018.
[2] Radford, Alec, et al. Improving language understanding by generative pre-training. (2018).
[3] Devlin, Jacob, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT. 2019.

From GPT、ELMo、Word2Vec to BERT



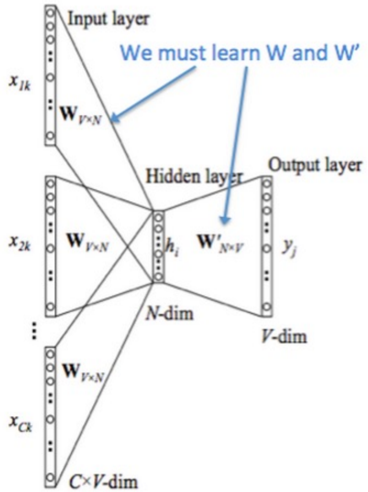
Bi-directional
(idea)





Transformer as feature extractor

Bi-directional language modeling
(method)



Contextualized Language Encoding

(Sentence/Contextual) Encoder as a Standard Network Block

- Word embeddings have changed NLP
- However, sentence is the least unit that delivers complete meaning as human use language
- Deep learning for NLP quickly found it is a frequent requirement on using a network component encoding a sentence input.
  - Encoder for encoding the complete sentence-level Context
- Encoder differs from sliding window input that it covers a full sentence.
- It especially matters when we have to handle passages in MRC tasks, where passage always consists of a lot of sentences (not words).
  - When the model faces passages, sentence becomes the basic unit
  - Usually building blocks for an encoder: RNN, especially LSTM

Traditional Contextualization:
Word embedding
+
Sentence Encoder

From Language Models to Language Representation

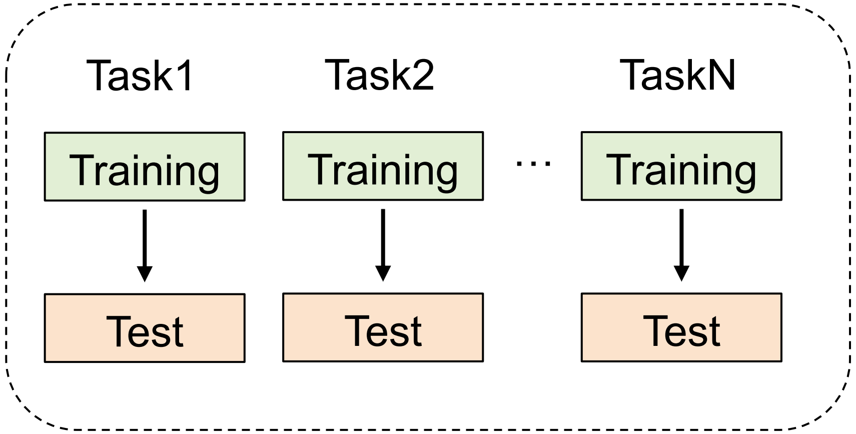
- MRC and other application NLP need a full sentence encoder,
  - Deep contextual information is required in MRC
  - Word and sentence should be represented as embeddings.
- Model can be trained in a style of *n*-gram language model
- So that there comes the language representation which includes
  - *n*-gram language model (**training object**), plus
  - Embedding (**representation form**), plus
  - Contextual encoder (**model architecture**)
  - Usage

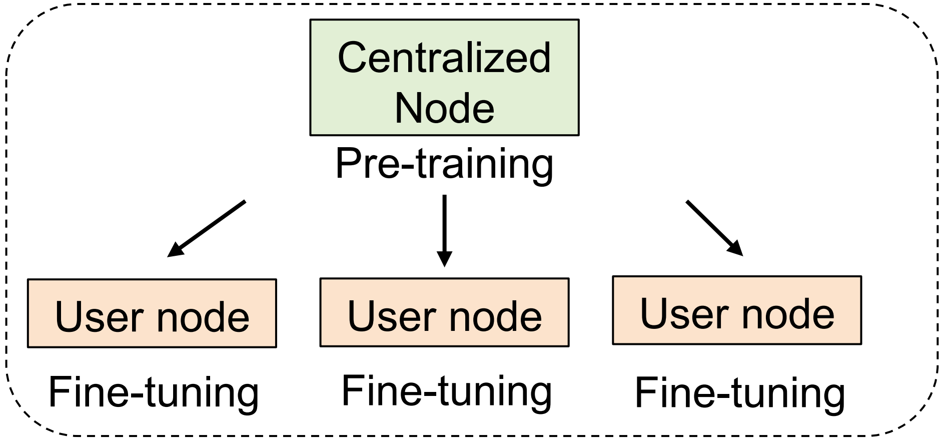
LM Contextualization:
Sentence -> Encoder -> Repr.

→ The representation for each word depends on the entire context in which it is used, **dynamic embedding**.

Model	Repr. form	Context	Training object	Usage
n-gram LM	One-hot	Sliding widow	*n*-gram LM (MLE)	Lookup
Word2vec/GloVe	Embedding	Sliding widow	*n*-gram LM (MLE)	Lookup
Contextualized LM	Embedding	Sentence	*n*-gram LM (MLE), +ext	Fine-tune

PrLM: New Paradigm





Previous

Each user trains individual machine learning models for each task.

Now

The central node trains the generalized language model (pre-training) and provides the nearly completed model for users as the standard module for task-specific fine-tuning.

Individual training ➡ Centralized pre-training + individual fine-tuning

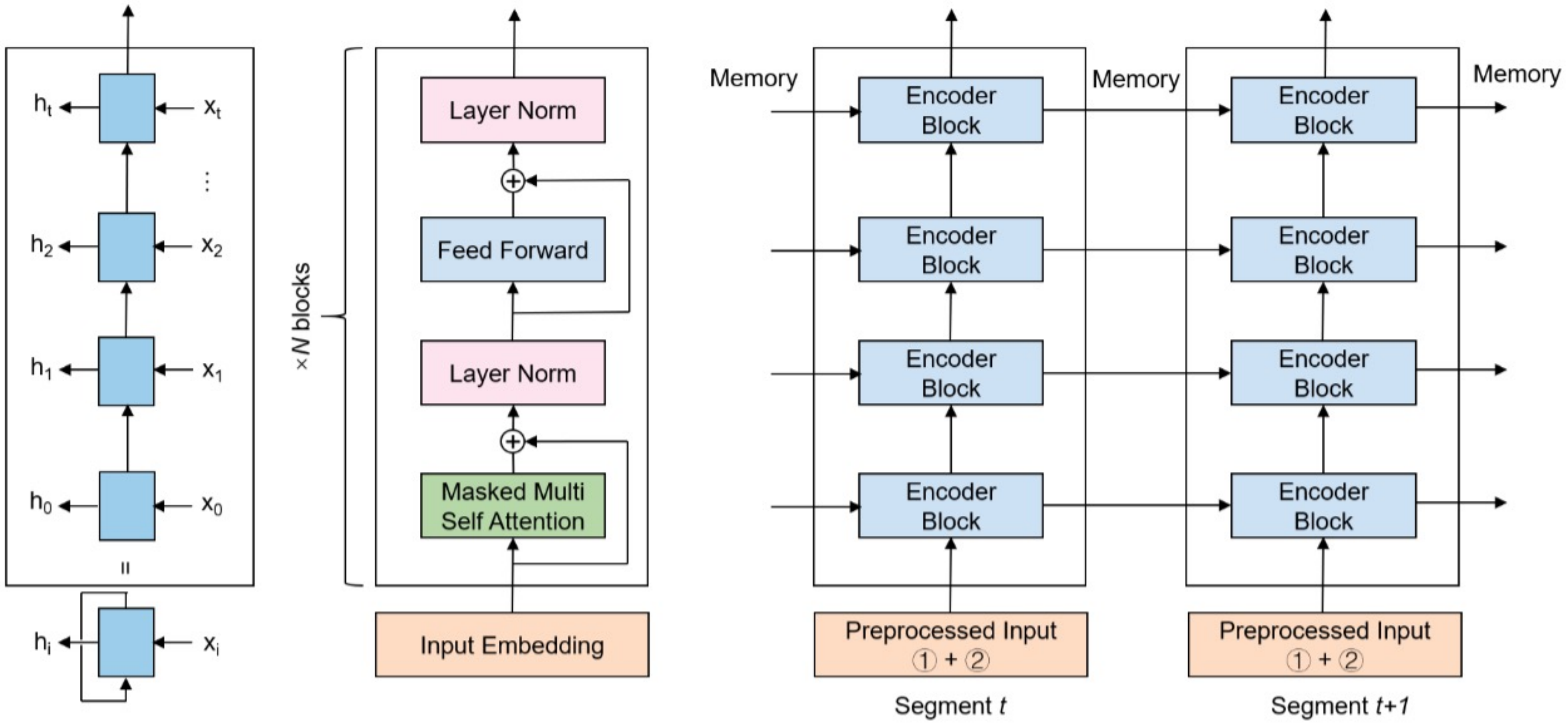
Extreme case： GPT3 gives predictions directly after pre-training, eliminating the fine-tuning process

The Elements of PrLMs

- Encoder architectures
  - RNN/Transformer/…
- Training objectives
  - (Autoregressive / denoising) task construction
- Sampling (training) methods

Architectures of PrLMs

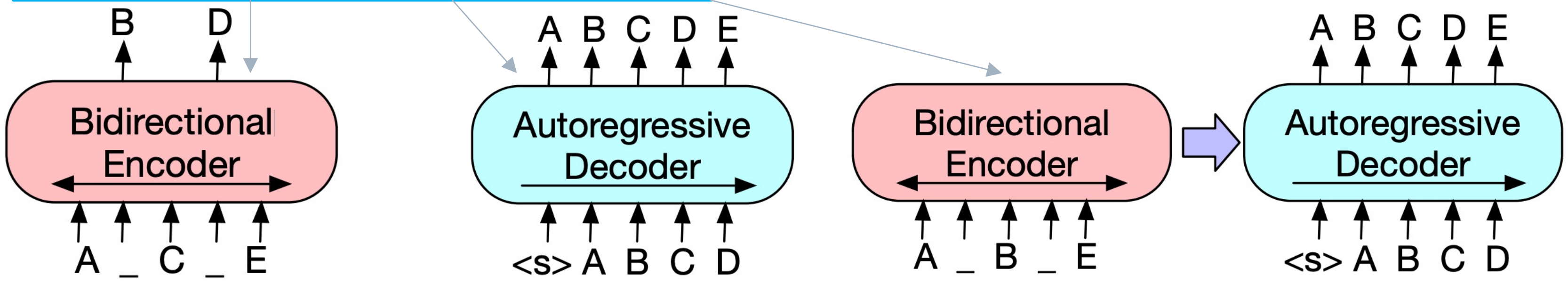
- RNN: GRU/LSTM
- Transformer
- Transformer-XL



(a) RNN (b) Transformer (c) Transformer-XL

Training Objectives

- Constructing the training samples with generalized **autoregressive** method
- Discriminative vs. Generative
  - **Discriminative**: Predict the corrupted tokens (BERT, ALBERT, ELECTRA, etc)
    → Useful for discriminative tasks like span-based MRC
  - **Generative** : Predict the complete sentence via Decoder (GPT 1-3, etc)
    → Helpful for generative tasks like machine translation
  - **Discriminative + Generative** : Predict the complete sentence via Decoder (BART)



(a) Discriminative (BERT) (b) Generative (GPT) (c) Discriminative + Generative (BART)

Lewis, Mike, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

The Evolution of PrLM Training Objectives

The core is the evolution of PrLM training objectives: n-gram, masked LM, permutation LM, etc.

The standard and common objective: n-gram LM.

An n-gram Language model yields a probability distribution over text (n-gram) sequences.

Sequence: (Sentence) $w_1 \quad w_2 \quad \cdots \quad \underbrace{w_i \quad \cdots \quad w_{i+n\text{-}1}}_{n\text{-gram}} \quad \cdots \quad w_L$

Probability of the sequence:

$$p(\mathbf{w}) = p(w_i \mid w_{i:i+n-2}),$$

Training objective:

$$\max_{\theta} \sum_{\mathbf{w}} \log p_{\theta}(\mathbf{w}),$$

Model	Repr. form	Context	Training object	Usage
n-gram LM	One-hot	Sliding widow	n-gram LM (MLE)	Lookup
Word2vec/GloVe	Embedding	Sliding widow	n-gram LM (MLE)	Lookup
Contextualized LM	Embedding	Sentence	n-gram LM (MLE), +ext	Fine-tune

Model	Loss	2^{nd} Loss	Direction	Encoder arch.	Input
ELMo	n-gram LM	-	Bi	RNN	Char
GPT_{v1}	n-gram LM	-	Uni	Transformer	Subword
BERT	Masked LM	NSP	Bi	Transformer	Subword
RoBERTa	Masked LM	-	Bi	Transformer	Subword
ALBERT	Masked LM	SOP	Bi	Transformer	Subword
XLNet	Permu. n-gram LM	-	Bi	Transformer-XL	Subword
ELECTRA	Masked LM	RTD	Bi	GAN	Subword

The Evolution of PrLM Training Objectives

When n expands to the maximum, the conditional context thus corresponds to the whole sequence

$$\sum_{k=c+1}^{L} \log p_\theta(w_k \mid w_{1:k-1}),$$

A **bidirectional form**:

$$\sum_{k=c+1}^{L} (\log p_\theta(w_k \mid w_{1:k-1}) + \log p_\theta(w_k \mid w_{k+1:L})),$$

ELMo

So, what are the Masked LM (MLM) and Permuted LM (PLM)?

MLM (BERT): tokens in a sentence are randomly replaced with a special mask symbol

$$\sum_{k \in \mathcal{D}} \log p_\theta(w_k \mid \mathbf{s}') \qquad \mathbf{s}' = \{w_1, [M], w_4, [M], w_5\}$$ where D denote the set of masked positions.

PLM (XLNet): maximize the expected log-likelihood of all possible permutations of the factorization order

-> Autoregressive n-gram LM! $$\mathbb{E}_{z \in \mathcal{Z}_L} \sum_{k=c+1}^{L} \log p_\theta(w_{z_k} \mid w_{z_{1:k-1}}).$$

where z means the permutation and c is the cutting point of a non-target conditional subsequence $z \leq c$ and a target subsequence $z > c$.

A Unified Form

MLM can be seen as a variant of n-gram LM to a certain extent --- bidirectional autoregressive n-gram LM (a).

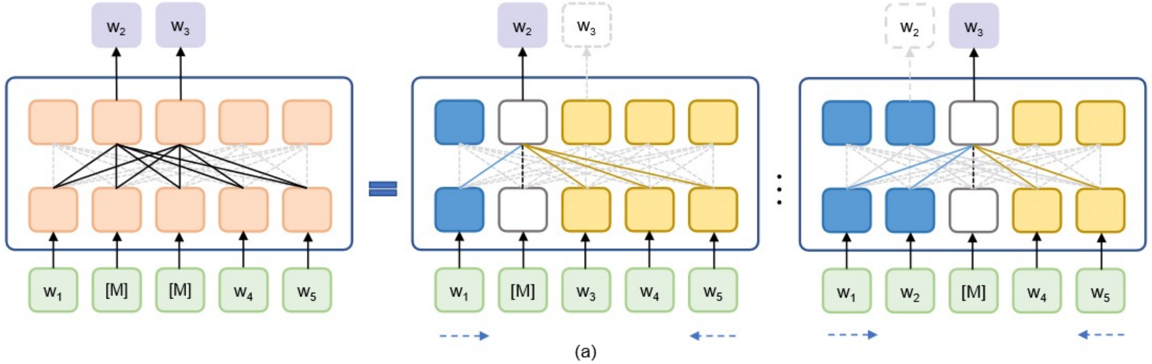
≈ BERT vs. ELMo

Naturally, the self-attention can attend to tokens from both sides.

MLM can be directly unified as PLM when the input sentence is permutable (with **insensitive word orders**) (b-c)

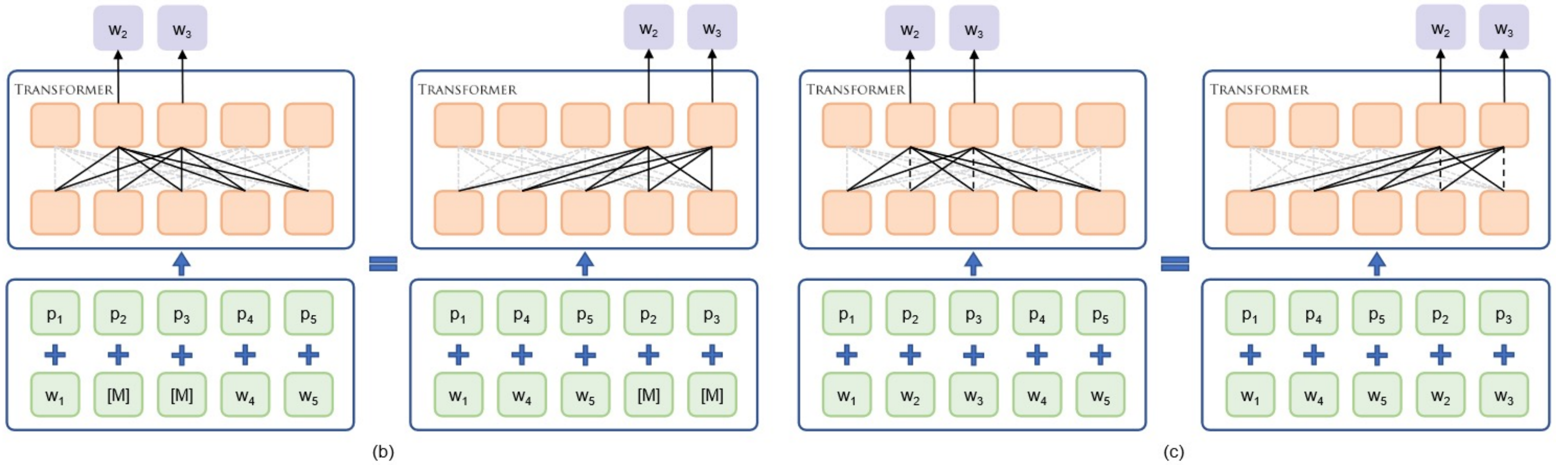
≈ BERT -> XLNet

$$\mathbb{E}_{z \in \mathcal{Z}_L} \sum_{k=c+1}^{L} \log p_\theta(w_{z_k} \mid w_{z_{1:c}}, M_{z_{k:L}}),$$



Transformer takes token positions in a sentence as inputs

-> **not sensitive to the absolute input order of these tokens.**



MPNet: Masked LM + Premuted LM

Training Objectives (Denoising)

- LM is an automatic denoising encoder in language
- Manually constructing different levels of corrupted units of natural language text
- ➔Edit Operations
  - deletion
  - addition
  - permutation/reordering
  - replacement
- Levels of language units:
  - word
  - sentence
  - passage

	word	sentence
deletion	Masking	NSP
replacement		
addition		
permutation	XLNet ?	SOP

- Training strategies:
  - direct prediction
  - generative-discriminative (Electra)

BERT

BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Huge Parameters:

BERT base: L=12, H=768, A=12, Total Parameters=110M

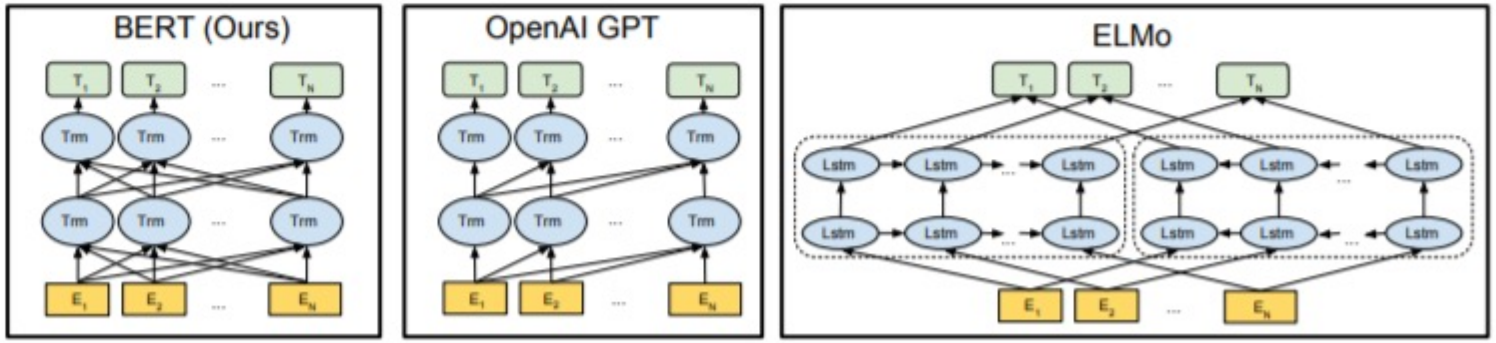
BERT large: L=24, H=1024, A=16, Total Parameters=340M

(L-transformer blocks, H - dimension of hidden state, A – self-attention heads)

Large corpus: BooksCorpus (800M words) + English Wikipedia (2,500M words)

Computing power: BERT base 16 TPU*4 day BERT large 64 TPU *4 day

BERT vs GPT vs ELMo



Devlin, Jacob, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT. 2019.

BERT Pre-training

Task #1: Masked LM
- replace the chosen words with [MASK] then predict it
- Not always replace the word with [MASK]

- 80% of the time: Replace the word with the [MASK] token, e.g., `my dog is hairy` → `my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy` → `my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy` → `my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.

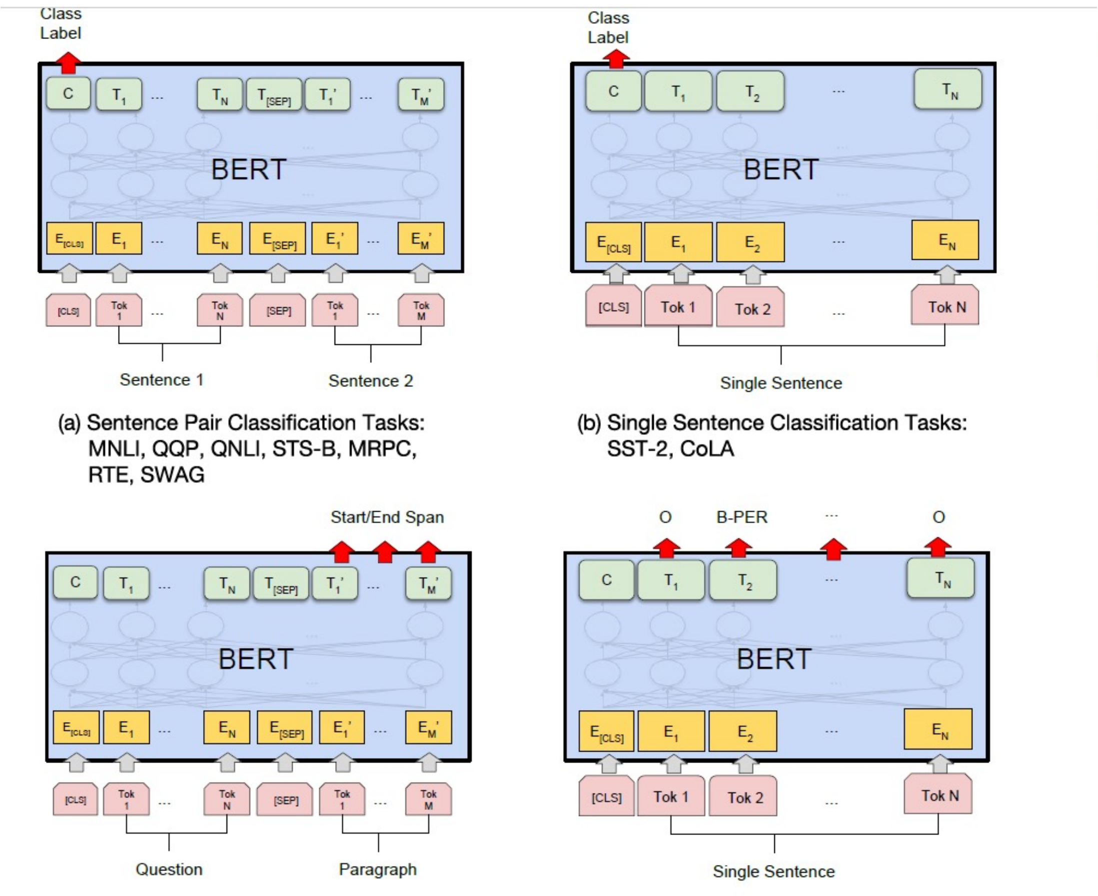
Task #2: Next Sentence Prediction
- [CLS] sentence A [SEP] sentence B [SEP]
- 50% of the time B is the actual next sentence that follows A, and 50% of the time it is a random sentence from the corpus

Input = `[CLS] the man went to [MASK] store [SEP]`
`he bought a gallon [MASK] milk [SEP]`
Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP]`
`penguin [MASK] are flight ##less birds [SEP]`
Label = `NotNext`

Devlin, Jacob, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT. 2019.

BERT Fine-tuning



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

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

(c) Question Answering Tasks: SQuAD v1.1

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

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	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD results. The BERT ensemble is 7x

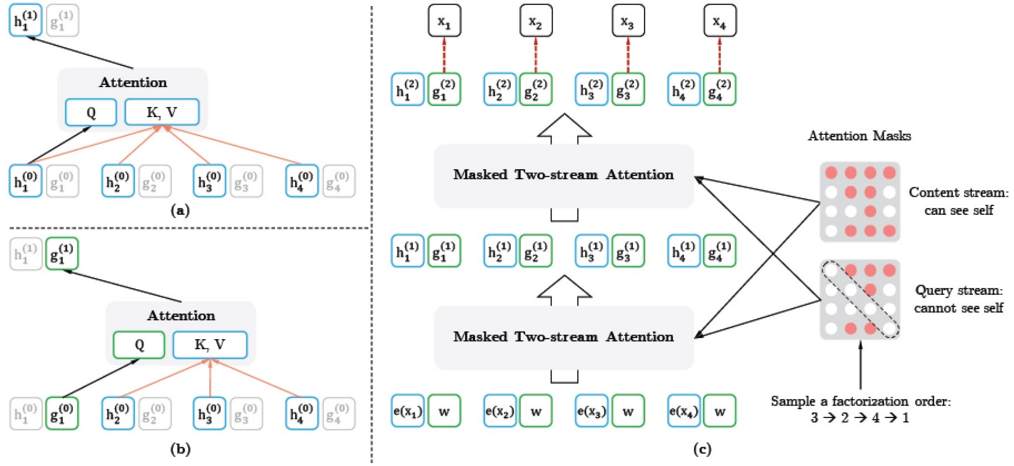
System	Dev F1	Test F1
ELMo+BiLSTM+CRF	95.7	92.2
CVT+Multi (Clark et al., 2018)	-	92.6
$\text{BERT}_{\text{BASE}}$	96.4	92.4
$\text{BERT}_{\text{LARGE}}$	**96.6**	**92.8**

Table 3: CoNLL-2003 Named Entity Recognition results. The hyperparameters were selected using the Dev set, and the reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.

Devlin, Jacob, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT. 2019.

XLNet: Token Permutation

- Token permutation + Two-stream Attention
  - Using autoregressive mechanism to overcome the shortcomings of BERT (Masked LM)
  - Permute the tokens in the sentence, and make the LM predictions



- Computation：512 TPU v3，500K steps, batch size = 2048, 2.5 days

Training corpus：
- 13G: BooksCorpus + English Wikipedia
- 16G: Giga5
- 19G: ClueWeb 2012-B
- 78G: Common Crawl

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding, NeurIPS 2019.

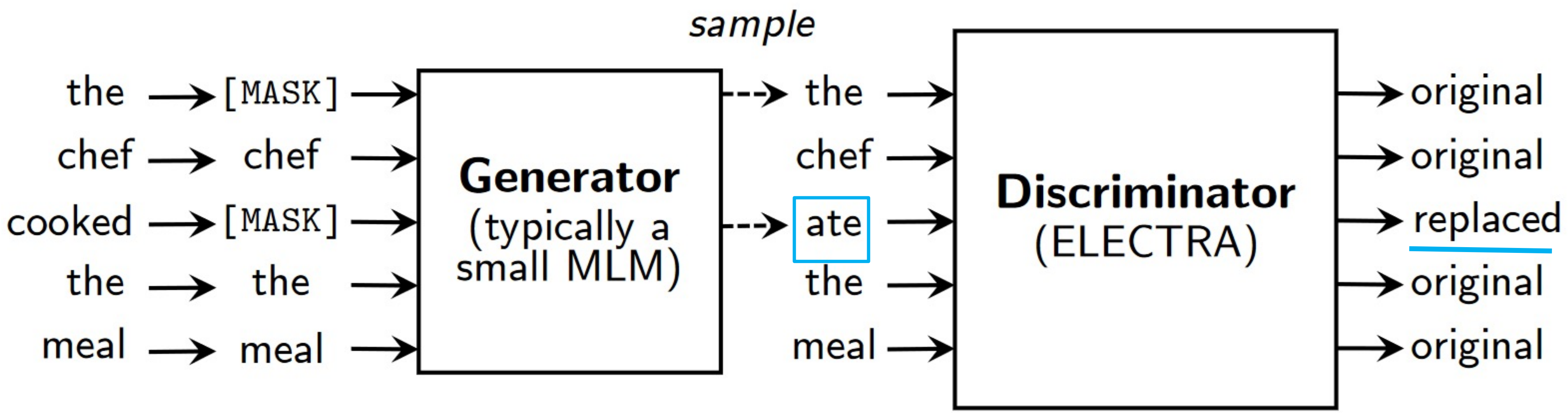
ALBERT: Sentence Order Prediction

- Three improvements：
  - Modify the Embedding (E) and hidden states (H) into the dimension H>>E， instead of E=H in BERT
  - Use full layer parameter sharing, including all forward networks and attention weights (significantly reduce the model size)
  - Modify the sentence training objective (NSP) of BERT to sentence order prediction (SOP)

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representation. *ICLR* 2020.

ELECTRA

- Predicts whether each token in the corrupted input was replaced by a generator sample or not.



Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ICLR* 2020.

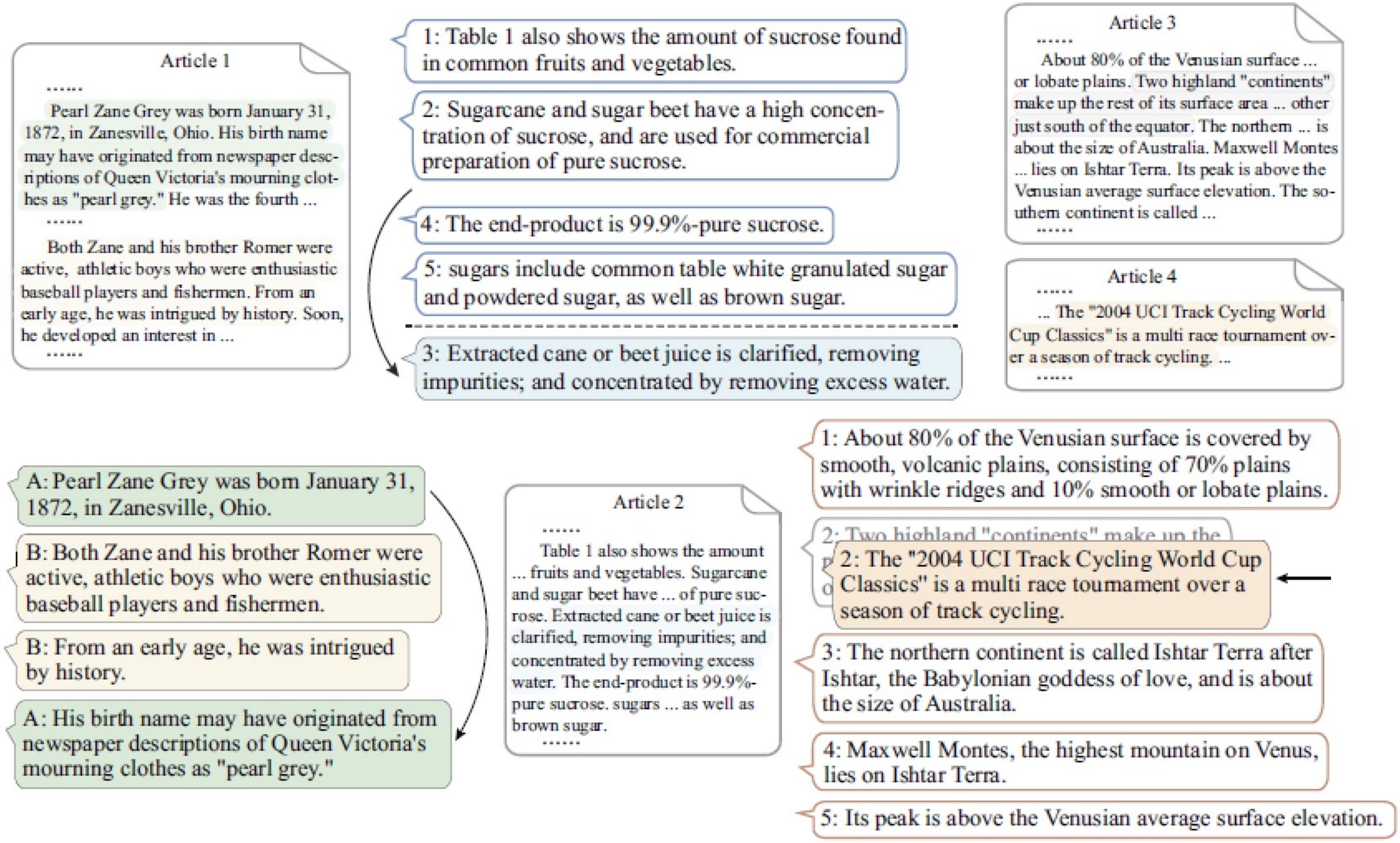
Dialogue-oriented Pre-training

- Yi Xu and Hai Zhao. 2021. Dialogue-oriented Pre-training. Findings of ACL: ACL-2021.

- Simulate the conversation features on general plain text to learn dialogue related features including speaker awareness, continuity and consistency:
  - Insertion： insert a sentence from another document
  - Deletion： delete a sentence in a document
  - Replacement： replace with a sentence from another document

Dialogue-oriented Pre-training

- Example：
  - Insertion
  - Deletion
  - Replacement



(a) Insertion　(b) Deletion　(c) Replacement

Dialogue-oriented Pre-training: Performance

- Results on benchmark datasets

	Model	E-commerce			Douban						Ubuntu		
		$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
	DialoGPT	-	-	-	-	-	-	-	-	-	79.0	88.5	97.1
	TOD-BERT	-	-	-	-	-	-	-	-	-	79.7	89.0	97.4
	BERT-[CLS]	62.7	82.2	96.2	58.7	62.7	45.1	27.6	45.8	82.7	81.9	90.4	97.8
◇	BERT-[SEP]	65.1	84.8	97.4	59.5	63.9	46.0	27.7	46.9	84.3	82.1	90.5	97.8
◇	Dialog-BERT	**66.2**	**85.5**	**97.6**	**60.0**	**64.1**	**46.9**	**28.9**	46.7	83.3	**82.3**	**90.6**	97.7
♣	BERT+multi-task	65.8	84.6	97.6	60.2	64.7	46.9	28.5	48.6	82.5	85.0	92.5	98.3
♣	Dialog-BERT+multi-task	**68.0**	**85.3**	**97.7**	**60.9**	**64.9**	**48.0**	**30.0**	47.9	**82.9**	**85.4**	**92.8**	**98.5**
	ELECTRA-[CLS]	58.2	79.6	96.9	59.0	63.2	44.8	27.6	47.3	82.8	82.5	90.7	97.8
♡	ELECTRA-[SEP]	60.4	80.6	96.3	58.8	62.5	44.2	26.9	46.3	84.1	82.2	90.7	97.8
♡	Dialog-ELECTRA	**61.1**	**81.4**	**96.9**	**59.8**	**64.1**	**46.5**	**28.3**	**47.7**	**84.1**	**83.5**	**91.4**	**98.0**
♠	ELECTRA+multi-task	68.1	86.8	97.9	61.4	65.3	47.5	29.6	50.6	83.8	86.6	93.4	98.5
♠	Dialog-ELECTRA+multi-task	**68.3**	86.3	**98.0**	**61.6**	**65.6**	**48.3**	**30.0**	49.8	**84.7**	**86.8**	**93.6**	**98.6**

Tokenization and masked units

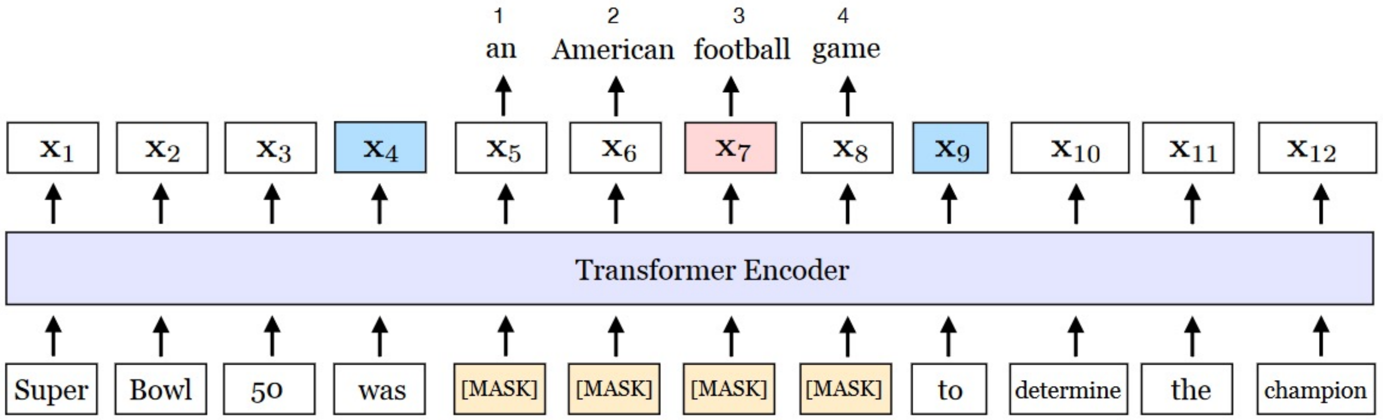
- Embedding units
  - character √ ELMo
  - subword √ BERT …
  - word ×

- Masked Units
  - Subword
  - Word/Span/Sentence
  - Knowledge pieces
  - Statistically meaningful units

BERT$_{\text{WWM}}$ vs. SpanBERT

- BERT$_{\text{WWM}}$: whole word masking
- SpanBERT
  - Mask continues spans
  - Span boundary objective

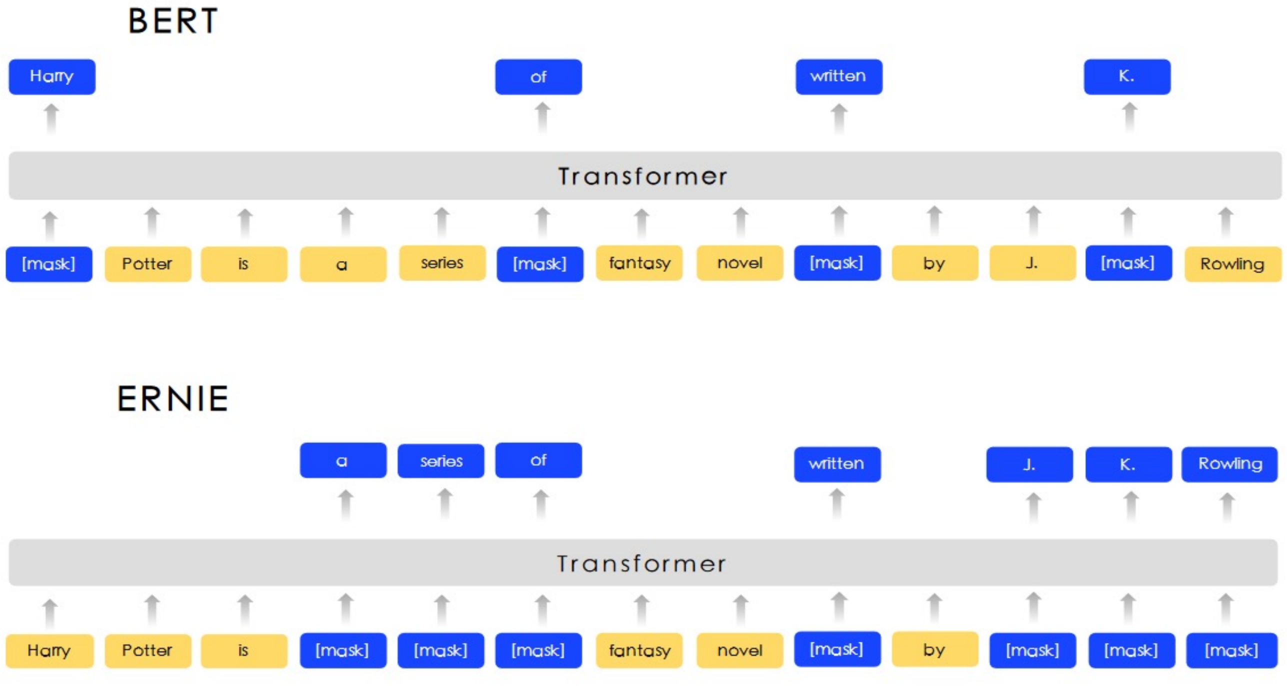
$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$



Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. TACL.

Masking Knowledge Units：ERNIE

- Knowledge-enhanced masking: **entities** + **phrases**



Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu. ERNIE: Enhanced Representation through Knowledge Integration. ACL 2020.

Masking Statistically Meaningful Units: BURT

- Construct the embedded representation in the same dimension for words, sentences and phrases
- Towards Universal Language Representation (ULR)
  - Calculate scores for all the n-grams according to point mutual information (PMI)
  - Only mask high-scored n-gram
  - MiSAD Objective

$$PMI(w) = \frac{1}{|w|}\left(\log P(w) - \sum_{k=1}^{|w|} \log P(x_k)\right) \qquad w = (x_1, \ldots, x_{|w|}) \qquad score_w = \frac{1}{|w|}\sum_{k=1}^{|w|} P(x_k | S^{\backslash w})$$

[1] Yian Li and Hai Zhao. 2021. Pre-training Universal Language Representation, ACL-2021.
[2] Yian Li and Hai Zhao. 2021. BURT: BERT-inspired Universal Representation from Learning Meaningful SegmenT, on TPAMI review

MiSAD

① *"London is"* + *"the capital of England"* = *"London is the capital of England"*

② vector(*"London is"*) + vector(*"the capital of England"*)

= vector(*"London is the capital of England"*)

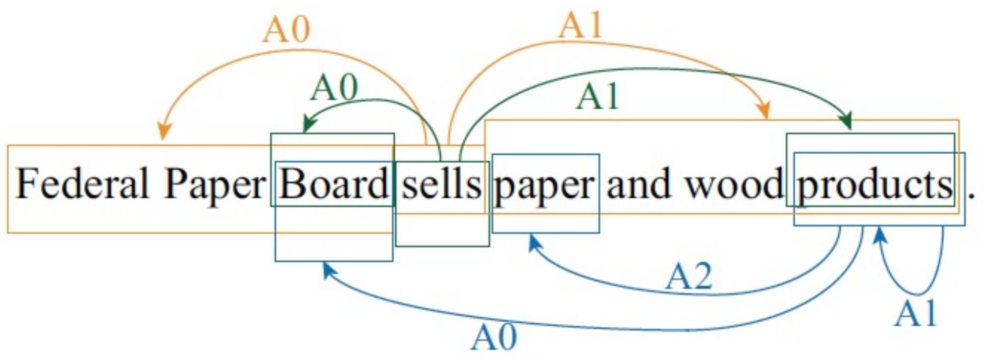
- Input sentence: S = $\{x_1, \ldots, x_m\}$ = $\{x_1, \ldots, x_{i-1}, w, x_{j+1}, \ldots, x_m\}$
- Extracted *n*-gram: w = $\{x_i, \ldots, x_j\}$, $1 \leq i < j \leq$ m
- The remaining tokens : R = $\{x_1, \ldots, x_{i-1}, x_{j+1}, \ldots, x_m\}$

$$\mathcal{L}_{MiSAD} = MSE(E^w + E^R, E^S)$$

[1] Yian Li and Hai Zhao. 2021. Pre-training Universal Language Representation, ACL-2021.
[2] Yian Li and Hai Zhao. 2021. BURT: BERT-inspired Universal Representation from Learning Meaningful SegmenT, on TPAMI review

Linguistic Mask：LIMIT-BERT

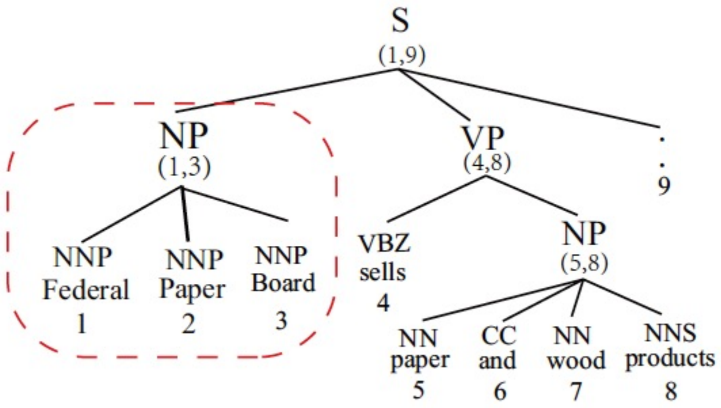
- Mask Strategy: syntactic and semantic masks
- Multitask Learning： improve the modeling performance of language model with linguistic tasks.



Span and Dependency SRL

federal paper board [MASK] paper and wood [MASK] .

(a) Semantic Phrase Masking.



Constituent Syntactic Tree

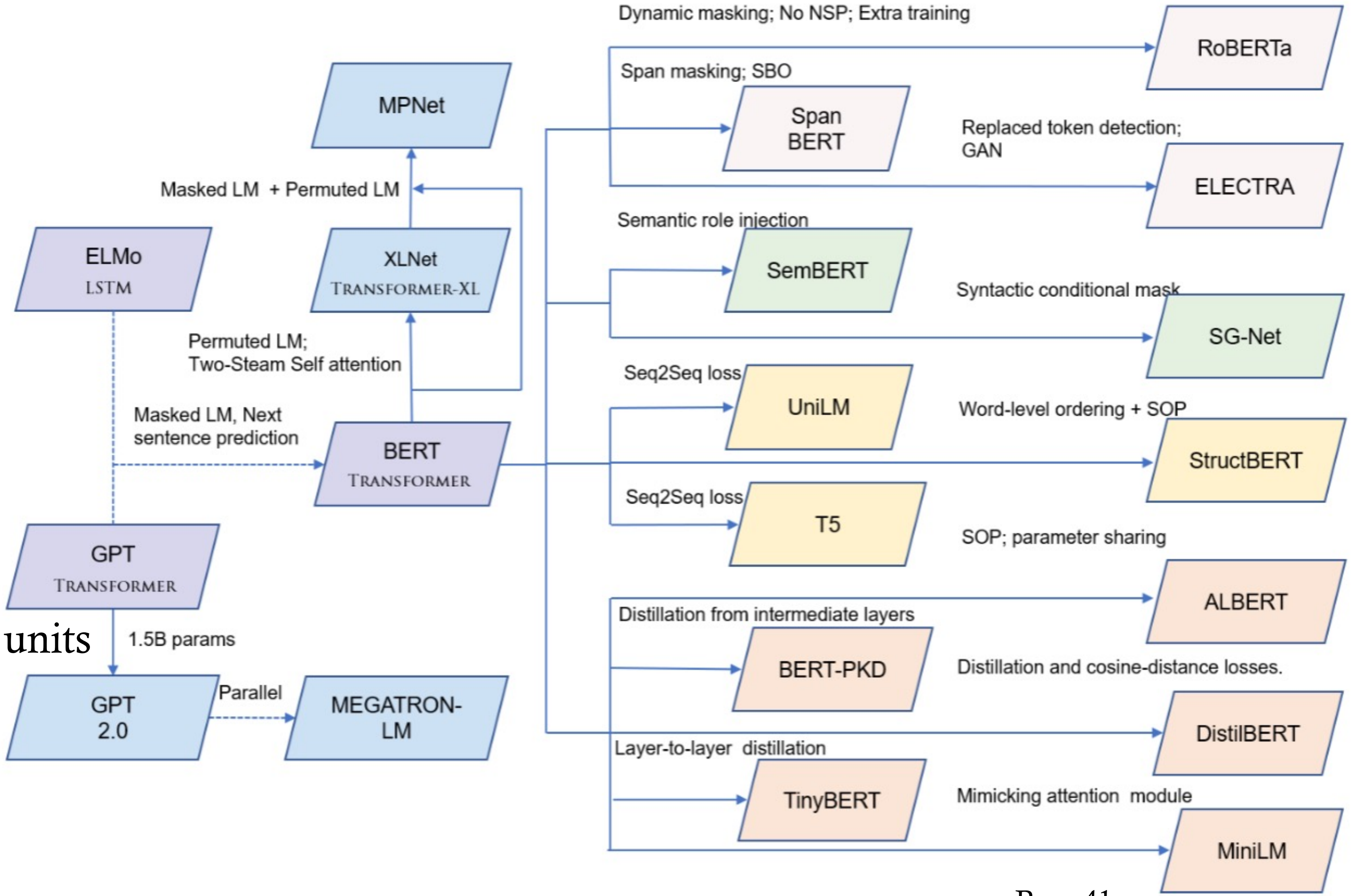
[MASK] [MASK] [MASK] sells paper and wood products .

(b) Syntactic Phrase Masking.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. LIMIT-BERT : Linguistics Informed Multi-Task BERT. EMNLP 2020. ACL Findings.

Derivative of PrLM

- Embedding Units
  - Character
  - Subword
  - Word
- Masked Units
  - Subword
  - Word/Span/Sentence
  - Knowledge pieces
  - Statistically meaningful units
- Sequence Prediction
  - Sentence relevance
  - Sentence order



Performance of PrLM derivatives

Method	SQuAD1.1 Dev	SQuAD1.1 ↑ Dev	SQuAD1.1 Test	SQuAD1.1 ↑ Test	SQuAD2.0 Dev	SQuAD2.0 ↑ Dev	SQuAD2.0 Test	SQuAD2.0 ↑ Test	RACE Acc	RACE ↑ Acc
ELMo	*85.6*	-	*85.8*	-	-		-		-	-
GPT_{v1}	-	-	-	-	-	-	-	-	*59.0*	-
BERT_{base}	88.5	2.9	-	-	*76.8*		-		65.3	6.3
BERT-PKD	85.3	-0.3	-	-	69.8	-7.0	-	-	60.3	1.3
DistilBERT	86.2	0.6	-	-	69.5	-7.3	-	-		
TinyBERT	87.5	1.9	-	-	73.4	-3.4	-	-	-	-
MiniLM	-	-	-	-	76.4	-0.4	-	-	-	-
Q-BERT	88.4	2.8	-	-	-	-	-	-	-	-
BERT_{large}	91.1*	5.5	91.8*	6	81.9	5.1	83.0	-	72.0†	-
SemBERT_{large}	-	-	-	-	83.6	6.8	85.2	2.2	-	-
SG-Net	-	-	-	-	88.3	11.5	87.9	4.9	74.2	15.2
SpanBERT_{large}	-	-	94.6	8.8	-	-	88.7	5.7	-	-
$\text{StructBERT}_{large}$	92.0	6.4	-	-	-	-	-	-	-	-
RoBERTa_{large}	94.6	9.0	-	-	89.4	12.6	89.8	6.8	83.2	24.2
$\text{ALBERT}_{xxlarge}$	94.8	9.2	-	-	90.2	13.4	90.9	7.9	86.5	27.5
XLNet_{large}	94.5	8.9	95.1*	9.3	88.8	12	89.1*	6.1	81.8	22.8
UniLM	-	-	-	-	83.4	6.6	-	-	-	-
ELECTRA_{large}	94.9	9.3	-	-	90.6	13.8	91.4	8.4	-	-
$\text{Megatron-LM}_{3.9B}$	95.5	9.9	-	-	91.2	14.4	-	-	89.5	30.5
T5_{11B}	95.6	10.0	-	-	-	-	-	-	-	-

Correlations Between MRC and PrLM

MRC and PrLM are **complementary** to each other.

MRC serves as an appropriate testbed for language representation, which is the focus of PrLMs.

The progress of PrLMs greatly promotes MRC tasks, achieving impressive gains of model performance.

The initial applications of PrLMs. The concerned NLU task can also be regarded as a special case of MRC

	NLU		MRC		
	SNLI	GLUE	SQuAD1.1	SQuAD2.0	RACE
ELMo	✓	✗	✓	✗	✗
GPT$_{v1}$	✓	✓	✗	✗	✓
BERT	✗	✓	✓	✓	✗
RoBERTa	✗	✓	✓	✓	✓
ALBERT	✗	✓	✓	✓	✓
XLNet	✗	✓	✓	✓	✓
ELECTRA	✗	✓	✓	✓	✗

Machine Reading Comprehension:

Technical Methods, Discussions, and Frontiers

Zhuosheng Zhang
zhangzs@sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhangzs

Two-stage Solving Architecture

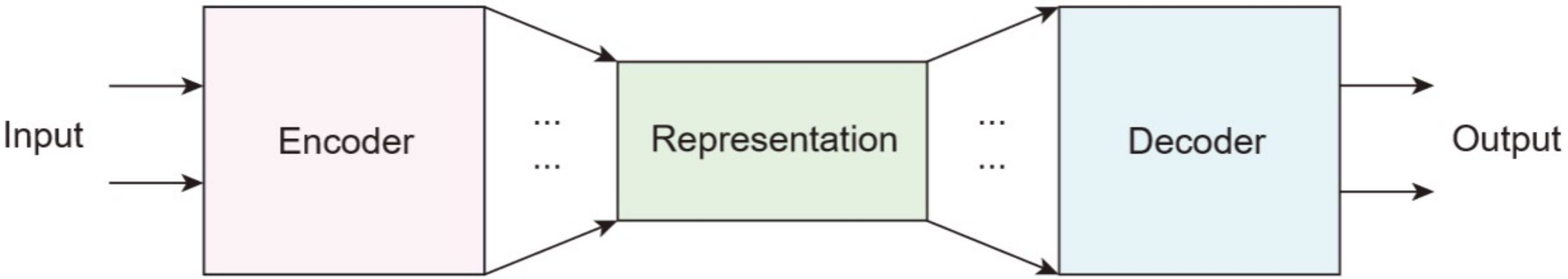
Inspired by Dual process theory of cognition psychology:

The cognitive process of human brains potentially involves two distinct types of procedures:

- **contextualized perception** (reading): gather information in an implicit process
- **analytic cognition** (comprehension): conduct the controlled reasoning and execute goals

Standard MRC system:

- building a PrLM as **Encoder**;
- designing ingenious mechanisms as **Decoder** according to task characteristics.



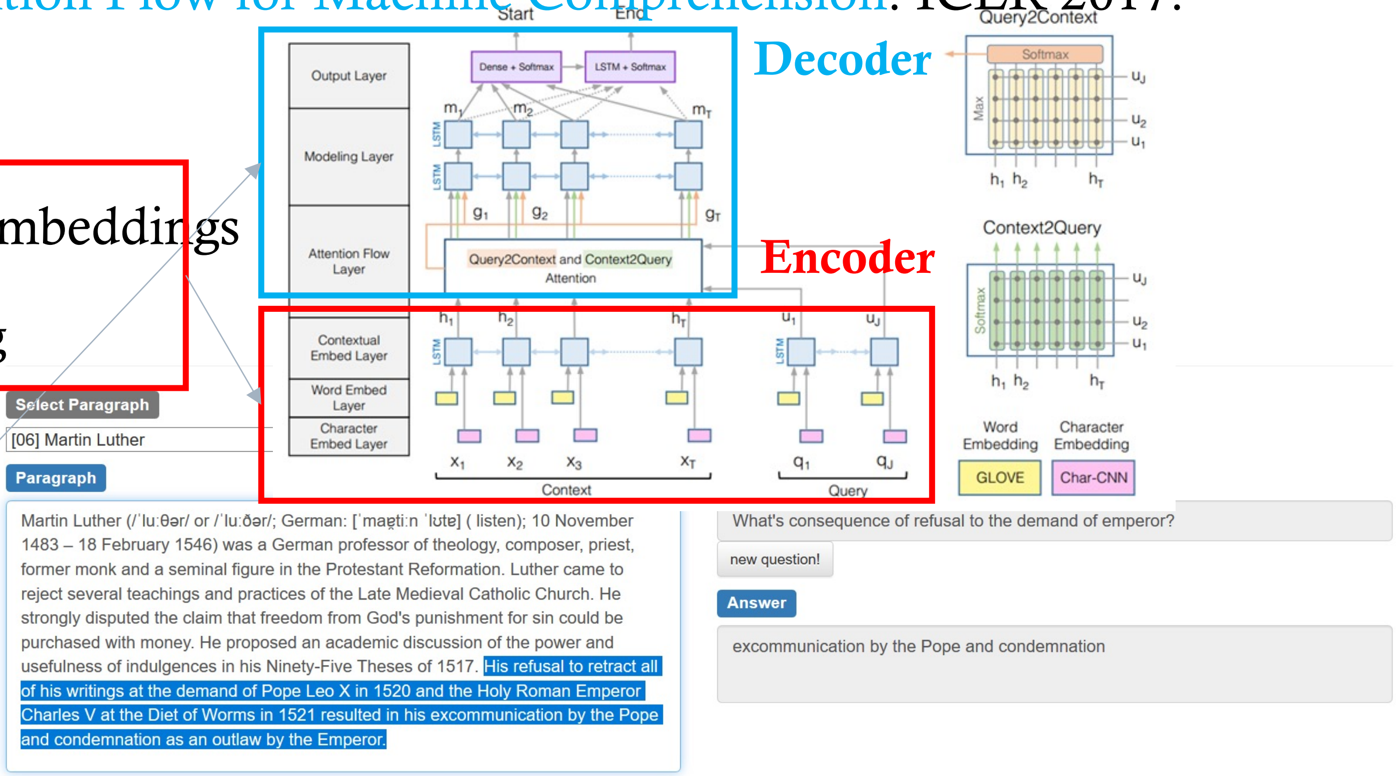
Typical MRC Architecture

☐ BiDAF

- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. ICLR 2017.

Hierarchical structure:

- Word + Char level embeddings
- Contextual encoding
- Attention modules
- Answer prediction



☐ Pre-trained PrLMs for Fine-tuning

Encoder: PrLM; **Decoder**: special modules for span prediction, answer verification, counting, reasoning.

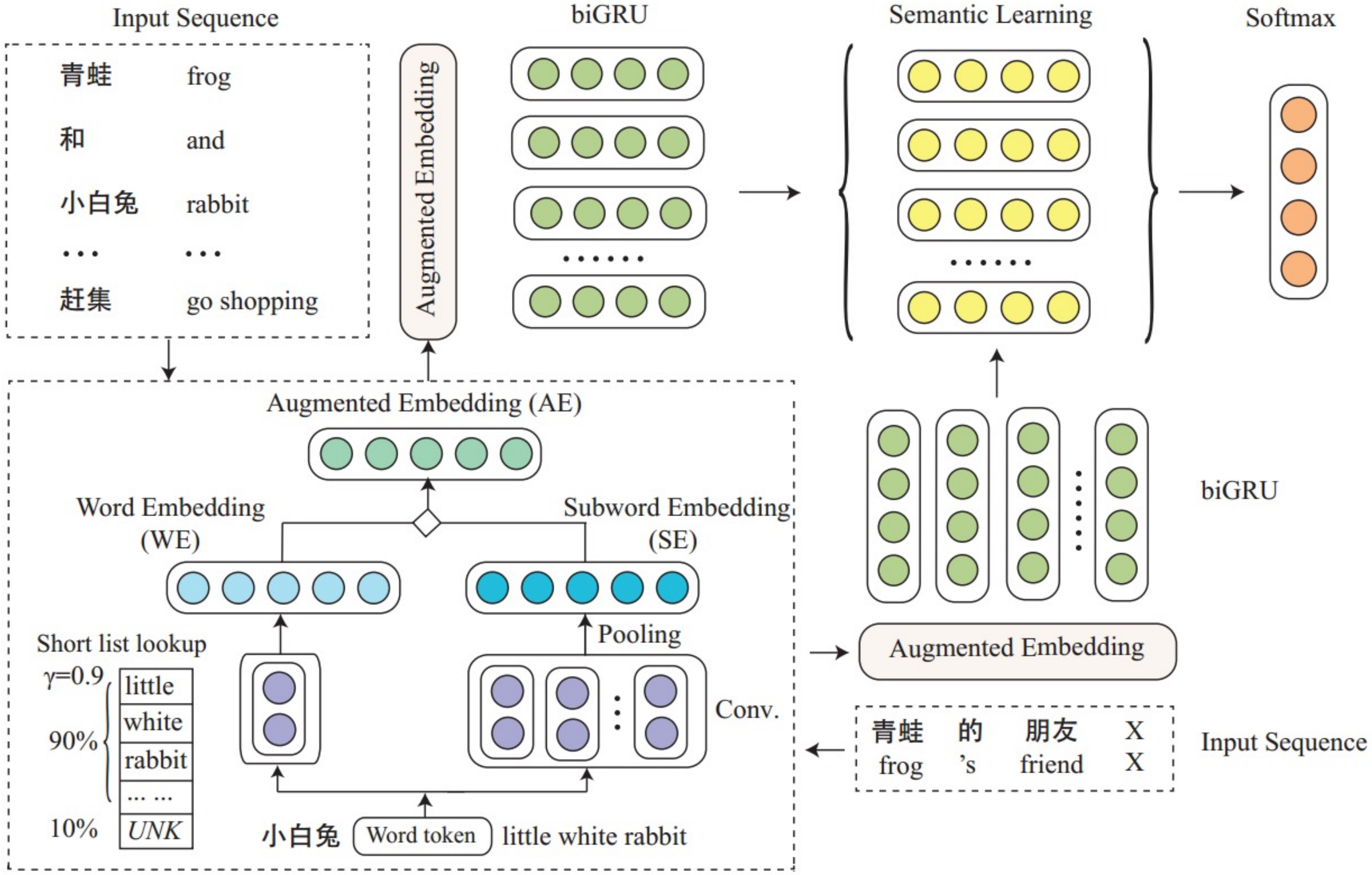
Encoder

- **Multiple Granularity Features**
  - Language Units: word, character, subword.
  - Salient Features: Linguistic features, such as part-of-speech, named entity tags, semantic role labeling tags, syntactic features, and binary Exact Match features.
- **Structured Knowledge Injection (Transformer/GNN)**
  - Linguistic Structures
  - Commonsense
- **Contextualized Sentence Representation**
  - Embedding pretraining

Encoder (our work: language units)

SubMRC: Subword-augmented Embedding

Zhuosheng Zhang, Yafang Huang, Hai Zhao. 2018. Subword-augmented Embedding for Cloze Reading Comprehension. COLING 2018



- Gold answers are often rare words.
- Error analysis shows that early MRC models suffer from out-of-vocabulary (OOV) issues.

We propose:

- Subword-level representation
- Frequency-based short list filtering

We investigate many **subword segmentation algorithms** and propose a unified framework composed of goodness measure and segmentation:

Zhuosheng Zhang, Hai Zhao, Kangwei Ling, Jiangtong Li, Shexia He, Guohong Fu (2019). Effective Subword Segmentation for Text Comprehension. IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP).

Encoder (our work: language units)

SubMRC: Subword-augmented Embedding

Zhuosheng Zhang, Yafang Huang, Hai Zhao. 2018. Subword-augmented Embedding for Cloze Reading Comprehension. COLING 2018

最佳单系统（Best Single System）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	上海交通大学仿脑计算与机器智能研究中心自然语言组 Shanghai Jiao Tong University (SJTU BCMI-NLP)	单系统	76.15%	77.73%

最终系统排名

填空类问题（Cloze-style Question）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	6ESTATES PTE LTD	多系统	81.85%	81.90%
		单系统	75.85%	74.73%
2	上海交通大学仿脑计算与机器智能研究中心自然语言组 Shanghai Jiao Tong University (SJTU BCMI-NLP)	多系统	78.35%	80.67%
		单系统	76.15%	77.73%
3	南京云思创智信息科技有限公司	多系统	79.20%	80.27%
		单系统	77.15%	77.53%
4	华东师范大学 East China Normal University (ECNU)	多系统	79.45%	79.70%
		单系统	77.95%	77.40%
5	鲁东大学 Ludong University	多系统	77.05%	77.07%
		单系统	74.75%	75.07%
6	武汉大学语言与信息研究中心 Wuhan University (WHU)	单系统	78.20%	76.53%

Best single model in CMRC 2017 shared task

Model	CMRC-2017 Valid	CMRC-2017 Test
Random Guess †	1.65	1.67
Top Frequency †	14.85	14.07
AS Reader †	69.75	71.23
GA Reader	72.90	74.10
SJTU BCMI-NLP †	76.15	77.73
6ESTATES PTE LTD †	75.85	74.73
Xinktech †	77.15	77.53
Ludong University †	74.75	75.07
ECNU †	77.95	77.40
WHU †	78.20	76.53
SAW Reader	**78.95**	**78.80**

Model	PD Valid	PD Test	CFT Test-human
AS Reader	64.1	67.2	33.1
GA Reader	67.2	69.0	36.9
CAS Reader	65.2	68.1	35.0
SAW Reader	**72.8**	**75.1**	**43.8**

Model	CBT-NE Valid	CBT-NE Test	CBT-CN Valid	CBT-CN Test
Human ‡	-	81.6	-	81.6
LSTMs ‡	51.2	41.8	62.6	56.0
MemNets ‡	70.4	66.6	64.2	63.0
AS Reader ‡	73.8	68.6	68.8	63.4
Iterative Attentive Reader ‡	75.2	68.2	72.1	69.2
EpiReader ‡	75.3	69.7	71.5	67.4
AoA Reader ‡	77.8	72.0	72.2	69.4
NSE ‡	78.2	73.2	74.3	71.9
FG Reader ‡	**79.1**	**75.0**	**75.3**	**72.0**
GA Reader ‡	76.8	72.5	73.1	69.6
SAW Reader	78.5	74.9	75.0	71.6

Encoder (our work: salient features)

SemBERT: Semantics-aware BERT

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, Xiang Zhou. 2020. Semantics-aware BERT for Language Understanding. AAAI-2020.

Passage

- *...Harvard was a founding member of the Association of American Universities in 1900. James Bryant Conant led the university through the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977merger with Radcliffe College.......*

Question

- *What was the name of the leader through the Great Depression and World War II?*

Semantic Role Labeling (SRL)

- *[James Bryant Conant]*$_{\mathrm{ARG0}}$ *[led]*$_{\mathrm{VERB}}$ *[the university]*$_{\mathrm{ARG1}}$ *through [the Great Depression and World War II]*$_{\mathrm{ARG2}}$

Answer

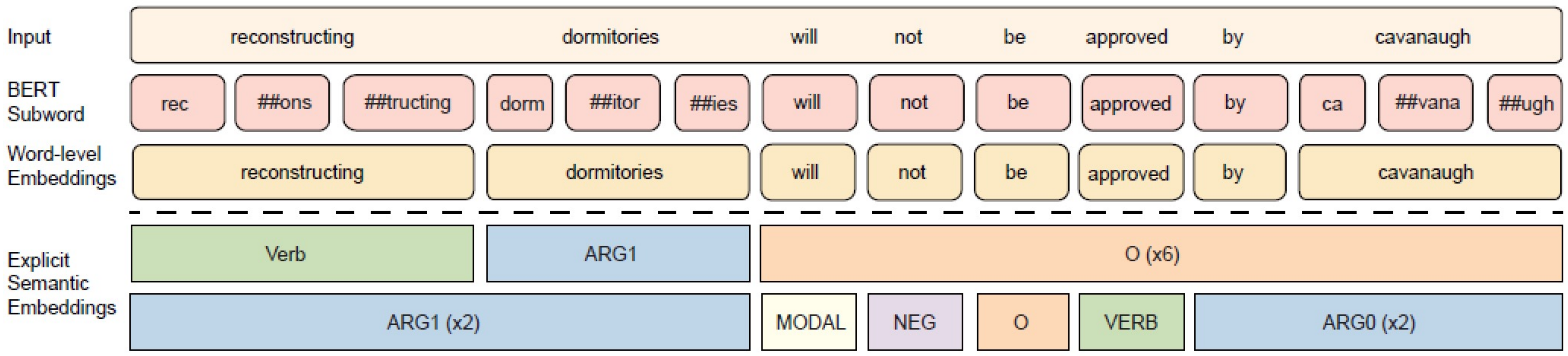
- *James Bryant Conant*

Problem: Who did what to whom, when and why?

Encoder (our work: salient features)

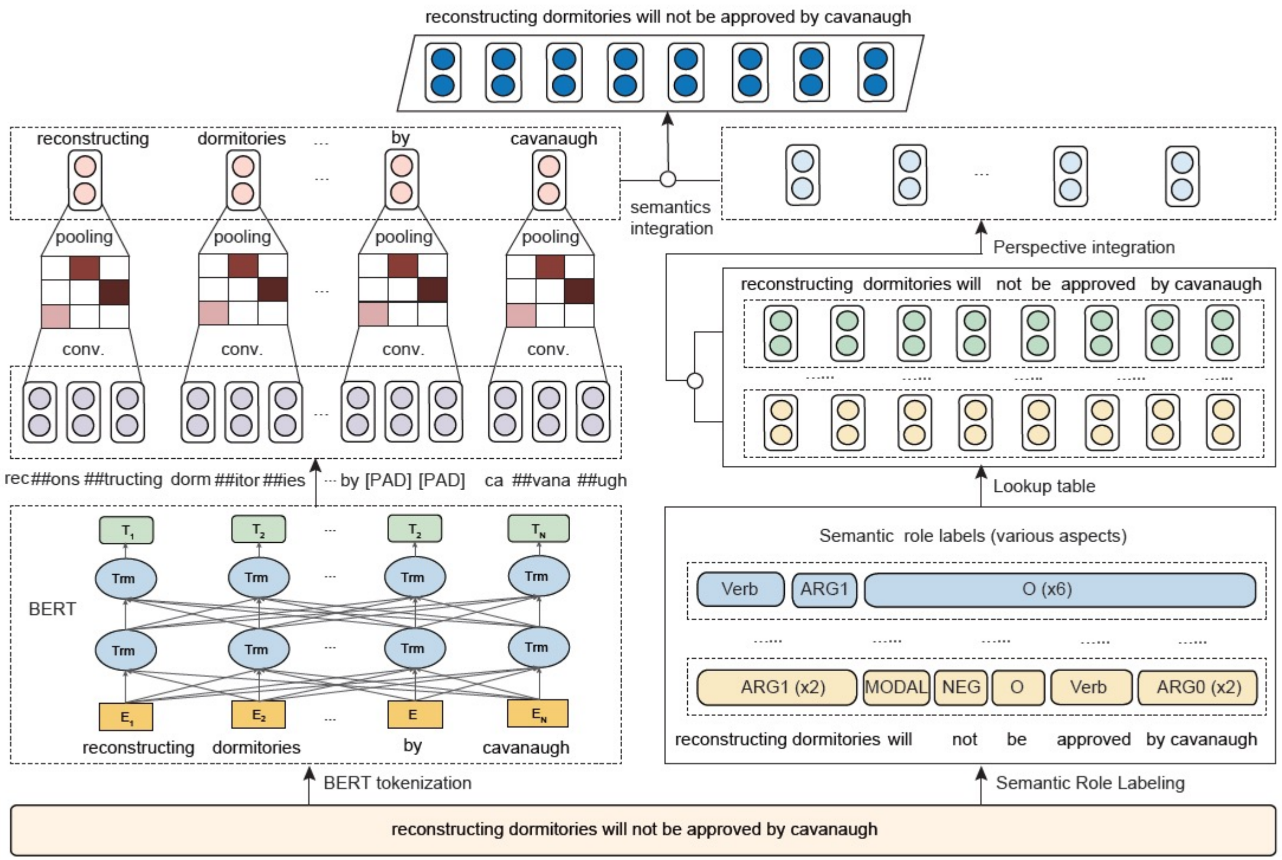
SemBERT: Semantics-aware BERT

- ELMo & BERT: only take Plain contextual features
- SemBERT: introduce Explicit contextual Semantics, Deeper representation?
  - Semantic Role Labeler + BERT encoder



Encoder (our work: salient features)

SemBERT: Semantics-aware



Method	Classification		Natural Language Inference			Semantic Similarity			Score
	CoLA (mc)	SST-2 (acc)	MNLI m/mm(acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)	STS-B (pc)	-
Leaderboard (September, 2019)									
ALBERT	69.1	97.1	91.3/91.0	99.2	89.2	93.4	74.2	92.5	89.4
RoBERTa	67.8	96.7	90.8/90.2	98.9	88.2	92.1	90.2	92.2	88.5
XLNET	67.8	96.8	90.2/89.8	98.6	86.3	93.0	90.3	91.6	88.4
In literature (April, 2019)									
BiLSTM+ELMo+Attn	36.0	90.4	76.4/76.1	79.9	56.8	84.9	64.8	75.1	70.5
GPT	45.4	91.3	82.1/81.4	88.1	56.0	82.3	70.3	82.0	72.8
GPT on STILTs	47.2	93.1	80.8/80.6	87.2	69.1	87.7	70.1	85.3	76.9
MT-DNN	61.5	95.6	86.7/86.0	-	75.5	90.0	72.4	88.3	82.2
$BERT_{BASE}$	52.1	93.5	84.6/83.4	-	66.4	88.9	71.2	87.1	78.3
$BERT_{LARGE}$	60.5	94.9	86.7/85.9	92.7	70.1	89.3	72.1	87.6	80.5
Our implementation									
$SemBERT_{BASE}$	57.8	93.5	84.4/84.0	90.9	69.3	88.2	71.8	87.3	80.9
$SemBERT_{LARGE}$	62.3	94.6	87.6/86.3	94.6	84.5	91.2	72.8	87.8	82.9

GLUE 实验结果

Model	EM	F1
#1 BERT + DAE + AoA†	85.9	88.6
#2 SG-Net†	85.2	87.9
#3 BERT + NGM + SST†	85.2	87.7
U-Net (Sun et al. 2018)	69.2	72.6
RMR + ELMo + Verifier (Hu et al. 2018)	71.7	74.2
Our implementation		
$BERT_{LARGE}$	80.5	83.6
$SemBERT_{LARGE}$	82.4	85.2
$SemBERT^{*}_{LARGE}$	84.8	87.9

SQuAD 实验结果

Model	Dev	Test
In literature		
DRCN (Kim et al. 2018)	-	90.1
SJRC (Zhang et al. 2019)	-	91.3
MT-DNN (Liu et al. 2019)†	92.2	91.6
Our implementation		
$BERT_{BASE}$	90.8	90.7
$BERT_{LARGE}$	91.3	91.1
$SemBERT_{BASE}$	91.2	91.0
$SemBERT_{LARGE}$	92.3	91.6

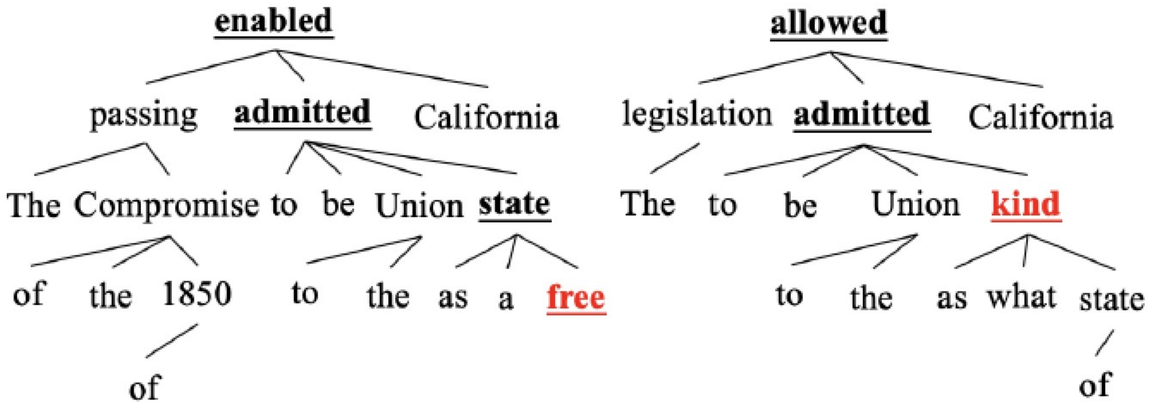
SNLI 实验结果

SNLI: The **best** among all submissions.
https://nlp.stanford.edu/projects/snli/
SQuAD2.0: The **best** among all the published work.
GLUE: substantial gains over all the tasks.

Encoder (our work: linguistic structures)

SG-Net: Syntax-guided Network

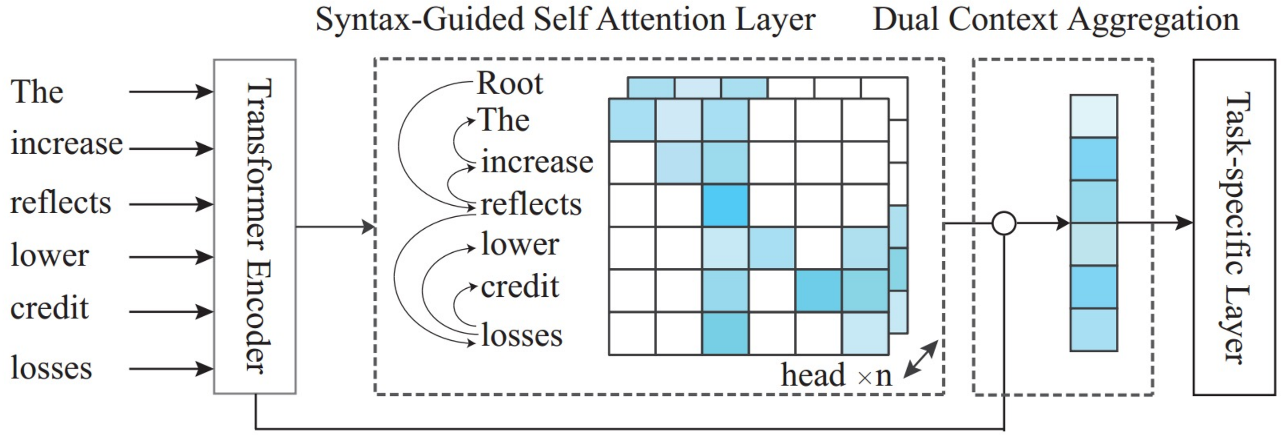
- Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, Rui Wang. 2020. Syntax-Guided Machine Reading Comprehension.AAAI-2020.
- Passage
  - *The passing of the Compromise of 1850 enabled California to be admitted to the Union as a free state, preventing southern California from becoming its own separate slave state …*
- Question:
  - *The legislation allowed California to be admitted to the Union as what kind of state?*
- Answer:
  - *free*

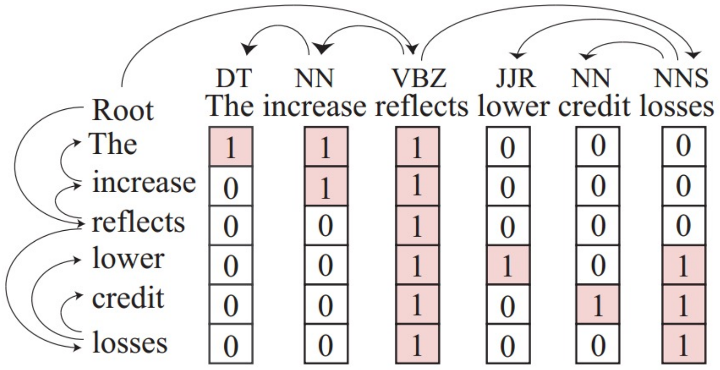


Encoder (our work: linguistic structures)

SG-Net: Syntax-guided Network

- Self-attention network (SAN) empowered Transformer-based encoder
- Syntax-guided self-attention network (SAN)
  - Syntactic dependency of interest (SDOI): regarding each word as a child node
  - SDOI consists all its ancestor nodes and itself in the dependency parsing tree
  - *Pi* : ancestor node set for each i_{th} word; M : SDOI mask $\mathcal{M}[i,j] = \begin{cases} 1, & \text{if } j \in P_i \text{ or } j = i \\ 0, & \text{otherwise.} \end{cases}$





Parser: Junru Zhou, Hai Zhao. 2019. Head-driven Phrase Structure Grammar Parsing on Penn Treebank. ACL 2019, pp.2396–2408.

Encoder (our work: linguistic structures)

SG-Net: Syntax-guided Network

- ☐ Our single model (XLNet + SG-Net Verifier) ranks **first**.
- ☐ The **first single model** to exceed **human performance**.

Model	Dev EM	Dev F1	Test EM	Test F1
Regular Track				
Joint SAN	69.3	72.2	68.7	71.4
U-Net	70.3	74.0	69.2	72.6
RMR + ELMo + Verifier	72.3	74.8	71.7	74.2
BERT Track				
Human	-	-	86.8	89.5
BERT + DAE + AoA†	-	-	85.9	88.6
BERT + NGM + SST†	-	-	85.2	87.7
BERT + CLSTM + MTL + V†	-	-	84.9	88.2
SemBERT†	-	-	84.8	87.9
Insight-baseline-BERT†	-	-	84.8	87.6
BERT + MMFT + ADA†	-	-	83.0	85.9
$\mathrm{BERT}_{\mathrm{LARGE}}$	-	-	82.1	84.8
Baseline	84.1	86.8	-	-
SG-Net	**85.1**	**87.9**	-	-
+Verifier	**85.6**	**88.3**	**85.2**	**87.9**

Model	RACE-M	RACE-H	RACE
Human Performance			
Turkers	85.1	69.4	73.3
Ceiling	95.4	94.2	94.5
Leaderboard			
DCMN	77.6	70.1	72.3
$\mathrm{BERT}_{\mathrm{LARGE}}$	76.6	70.1	72.0
OCN	76.7	69.6	71.7
Baseline	78.4	70.4	72.6
SG-Net	**78.8**	**72.2**	**74.2**

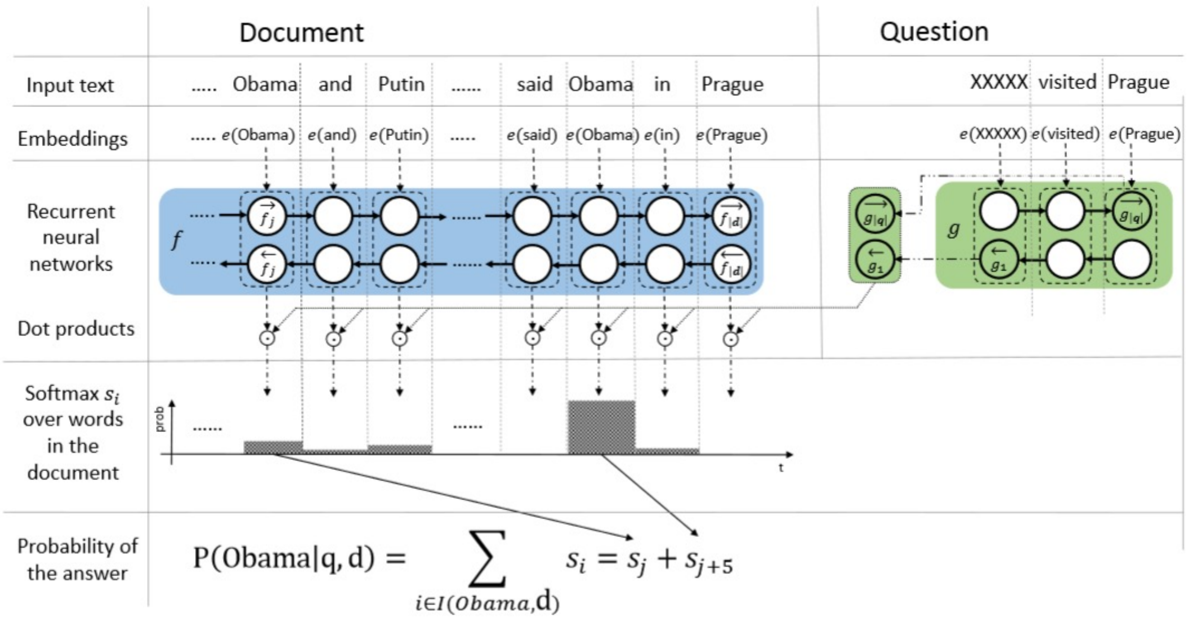
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**88.592**	**90.859**
2 Jul 19, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk*	88.050	90.645
3 Jul 19, 2019	XLNet + SG-Net Verifier (single model) *Shanghai Jiao Tong University & CloudWalk*	87.035	89.897
3 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
3 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795
4 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language*	86.673	89.147
6 May 21, 2019	XLNet (single model) *Google Brain & CMU*	86.346	89.133
7 May 14, 2019	SG-Net (ensemble) *Shanghai Jiao Tong University*	86.211	88.848
7 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886
8	BERT + DAE + AoA (single model)	85.884	88.621

Decoder

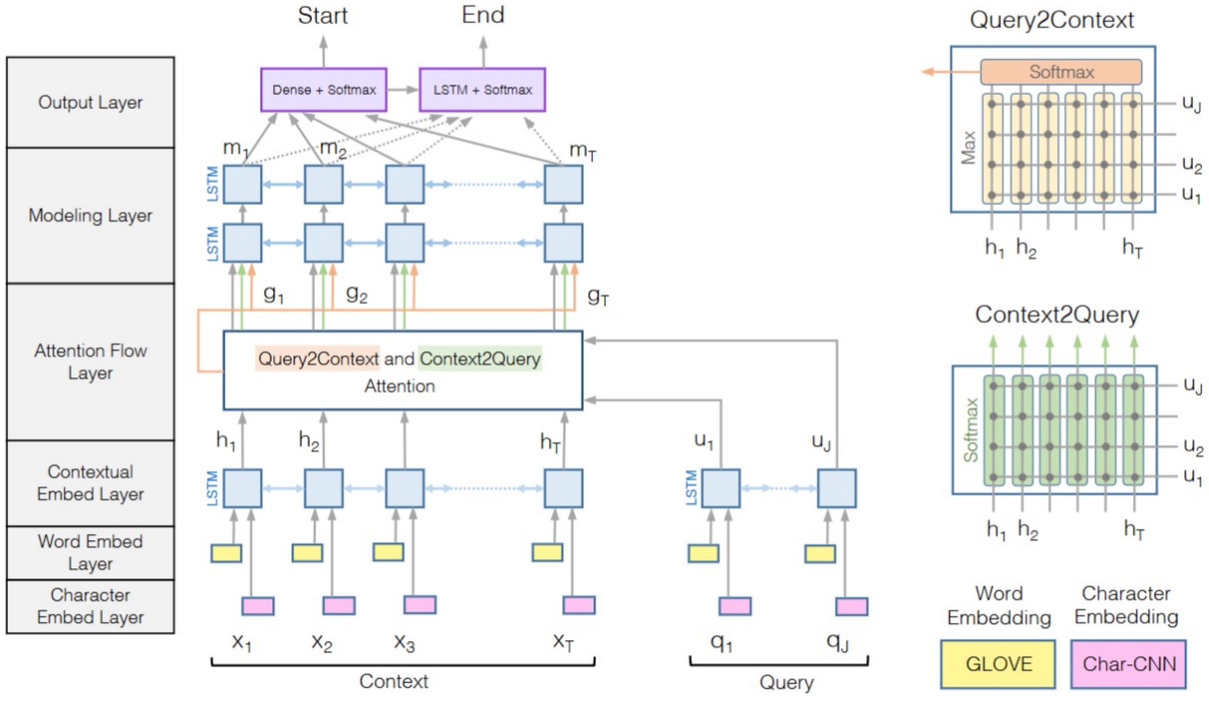
- Matching Network:
  - Attention Sum, Gated Attention, Self-matching, Attention over Attention, Co-match Attention, Dual Co-match Attention, etc.
- Fine-grained Reasoning Network:
  - Decouple the context into multiple elements and measure the relationships for reasoning
- Answer Pointer:
  - Pointer Network for span prediction
  - Reinforcement learning based self-critical learning to predict more acceptable answers
- Answer Verifier:
  - Threshold-based answerable verification
  - Multitask-style verification
  - External parallel verification
- Answer Type Predictor for multi-type MRC tasks

Decoder

- Matching Network:
  - Attention Sum, Gated Attention, Self-matching, Attention over Attention, BiDAF, etc.
- Attention weights: sum, dot, gating, etc.
- Attention Direction: question-aware, passage aware, self-attention, bidirectional, etc.



(AS Reader)



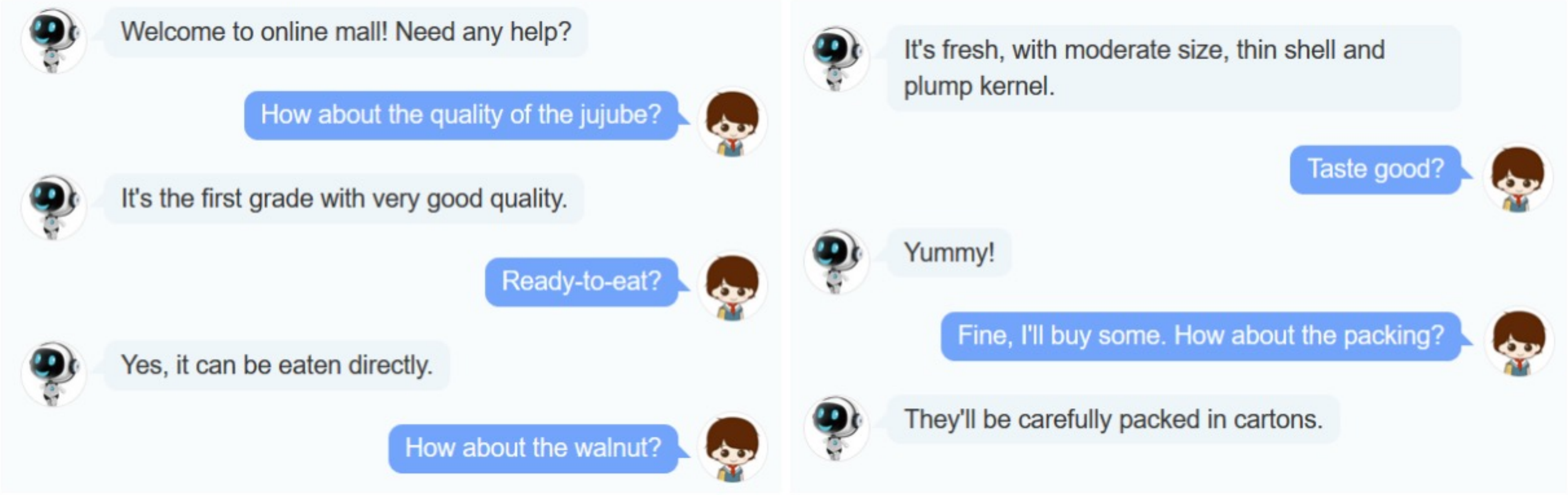
BiDAF

- Attention Granularity : word-level, sequence-level, hierarchical, etc.

Decoder (Deep Utterance Aggregation)

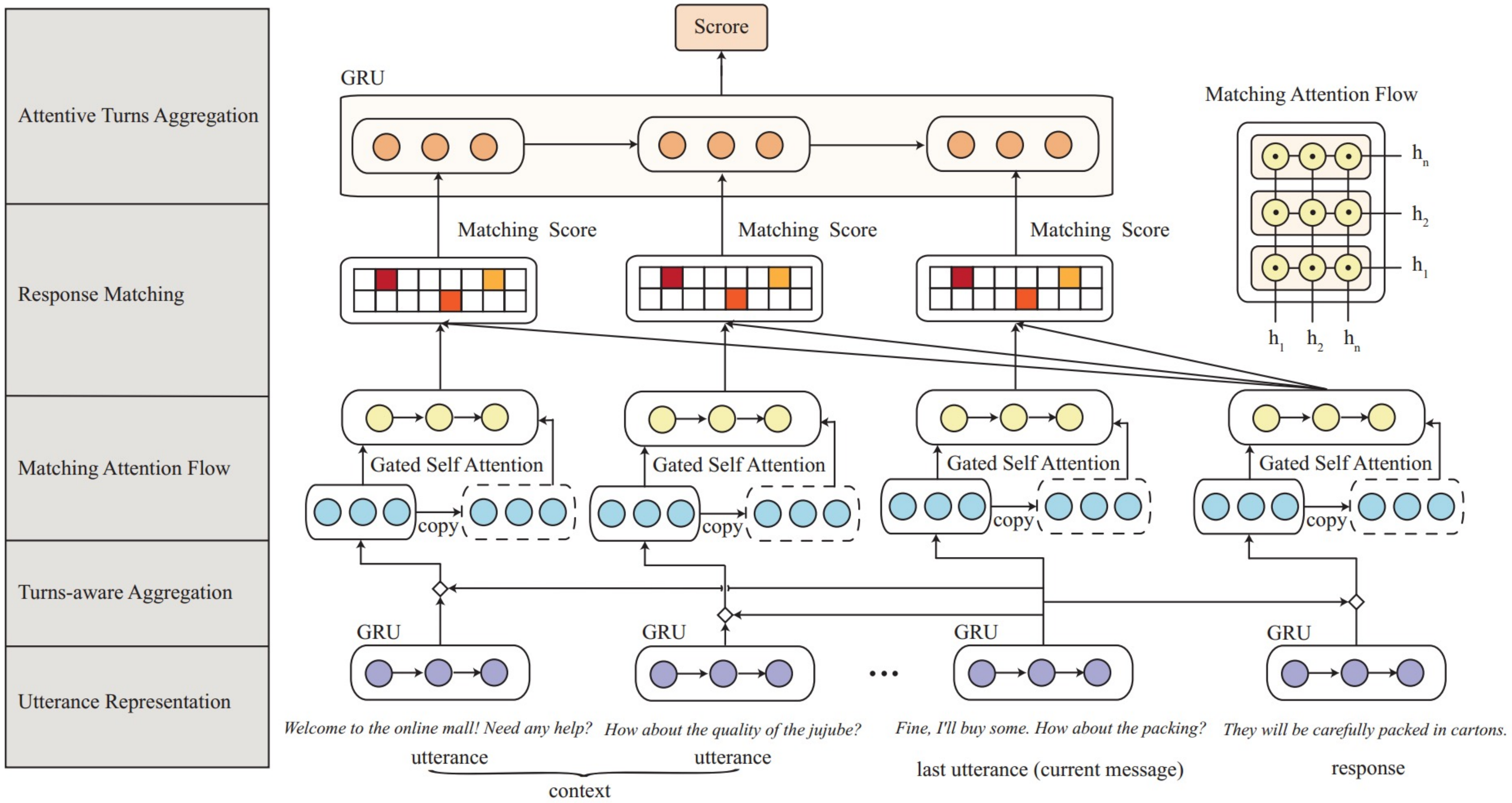
Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao and Gongshen Liu. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. COLING 2018.

- Challenge: **long utterances, multiple intentions, topic shift**, etc.
- Aim: recognize the **key information** from complex dialogue history
- Solution: deep utterance aggregation framework (**DUA**)
- Corpus: a new **E-commerce Dialogue Corpus**



Decoder (Deep Utterance Aggregation)

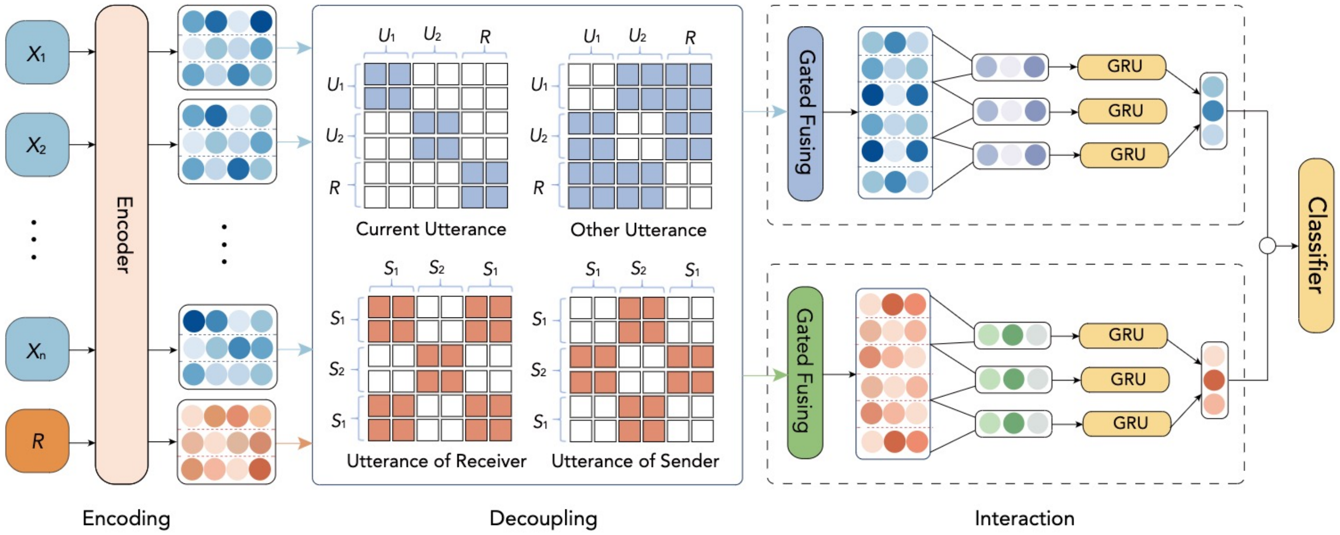
- Capture the main information in each utterance (**self attention,** first introduced)
- Model the **information flow through the utterances** in dialogue history
- Match the relationship **between utterance and candidate response**



Highlight the importance of **the last utterance**.

Decoder (MDFN)

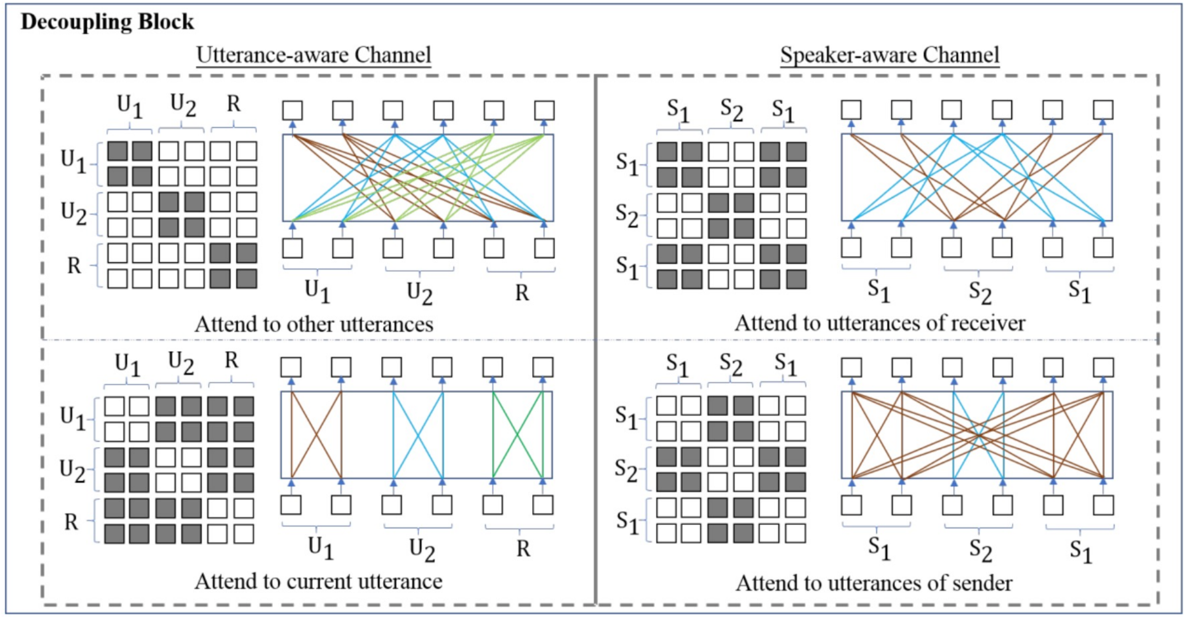
- **Challenges for Multi-turn Dialogue Comprehension:**
  - **Transition (multi-round):** speaker role transitions
  - **Inherency:** Utterances have their own inherent meaning and contextual meaning.
- Motivation
  - Different people have different speaking styles and intents.
  - The **hierarchical information** on either **utterance interrelation** or **speaker identity** is not well addressed.



Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, Xiang Zhou. 2021. Filling the Gap of Utterance-aware and Speaker-aware Representation for Multi-turn Dialogue, AAAI-2021

Decoder (MDFN)

- A novel end-to-end Mask-based Decoupling-Fusing Network (MDFN)
  - Decoupled the contextualized words representations into four parts
  - Fused the representations after sufficient interactions
- **Two Channels:** Four independent self-attention blocks with the same inputs and **different masks**
- **Focus:** current utterance, other utterances, utterances of sender and utterances of receiver



$$M_1[i,j] = \begin{cases} 0, & \text{if } \mathbb{T}_i = \mathbb{T}_j \\ -\infty, & \text{otherwise} \end{cases}$$

$$M_2[i,j] = \begin{cases} 0, & \text{if } \mathbb{T}_i \neq \mathbb{T}_j \\ -\infty, & \text{otherwise} \end{cases}$$

$$M_3[i,j] = \begin{cases} 0, & \text{if } \mathbb{S}_i = \mathbb{S}_j \\ -\infty, & \text{otherwise} \end{cases}$$

$$M_4[i,j] = \begin{cases} 0, & \text{if } \mathbb{S}_i \neq \mathbb{S}_j \\ -\infty, & \text{otherwise} \end{cases}$$

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, Xiang Zhou. 2021. Filling the Gap of Utterance-aware and Speaker-aware Representation for Multi-turn Dialogue, AAAI-2021

Decoder

- Answer Pointer:
  - Pointer Network for span prediction (start and end positions):

$$p(\mathbf{a}|\mathbf{H}^{\mathrm{r}}) = p(a_{\mathrm{s}}|\mathbf{H}^{\mathrm{r}})p(a_{\mathrm{e}}|a_{\mathrm{s}}, \mathbf{H}^{\mathrm{r}}).$$

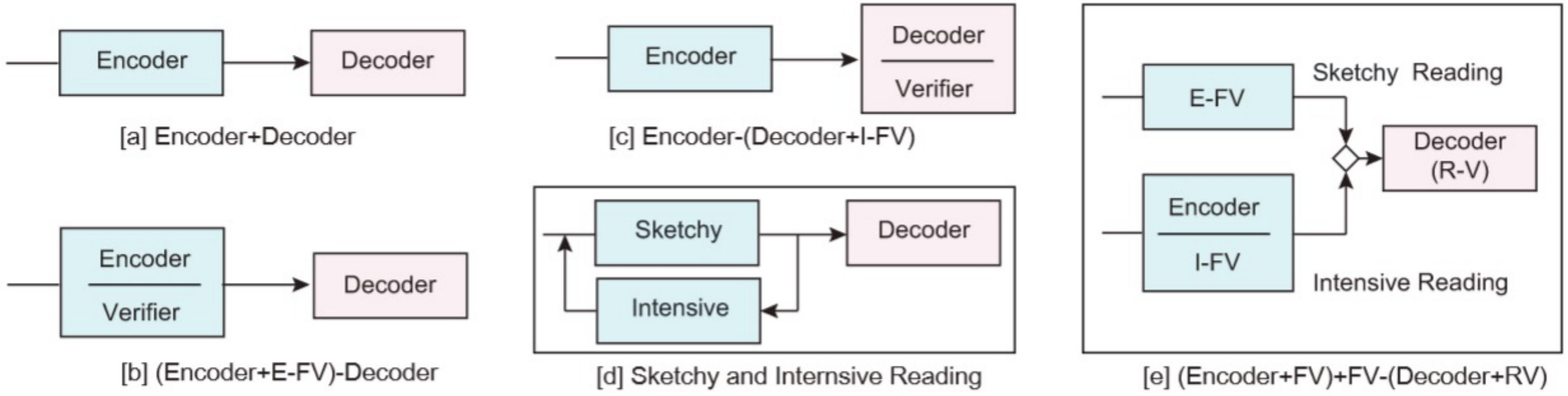
  - Reinforcement learning based self-critical learning to predict more acceptable answers:

Vanilla: maximize the log probabilities of the ground truth answer positions (exact match)

RL: Measure word overlap between predicted answer and ground truth.

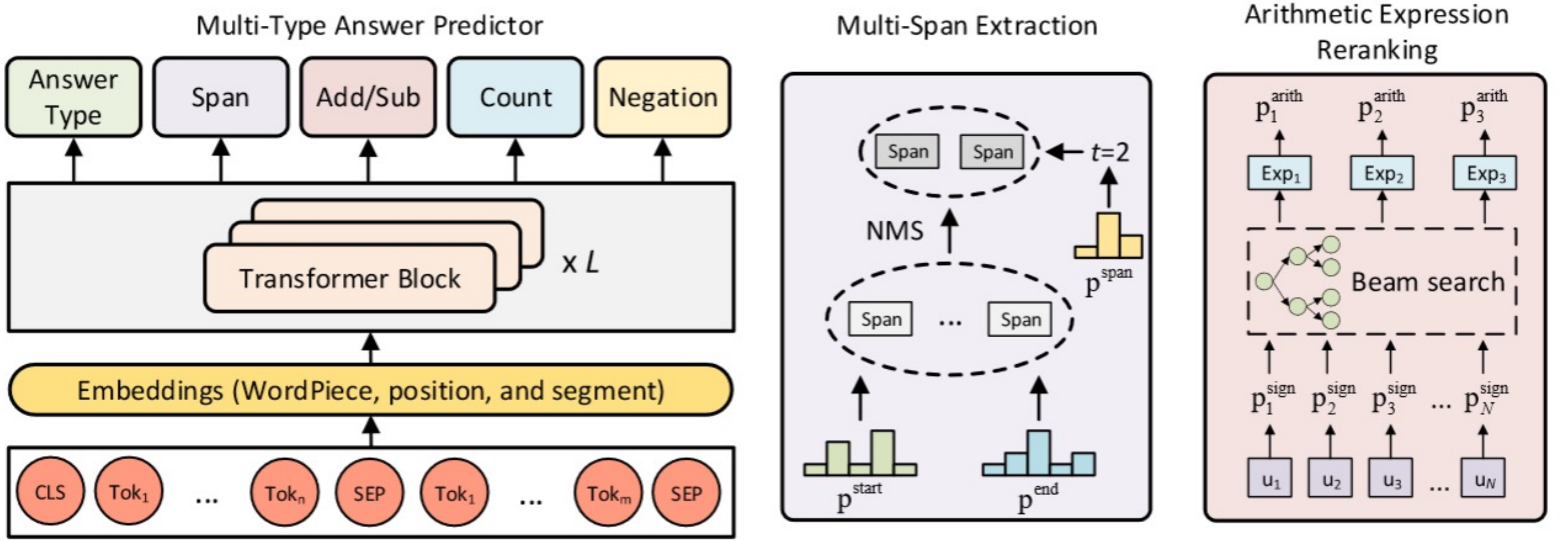
Decoder

- Answer Verifier:
  - Threshold-based answerable verification
  - Multitask-style verification
  - External parallel verification



Decoder

- Answer Type Predictor for multi-type MRC tasks

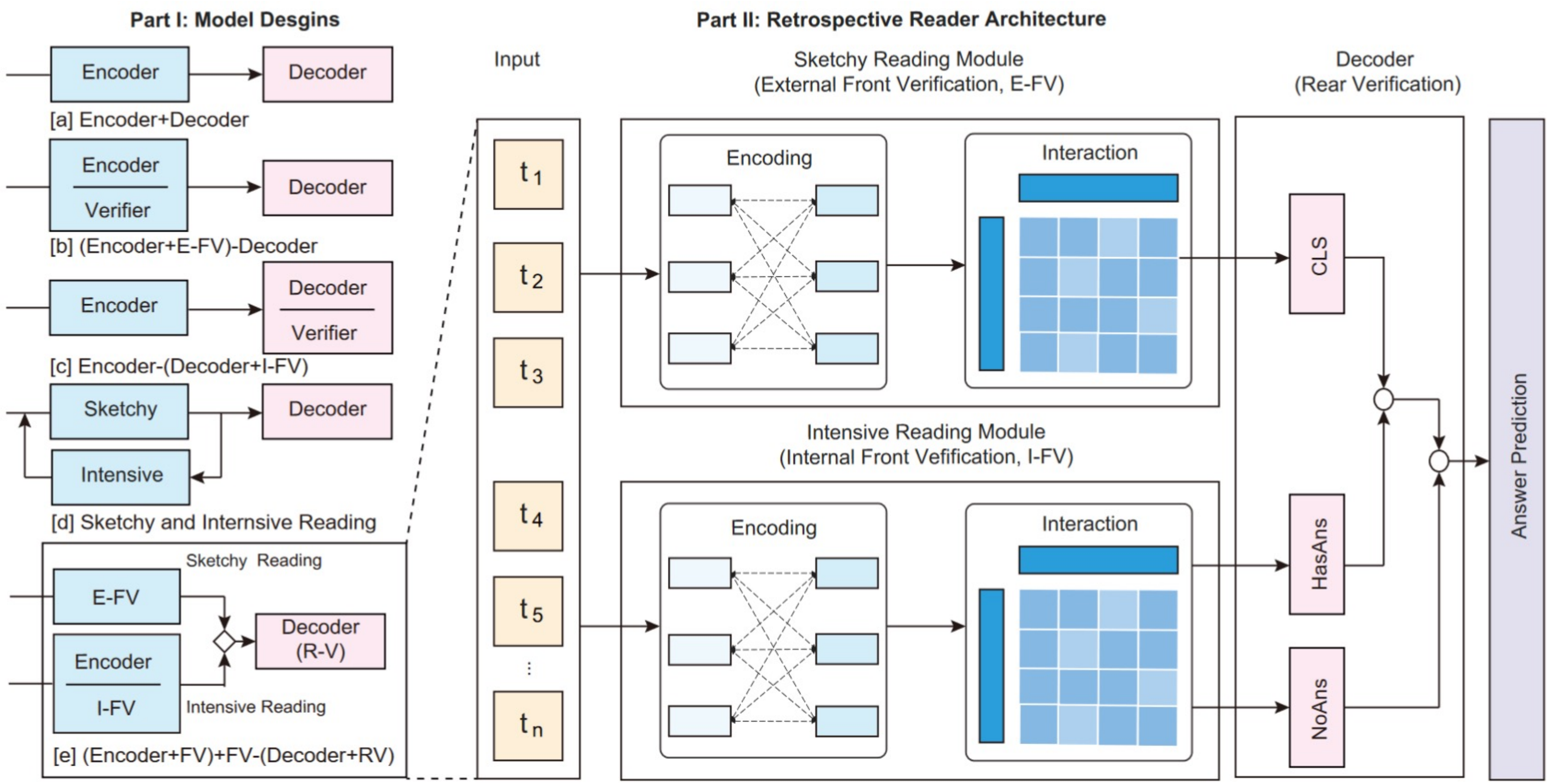


(MTMSN model from Hu et at., 2019)

Decoder (our work: answer verifier)

- Retro-Reader

Zhuosheng Zhang, Junjie Yang, Hai Zhao. Retrospective Reader for Machine Reading Comprehension. AAAI 2021.



Sketchy reading:

- Parallel External Verification

Intensive reading:

- Multitask Internal Verification

Rear Verification

Decoder (our work: answer verifier)

- Retro-Reader

SOTA results on SQuAD 2.0 and NewsQA

Passage: *Southern California consists of a heavily developed urban environment, home to some of the largest urban areas in the state, along with vast areas that have been left undeveloped. It is the third most populated megalopolis in the United States, after the Great Lakes Megalopolis and the Northeastern megalopolis. Much of southern California is famous for its large, spread-out, suburban communities and use of automobiles and highways...*
Question: *What are the second and third most populated megalopolis after Southern California?*
Answer: **Gold:** ⟨no answer⟩ **ALBERT (+TAV):** Great Lakes Megalopolis and the Northeastern megalopolis. **Retro-Reader over ALBERT:** ⟨no answer⟩ $score_{has} = 0.03, score_{na} = 1.73, \lambda = -0.98$

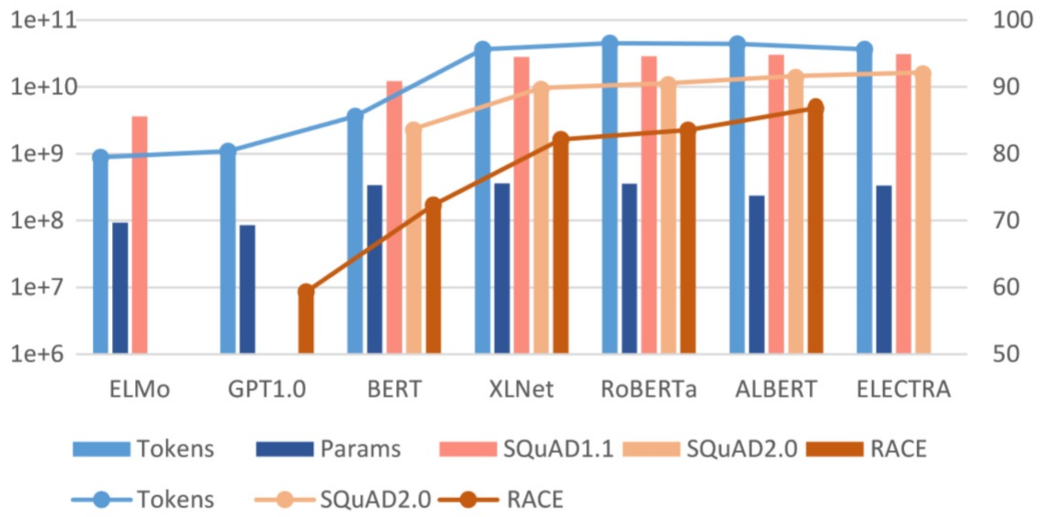
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University*	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Dec 08, 2019	ALBERT+Entailment DA (ensemble) *CloudWalk*	88.761	91.745
5 Jan 19, 2020	Retro-Reader on ALBERT (single model) *Shanghai Jiao Tong University*	88.107	91.419
5 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
5 Nov 22, 2019	albert+verifier (single model) *Ping An Life Insurance Company AI Team*	88.355	91.019

PrLMs greatly boost the benchmark of current MRC

Models	**Encoder**	**EM**	**F1**	**↑ EM**	**↑ F1**
Human (Rajpurkar, Jia, and Liang 2018)	-	82.304	91.221	-	-
Match-LSTM (Wang and Jiang 2016)	RNN	*64.744*	*73.743*	-	-
DCN (Xiong, Zhong, and Socher 2016)	RNN	66.233	75.896	1.489	2.153
Bi-DAF (Seo et al. 2017)	RNN	67.974	77.323	3.230	3.580
Mnemonic Reader (Hu, Peng, and Qiu 2017)	RNN	70.995	80.146	6.251	6.403
Document Reader (Chen et al. 2017)	RNN	70.733	79.353	5.989	5.610
DCN+ (Xiong, Zhong, and Socher 2017)	RNN	75.087	83.081	10.343	9.338
r-net (Wang et al. 2017)	RNN	76.461	84.265	11.717	10.522
MEMEN (Pan et al. 2017)	RNN	78.234	85.344	13.490	11.601
QANet (Yu et al. 2018)*	TRFM	80.929	87.773	16.185	14.030
CLMs					
ELMo (Peters et al. 2018)	RNN	78.580	85.833	13.836	12.090
BERT (Devlin et al. 2018)*	TRFM	85.083	91.835	20.339	18.092
SpanBERT (Joshi et al. 2020)	TRFM	88.839	94.635	24.095	20.892
XLNet (Yang et al. 2019c)	TRFM-XL	89.898	95.080	25.154	21.337

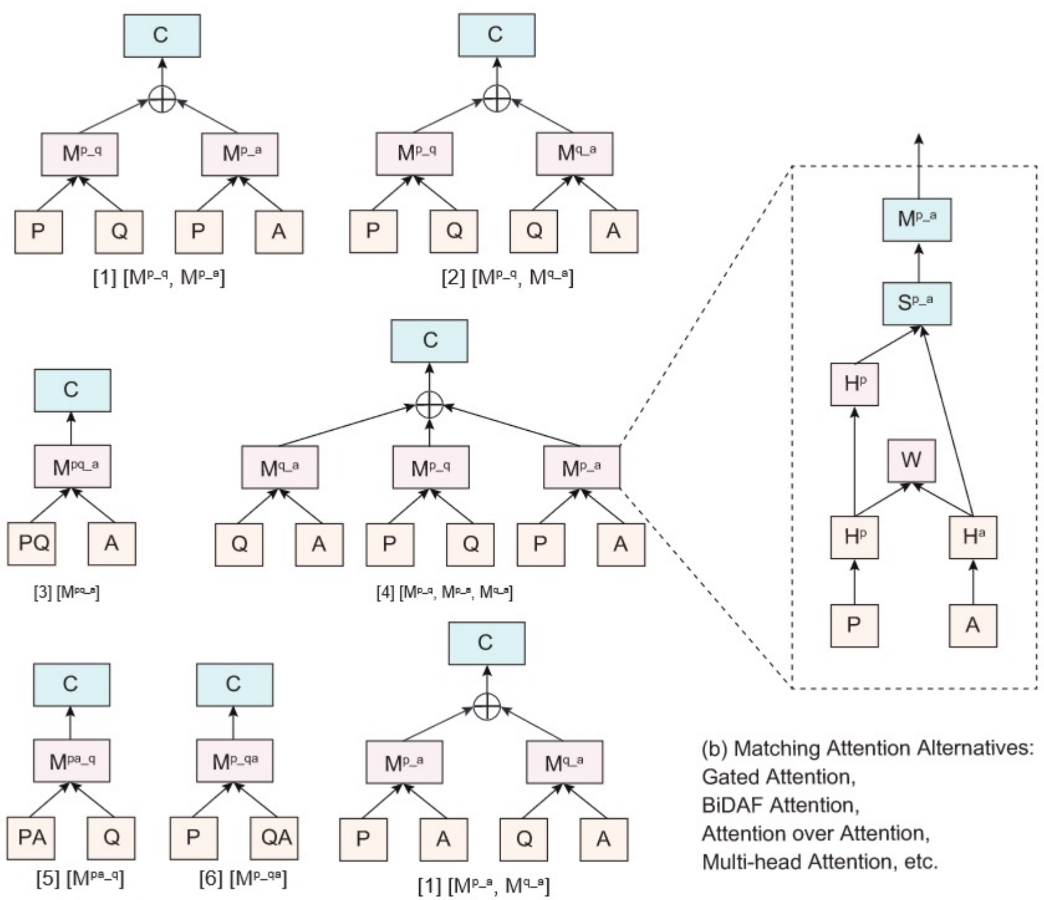
Models	**Encoder**	**SQuAD 2.0**	**↑ F1**	**RACE**	**↑ Acc**
Human (Rajpurkar, Jia, and Liang 2018)	-	91.221	-	-	
GPT_{v1} (Radford et al. 2018)	TRFM	-	-	*59.0*	-
BERT (Devlin et al. 2018)	TRFM	*83.061*	-	72.0	-
SemBERT (Zhang et al. 2020b)	TRFM	87.864	4.803	-	-
SG-Net (Zhang et al. 2020c)	TRFM	87.926	4.865	-	-
RoBERTa (Liu et al. 2019c)	TRFM	89.795	6.734	83.2	24.2
ALBERT (Lan et al. 2019)	TRFM	90.902	7.841	86.5	27.5
XLNet (Yang et al. 2019c)	TRFM-XL	90.689	7.628	81.8	22.8
ELECTRA (Clark et al. 2019c)	TRFM	91.365	8.304	-	-

Method	Tokens	Size	Params	SQuAD1.1 Dev	SQuAD1.1 Test	SQuAD2.0 Dev	SQuAD2.0 Test	RACE
ELMo	800M	-	93.6M	85.6	85.8	-	-	-
GPT_{v1}	985M	-	85M	-	-	-	-	59.0
$XLNet_{large}$	33B	-	360M	94.5	95.1*	88.8	89.1*	81.8
$BERT_{large}$	3.3B	13GB	340M	91.1	91.8*	81.9	83.0	72.0†
$RoBERTa_{large}$	-	160GB	355M	94.6	-	89.4	89.8	83.2
$ALBERT_{xxlarge}$	-	157GB	235M	94.8	-	90.2	90.9	86.5
$ELECTRA_{large}$	33B	-	335M	94.9	-	90.6	91.4	-



- Knowledge from large-scale copora
- Deep architectures

Decline of Matching Attention



(b) Matching Attention Alternatives: Gated Attention, BiDAF Attention, Attention over Attention, Multi-head Attention, etc.

(a) sequence-aware interaction patterns

Method	Att. Type	CNN val	CNN test	DailyMail val	DailyMail test
Attentive Reader (Hermann et al. 2015)	UA	61.6	63.0	70.5	69.0
AS Reader (Kadlec et al. 2016)	UA	68.6	69.5	75.0	73.9
Iterative Attention (Sordoni et al. 2016)	UA	72.6	73.3	-	-
Stanford AR (Chen, Bolton, and Manning 2016)	UA	73.8	73.6	77.6	76.6
GAReader (Dhingra et al. 2017)	UA	73.0	73.8	76.7	75.7
AoA Reader (Cui et al. 2017)	BA	73.1	74.4	-	-
BiDAF (Seo et al. 2017)	BA	76.3	76.9	80.3	79.6

Model	Matching	M	H	RACE
Human Ceiling Performance (Lai et al. 2017)		95.4	94.2	94.5
Amazon Mechanical Turker (Lai et al. 2017)		85.1	69.4	73.3
HAF (Zhu et al. 2018a)	$[M^{P\_A}; M^{P\_Q}; M^{Q\_A}]$	45.0	46.4	46.0
MRU (Tay, Tuan, and Hui 2018)	$[M^{P\_Q\_A}]$	57.7	47.4	50.4
HCM (Wang et al. 2018a)	$[M^{P\_Q}; M^{P\_A}]$	55.8	48.2	50.4
MMN (Tang, Cai, and Zhuo 2019)	$[M^{Q\_A}; M^{A\_Q}; M^{P\_Q}; M^{P\_A}]$	61.1	52.2	54.7
GPT (Radford et al. 2018)	$[M^{P\_Q\_A}]$	62.9	57.4	59.0
RSM (Sun et al. 2019b)	$[M^{P\_QA}]$	69.2	61.5	63.8
DCMN (Zhang et al. 2019a)	$[M^{PQ\_A}]$	77.6	70.1	72.3
OCN (Ran et al. 2019a)	$[M^{P\_Q\_A}]$	76.7	69.6	71.7
BERT$_{large}$ (Pan et al. 2019b)	$[M^{P\_Q\_A}]$	76.6	70.1	72.0
XLNet (Yang et al. 2019c)	$[M^{P\_Q\_A}]$	85.5	80.2	81.8
+ DCMN+ (Zhang et al. 2020a)	$[M^{P\_Q}; M^{P\_O}; M^{Q\_O}]$	86.5	81.3	82.8
RoBERTa (Liu et al. 2019c)	$[M^{P\_Q\_A}]$	86.5	81.8	83.2
+ MMM (Jin et al. 2019a)	$[M^{P\_Q\_A}]$	89.1	83.3	85.0
ALBERT (Jin et al. 2019a)	$[M^{P\_Q\_A}]$	89.0	85.5	86.5
+ DUMA (Zhu, Zhao, and Li 2020)	$[M^{P\_QA}; M^{QA\_P}]$	90.9	86.7	88.0
Megatron-BERT (Shoeybi et al. 2019)	$[M^{P\_Q\_A}]$	91.8	88.6	89.5

Optimizing the decoder strategies also works

Reading Strategy based on human reading patterns

- Learning to skim text
- Learning to stop reading
- Retrospective reading
- Back and forth reading, highlighting, and self-assessment

Tactic Optimization:

- The **objective** of answer verification
- The **dependency** inside answer span
- **Re-ranking** of candidate answers

Data Augmentation

- Most high-quality MRC datasets are human-annotated and inevitably relatively small.
- Training Data Augmentation:
  - Combining various MRC datasets as training data augmentation
  - Multi-tasking
  - Automatic question generation, such as back translation and synthetic generation
- Large-scale Pre-training
  - Recent studies showed that PrLMs well acquired linguistic information through pre-training
  - Some commonsense would be also entailed after pre-training.

Our Empirical Analysis

- Interaction: Dot Attention (DT-ATT); Multi-head Attention (MH-ATT)
- Verification: parallel external verifier (E-FV); multi-task based internal front verifier (I-FV); Rear verifier (I-FV+E-FV)
- Answer Dependency: using start logits and final sequence hidden states to obtain the end logits (SED).

Method	BERT EM	BERT F1	ALBERT EM	ALBERT F1
Baseline	78.8	81.7	87.0	90.2
Interaction				
+ MH-ATT	78.8	81.7	87.3	90.3
+ DT-ATT	78.3	81.4	86.8	90.0
Verification				
+ E-FV	79.1	82.1	87.4	90.6
+ I-FV-CE	78.6	82.0	87.2	90.3
+ I-FV-BE	78.8	81.8	87.2	90.2
+ I-FV-MSE	78.5	81.7	87.3	90.4
+ All I-FVs	79.4	82.1	87.5	90.6
+ All I-FVs + E-FV	79.6	82.5	87.7	90.8
Answer Dependency				
+ SED	79.1	81.9	87.3	90.3

Findings:

- Adding extra matching interaction layers heuristically after the strong PrLMs would be trivial.
- Either of the front verifiers boosts the baselines, and integrating all the verifiers can yield even better results
- Answer dependency can effectively improve the exact match score, yielding a more exactly matched answer span.

Interpretability of Human-parity Performance

- What kind of knowledge or reading comprehension skills the systems have grasped?
- For PrLM encoder side:
  - good at linguistic notions of syntax and coreference.
  - struggles with challenging inferences and role-based event prediction
  - obvious failures with the meaning of negation
- For MRC model side
  - overestimated ability of MRC systems that do not necessarily provide human-level understanding
  - unprecise benchmarking on the existing datasets.
  - suffers from adversarial attacks
- Decomposition of Prerequisite Skills
  - decompose the skills required by the dataset and take skill-wise evaluations
  - provide more explainable and convincing benchmarking of model capacity

New Frontiers

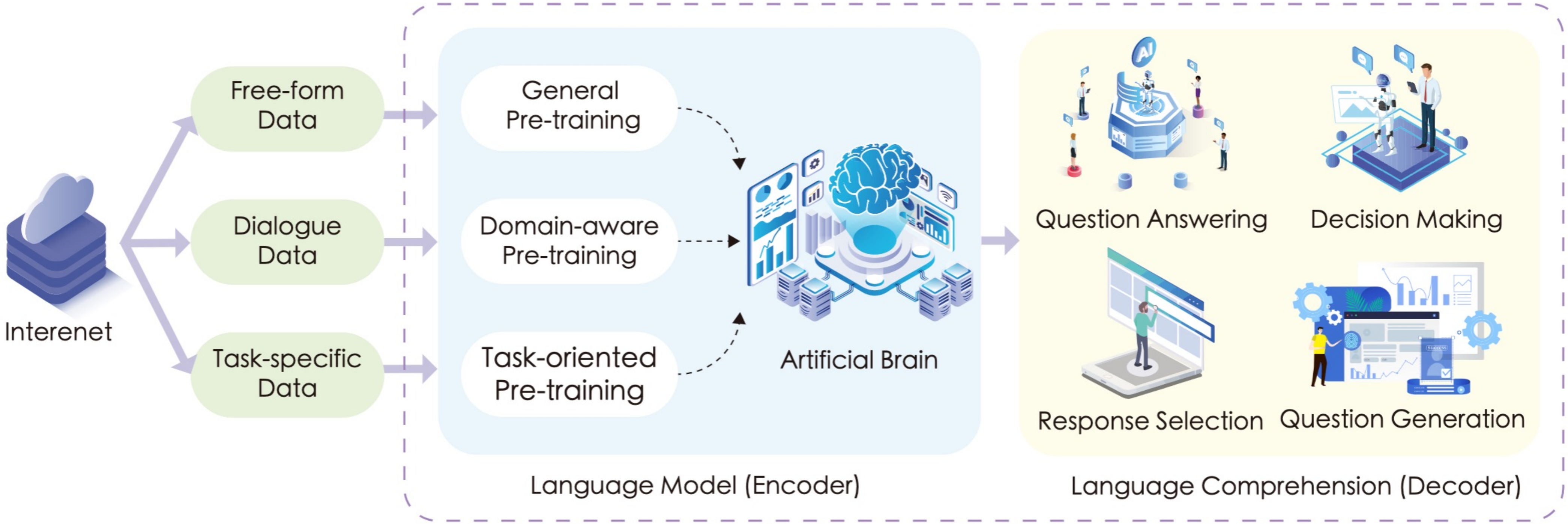
- Techniques
  - Domain/Task-adaptive Pre-training
  - Graph-aware Knowledge Structure Modeling
- Tasks
  - Multi-turn Dialogue Comprehension
  - Logical Reasoning
  - Commonsense Reasoning
- Applications
  - Open-domain QA
  - Multilingual, Multimodal, Multitask

New Frontiers

- Techniques
  - Domain/Task-adaptive Pre-training
  - Graph-aware Knowledge Structure Modeling
- Tasks
  - Multi-turn Dialogue Comprehension
  - Logical Reasoning
  - Commonsense Reasoning
- Applications
  - Open-domain QA
  - Multilingual, Multimodal, Multitask

Domain/Task-adaptive Pre-training

- General-purpose Pre-training (e.g., mask language modeling)
- Domain-aware Pre-training (e.g., science, news, medical domains)
- Task-oriented Pre-training (e.g., dialogue/discourse structure modeling)



Zhuosheng Zhang, Hai Zhao, 2021.* Advances in Multi-turn Dialogue Comprehension: A Survey.

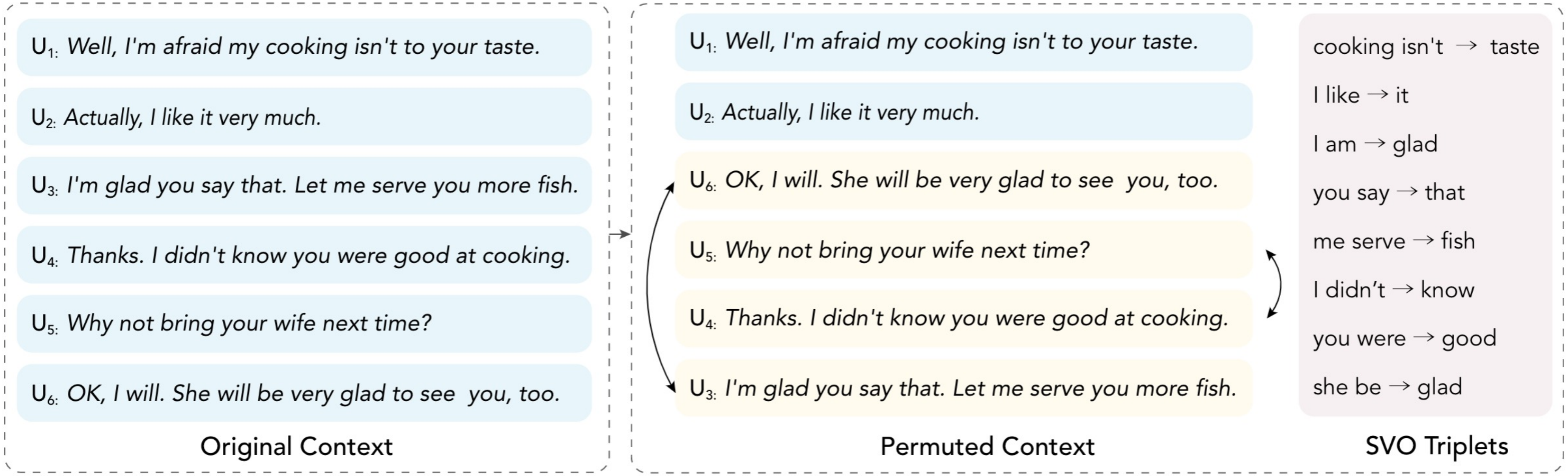
Dialogue-aware Pre-training (SPIDER)

- **Background:** How to train language models on dialogue scenarios
  - open-domain pre-training
  - domain-adaptive post-training
- **Motivation:**
  - The pre-trained models handle the whole input text as a linear sequence of successive tokens
  - It is challenging to effectively capture task-related knowledge from dialogue texts
  - Dialogue contexts are composed of many utterances from different speakers
  - Dialogues are rich in complex discourse structures and correlations

Zhuosheng Zhang, Hai Zhao. Structural Pre-training for Dialogue Comprehension. ACL 2021.

Dialogue-aware Pre-training (SPIDER)

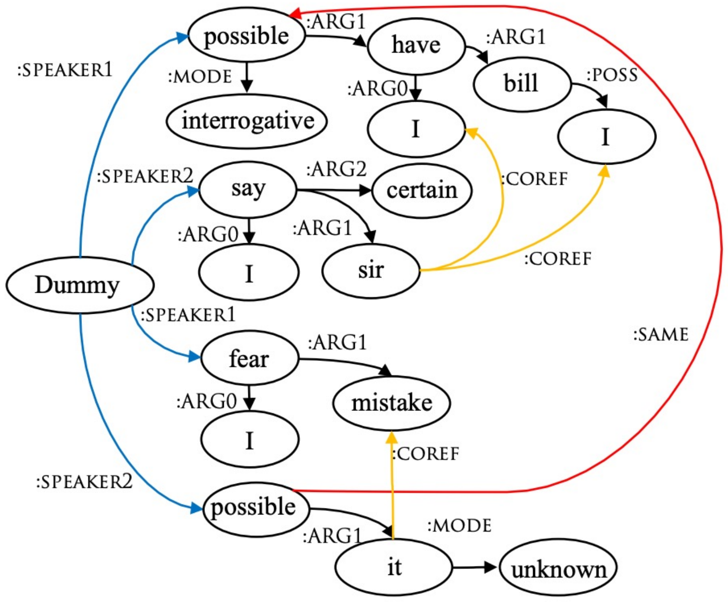
- **SPIDER**: Structural Pre-trained Dialogue Reader
  - **utterance order restoration:** predicts the order of the permuted utterances
  - **sentence backbone regularization:** improve the factual correctness of SVO triples
- Efficiently and explicitly model the coherence among utterances and the key facts in utterances



Zhuosheng Zhang, Hai Zhao. Structural Pre-training for Dialogue Comprehension. ACL 2021.

Graph-aware Knowledge Structure Modeling

- Technical trend: Graph Neural Network (GNN)
  - Injecting extra commonsense from knowledge graphs
  - Modeling entity relationships
  - Graph-attention can be considered as a special case of self-attention
- Application Scenarios
  - Entity linking and coreference modeling
  - Dialogue discourse structure
  - Abstract meaning representation (AMR)



AMR

Bai, Xuefeng, et al. Semantic Representation for Dialogue Modeling. ACL2021.

New Frontiers

- Techniques
  - Domain/Task-adaptive Pre-training
  - Graph-aware Knowledge Structure Modeling
- Tasks
  - Multi-turn Dialogue Comprehension
  - Logical Reasoning
  - Commonsense Reasoning
- Applications
  - Open-domain QA
  - Multilingual, Multimodal, Multitask

Discourse Graph Modeling for Dialogue Comprehension

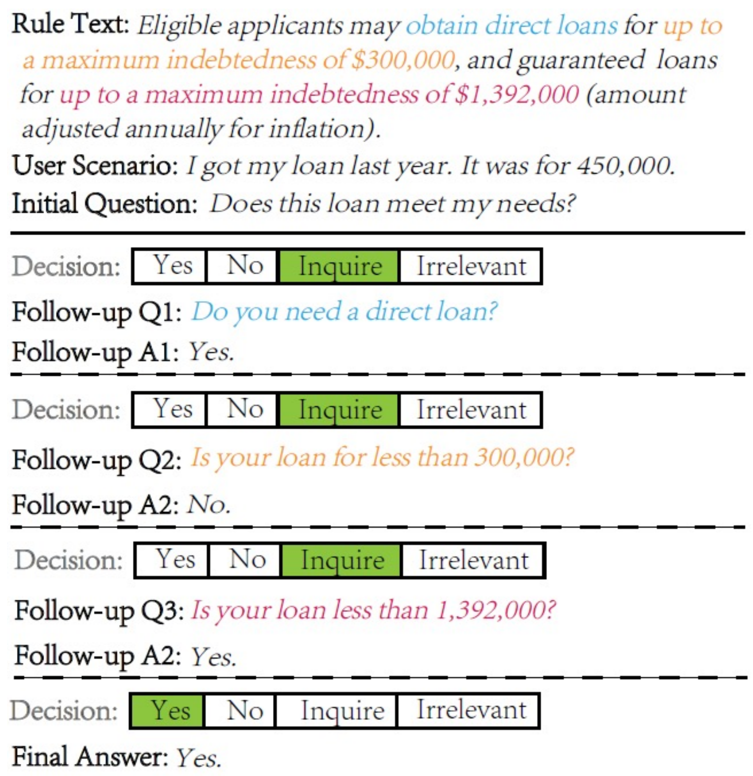
- Task: Conversational Machine Reading

  Input x = (r, s, q, h)

  - r: Rule Text
  - s: User Scenario
  - q: Initial Question
  - h: Dialogue History

  Output (divided into two subtasks):

  - A decision ∈ (yes, no, inquire, irrelevant)
  - If *inquire*, ask a follow-up question



An example taken from the ShARC (Saeidi et al., 2018) benchmark

Siru Ouyang, Zhuosheng Zhang, Hai Zhao*, 2021. Dialogue Graph Modeling for Conversational Machine Reading. Findings of ACL 2021.

Discourse Graph Modeling for Dialogue Comprehension

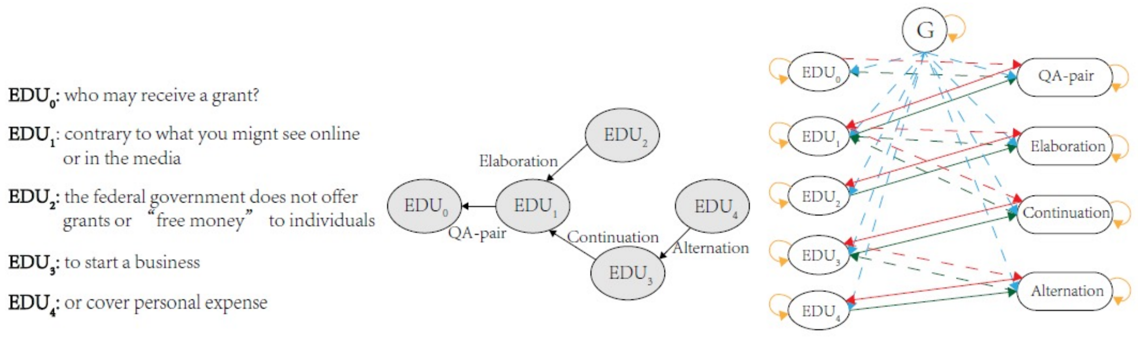
- Interpreting rule document
  - Identify rule conditions
  - Discourse relations among rule conditions
  - Interactions among all the elements (scenario, question, etc.)
- Make decisions as the conversation flows
  - Track fulfillment over identified rule conditions
  - jointly consider fulfillment states to make the final decision

Siru Ouyang, Zhuosheng Zhang, Hai Zhao*, 2021. Dialogue Graph Modeling for Conversational Machine Reading. Findings of ACL 2021.

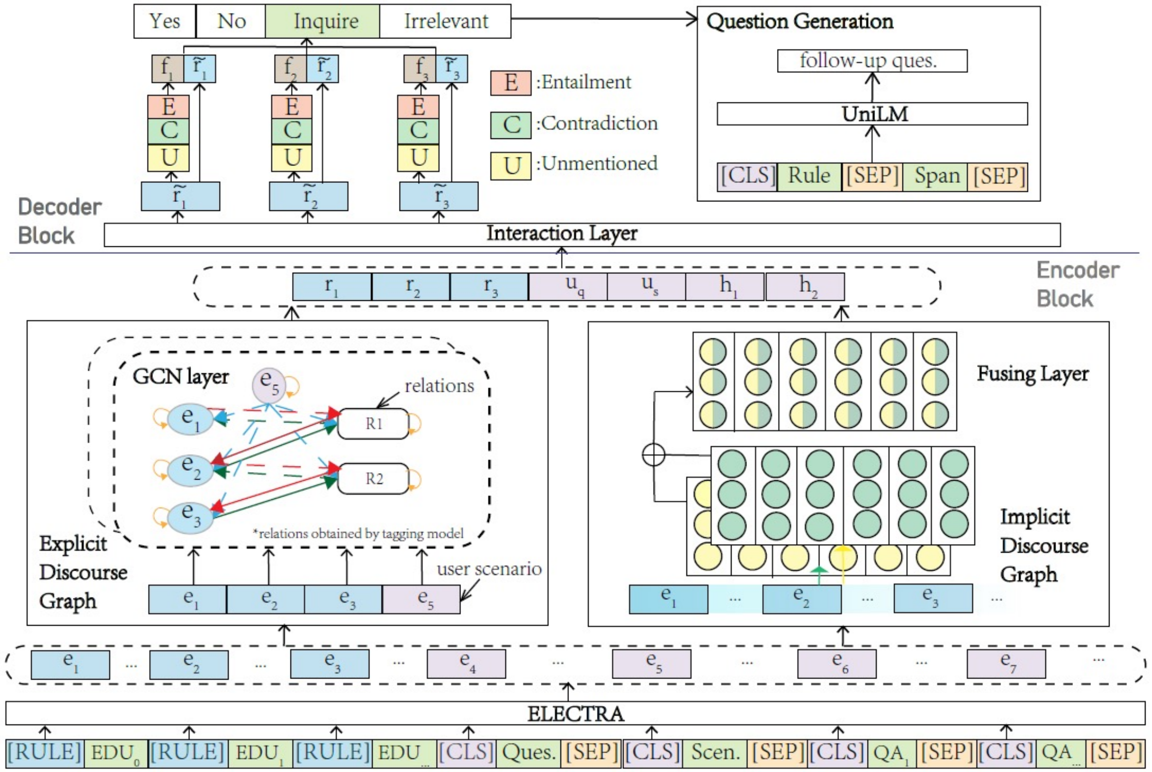
Discourse Graph Modeling for Dialogue Comprehension

- Explicit Discourse Graph:

  injects the discourse relations via open-source tagging tool



(a) separated EDUs (b) the original graph (c) the Levi graph



Using RGCN models to encode the graph.

Siru Ouyang, Zhuosheng Zhang, Hai Zhao*, 2021. Dialogue Graph Modeling for Conversational Machine Reading. Findings of ACL 2021.

Discourse Graph Modeling for Dialogue Comprehension

Model	Dev Set				Test Set			
	Decision Making		Question Gen.		Decision Making		Question Gen.	
	Micro	Macro	BLEU1	BLEU4	Micro	Macro	BLEU1	BLEU4
NMT (Saeidi et al., 2018)	-	-	-	-	44.8	42.8	34.0	7.8
CM (Saeidi et al., 2018)	-	-	-	-	61.9	68.9	54.4	34.4
BERTQA (Zhong and Zettlemoyer, 2019)	68.6	73.7	47.4	54.0	63.6	70.8	46.2	36.3
UcraNet (Verma et al., 2020)	-	-	-	-	65.1	71.2	60.5	46.1
BiSon (Lawrence et al., 2019)	66.0	70.8	46.6	54.1	66.9	71.6	58.8	44.3
E^3 (Zhong and Zettlemoyer, 2019)	68.0	73.4	67.1	53.7	67.7	73.3	54.1	38.7
EMT (Gao et al., 2020a)	73.2	78.3	67.5	53.2	69.1	74.6	63.9	**49.5**
DISCERN (Gao et al., 2020b)	74.9	79.8	65.7	52.4	73.2	78.3	**64.0**	49.1
DGM (ours)	**78.6**	**82.2**	**71.8**	**60.2**	**77.4**	**81.2**	63.3	48.4

Evaluation Metrics

- Decision Making: Micro-accuracy and Macro-accuracy
- Question Generation: BLEU1 and BLEU4

Siru Ouyang, Zhuosheng Zhang, Hai Zhao*, 2021. Dialogue Graph Modeling for Conversational Machine Reading. Findings of ACL 2021.

Fact-driven Logical Reasoning

- Task: Logical Reasoning
  - Challenges: entity-aware commonsense, perception of facts or events.
  - Logical supervision is rarely available during language model pre-training.

Question	***Passage***	***Answer***
Example 1 From this we know	Xiao Wang is taller than Xiao Li, Xiao Zhao is taller than Xiao Qian, Xiao Li is shorter than Xiao Sun, and Xiao Sun is shorter than Xiao Qian.	✓ A. Xiao Li is shorter than Xiao Zhao. B. Xiao Wang is taller than Xiao Zhao. C. Xiao Sun is shorter than Xiao Wang. D. Xiao Sun is taller than Xiao Zhao.
Example 2 Which one of the following statements, most seriously weakens the argument?	.... A large enough comet colliding with Earth could have caused a cloud of dust that enshrouded the planet and cooled the climate long enough to result in the dinosaurs' demise.	A. Many other animal species from same era did not become extinct at the same time the dinosaurs did. B. It cannot be determined from dinosaur skeletons whether the animals died from the effects of a dust cloud. C. The consequences for vegetation and animals of a comet colliding with Earth are not fully understood. ✓ D. Various species of animals from the same era and similar to them in habitat and physiology did not become extinct.

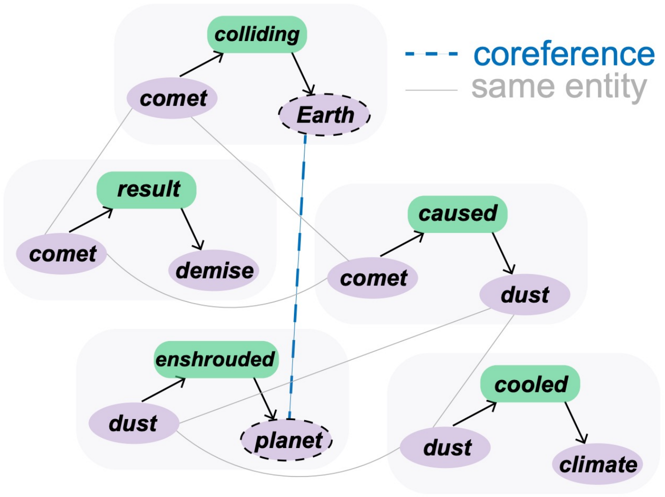
Siru Ouyang, Zhuosheng Zhang, Hai Zhao, 2021. Fact-driven Logical Reasoning.

Fact-driven Logical Reasoning

- Natural logic units would be the group of backbone constituents of the sentence such as subject, verb and object that cover both global and local knowledge pieces.

Definition 1 ***(Fact Unit)*** *Given an triplet $T = \{E_1, R, E_2\}$, where E_1 and E_2 are entities, P is the predicate between them, a fact unit F is the set of all entities in T and their corresponding relations.*

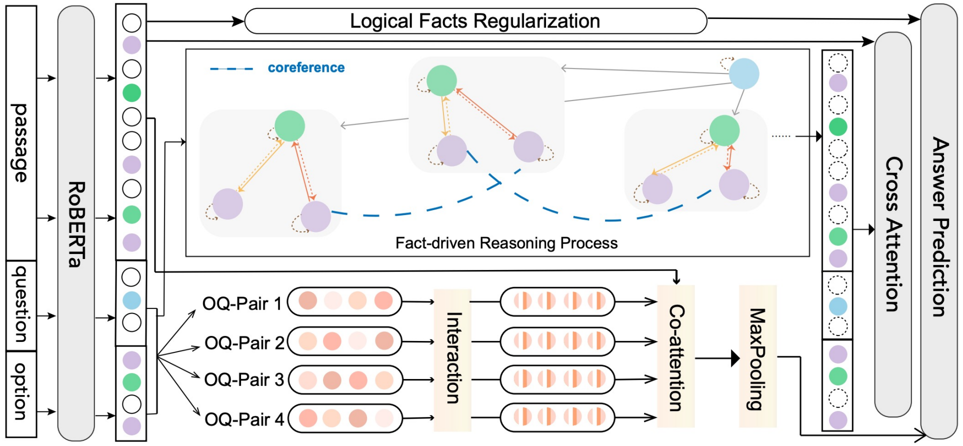
Definition 2 ***(Supergraph)*** *A supergraph is a structure made of fact units (regarded as subgraphs) as the vertices, and the coreference relations as undirected edges.*

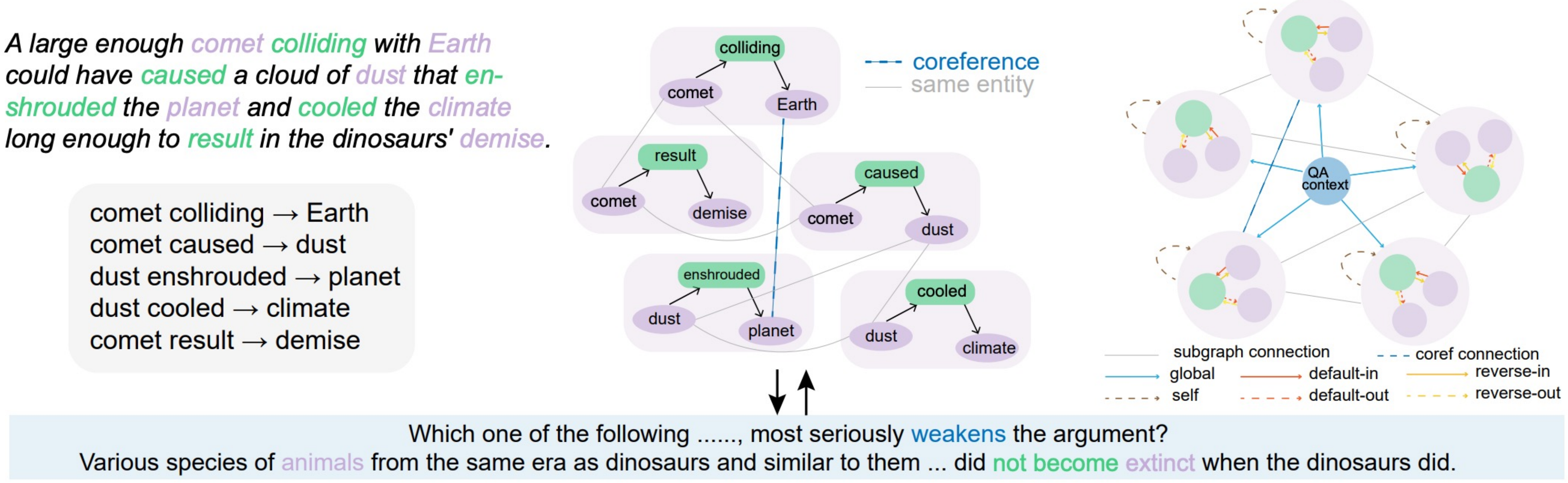


Siru Ouyang, Zhuosheng Zhang, Hai Zhao, 2021. Fact-driven Logical Reasoning.

Fact-driven Logical Reasoning

- Supergraph Modeling
  - Build a supergraph on our newly defined fact units
  - Question-Option-aware Interaction
  - Logical Fact Regularization





Siru Ouyang, Zhuosheng Zhang, Hai Zhao, 2021. Fact-driven Logical Reasoning.

Fact-driven Logical Reasoning

- Dramatic improvements on the logical reasoning benchmarks
- FOCAL REASONER makes better use of logical structure inherent in the given context to perform reasoning than existing methods.

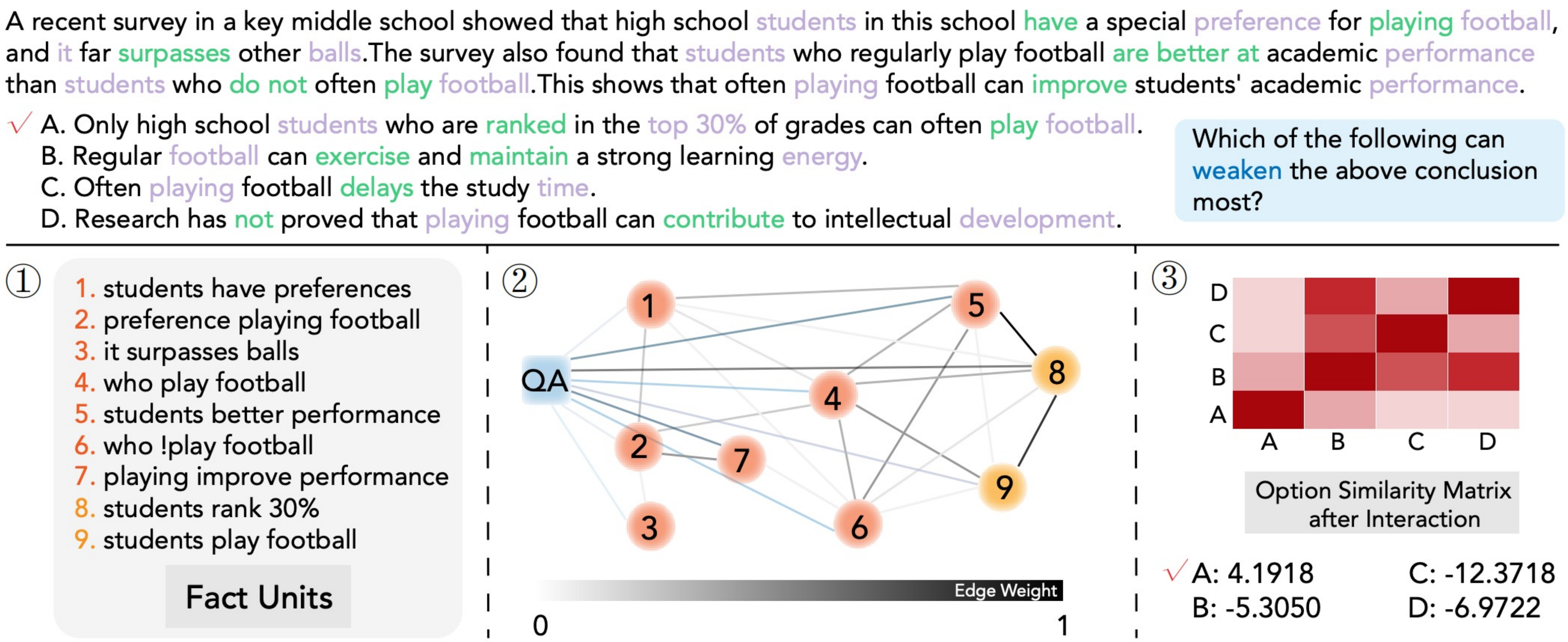
Model	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
Human [3]	-	63.00	57.10	67.20	-	86.00
BERT-Large [3]	53.80	49.80	72.00	32.30	34.10	31.03
XLNet-Large [3]	62.00	56.00	75.70	40.50	-	-
RoBERTa-Large [3]	62.60	55.60	75.50	40.00	35.02	35.33
DAGN [6]	65.20	58.20	76.14	44.11	35.48	38.71
DAGN (Aug) [6]	65.80	58.30	75.91	44.46	36.87	39.32
FOCAL REASONER	**66.80**	**58.90**	**77.05**	**44.64**	**41.01**	**40.25**

Model	MuTual						MuTual^{plus}					
	Dev Set			Test Set			Dev Set			Test Set		
	$R_4@1$	$R_4@2$	MRR	$R_4@1$	$R_4@2$	MRR	$R_4@1$	$R_4@2$	MRR	$R_4@1$	$R_4@2$	MRR
RoBERTa$_{base}$ [38]	69.5	87.8	82.4	71.3	89.2	83.6	62.2	85.3	78.2	62.6	**86.6**	78.7
-MC [38]	69.3	88.7	82.5	68.6	88.7	82.2	62.1	83.0	77.8	64.3	84.5	79.2
FOCAL REASONER	**73.4**	**90.3**	**84.9**	**72.7**	**91.0**	**84.6**	**63.7**	**86.1**	**79.1**	**65.5**	84.3	**79.7**

Siru Ouyang, Zhuosheng Zhang, Hai Zhao, 2021. Fact-driven Logical Reasoning.

Fact-driven Logical Reasoning

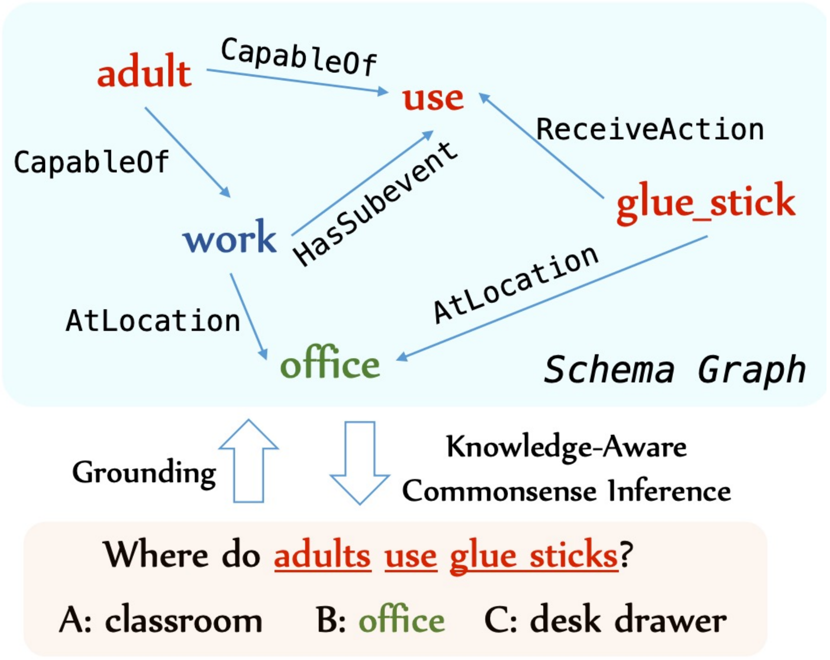
- An example of how our model reasons to get the final answer



Siru Ouyang, Zhuosheng Zhang, Hai Zhao, 2021. Fact-driven Logical Reasoning.

Commonsense Reasoning

- Resources (in natural language)
  - ConceptNet: semantic knowledge in natural language form
  - ATOMIC: knowledge of cause and effect
- Injecting commonsense into neural networks
  - Inserting into the texts
  - Attention-based interaction
  - Multi-task learning
- Temporal commonsense
  - Understand temporal relations: order, duration, frequency, …, of events

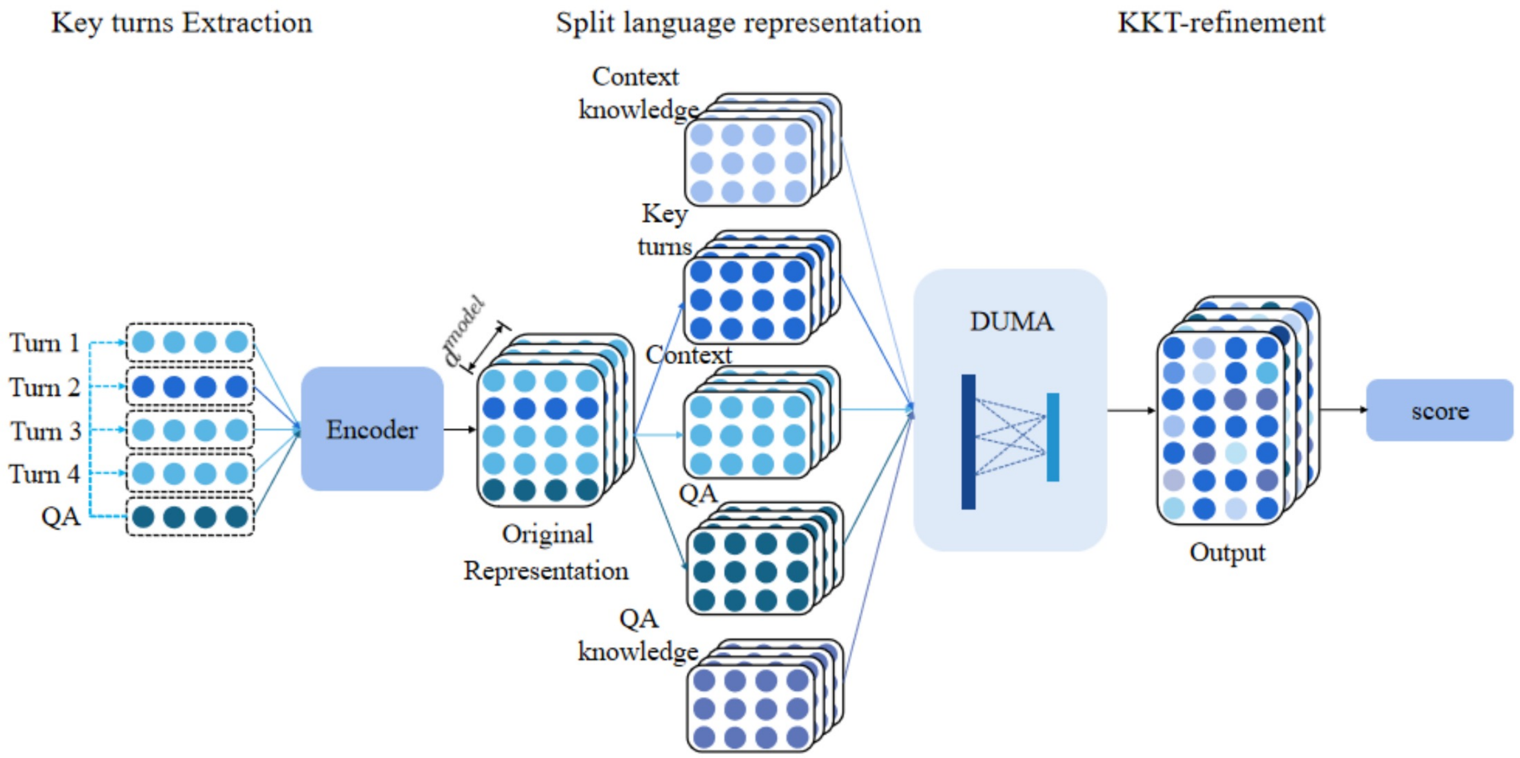


(From KagNet)

[1] Lin, Bill Yuchen, et al. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. EMNLP 2019.
[2] https://homes.cs.washington.edu/~msap/acl2020-commonsense/

Commonsense Reasoning (KKT)

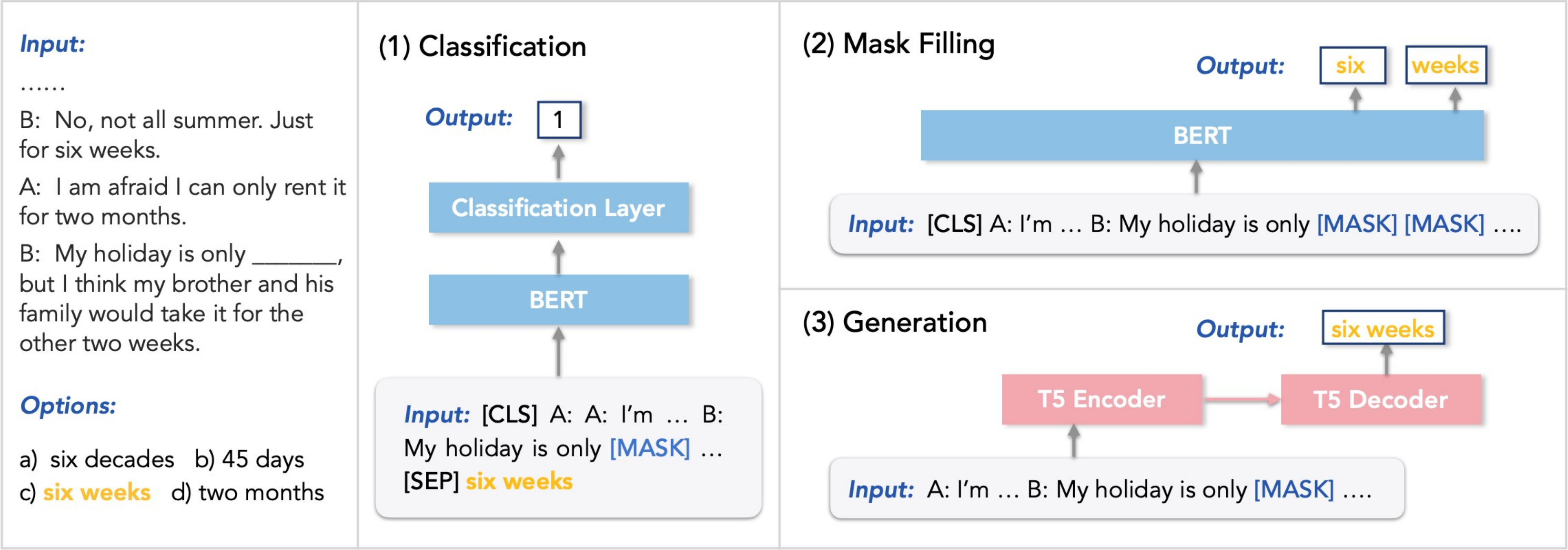
- Retrieve Relevant Knowledge from ConceptNet
- Filter the informative knowledge and use the selected knowledge to enhance the context



Junlong Li, Zhuosheng Zhang, Hai Zhao. Multi-turn Dialogue Reading Comprehension with Pivot Turns and Knowledge. TASLP.

Temporal Commonsense

- Understand temporal relations: order, duration, frequency, …, of events
- Humans can easily answer these questions (97.8% accuracy)
- The best model variant (T5-large with in-domain training) struggles on this challenge set (73%)



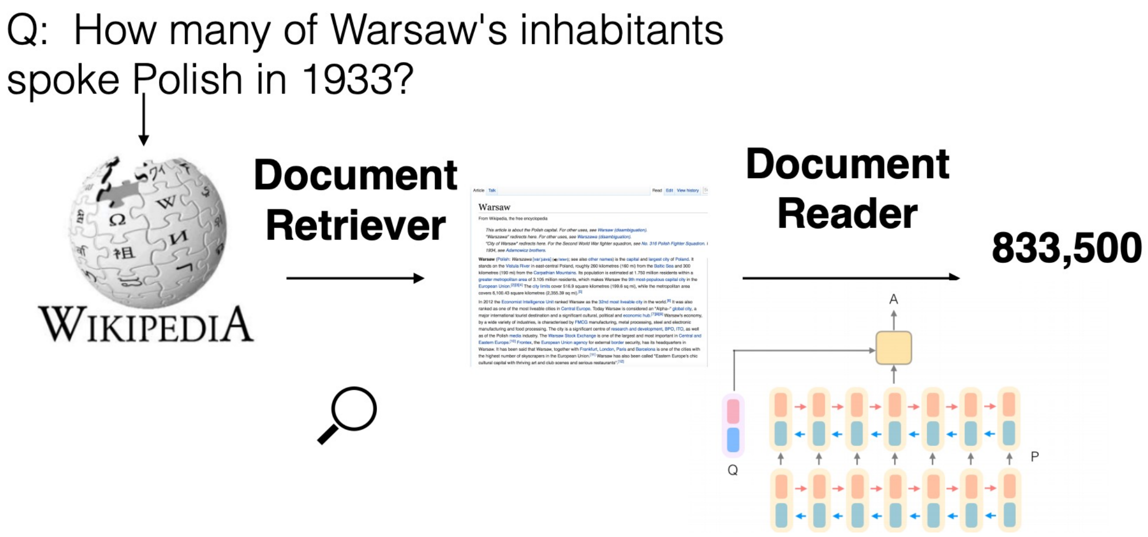
Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi and Manaal Faruqui. TIMEDIAL: Temporal Commonsense Reasoning in Dialog. ACL 2021.

New Frontiers

- Techniques
  - Domain/Task-adaptive Pre-training
  - Graph-aware Knowledge Structure Modeling
- Tasks
  - Multi-turn Dialogue Comprehension
  - Logical Reasoning
  - Commonsense Reasoning
- Applications
  - Open-domain QA
  - Multilingual, Multimodal, Multitask

Open-Domain QA

- Reading Comprehension = Document-level Modeling + QA
- Open-Domain QA= Open-Domain Reading Comprehension = Open-Domain Document Modeling + QA
  - Machine Reading Comprehension over the whole internet
- Typical architecture
  - Traditional Retriever-Reader architecture
  - Dense Retrieval vs. BM25
  - Span extraction based on the retrieved documents
- Next-generation Search Engine



Chen, Danqi, et al. 2017. Reading wikipedia to answer open-domain questions. ACL 2017.

Open-Domain QA: DPR

- Dense Passage Retriever (DPR)
  - maps any text passage to a fixed dimension of real-valued vectors
  - builds an index for all the passages that we will use for retrieval.

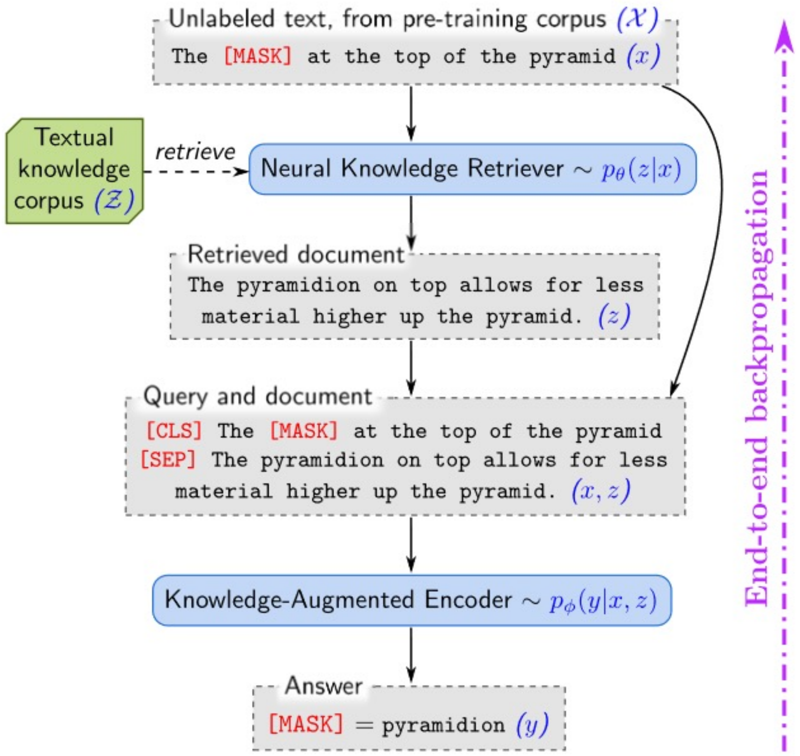
Training	**Retriever**	**Top-20** NQ	TriviaQA	WQ	TREC	SQuAD	**Top-100** NQ	TriviaQA	WQ	TREC	SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
Single	DPR	78.4	79.4	73.2	79.8	63.2	85.4	**85.0**	81.4	89.1	77.2
	BM25 + DPR	76.6	79.8	71.0	85.2	**71.5**	83.8	84.5	80.5	92.7	**81.3**
Multi	DPR	**79.4**	78.8	**75.0**	**89.1**	51.6	**86.0**	84.7	**82.9**	93.9	67.6
	BM25 + DPR	78.0	**79.9**	74.7	88.5	66.2	83.9	84.4	82.3	**94.1**	78.6

Table 2: Top-20 & Top-100 retrieval accuracy on test sets, measured as the percentage of top 20/100 retrieved passages that contain the answer. *Single* and *Multi* denote that our Dense Passage Retriever (DPR) was trained using individial or combined training datasets (all the datasets excluding SQuAD). See text for more details.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

Open-Domain QA: REALM

- Two stages：Knowledge Retrieval + Language Modeling
- Retrieve and attend over documents from a large corpus such as Wikipedia
- Training Strategies：
  - Only mask “knowledge” tokens (entities, dates, etc.)
  - Add a special empty documents beyond the top-k ones
  - Avoid duplication of pre training documents and knowledge base documents
  - Warmup task：Inverse Cloze Task, retrieve the original document for the sentence



Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. ICML 2020.

Open-Domain QA: REALM

Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)	# params
BERT-Baseline (Lee et al., 2019)	Sparse Retr.+Transformer	BERT	26.5	17.7	21.3	110m
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1	-	223m
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2	-	738m
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4	-	11318m
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-	110m
ORQA (Lee et al., 2019)	Dense Retr.+Transformer	ICT+BERT	33.3	36.4	30.1	330m
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2	**46.8**	330m
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**	42.9	330m

- For open-domain QA
- Outperforms previous models
- When MRC requires information retrieval and language modeling, we can train
  - Retrieval-based langauge models
  - Pre-training LMs on the whole internet

Ablation	Exact Match	Zero-shot Retrieval Recall@5
REALM	38.2	38.5
REALM retriever+Baseline encoder	37.4	38.5
Baseline retriever+REALM encoder	35.3	13.9
Baseline (ORQA)	31.3	13.9
REALM with random uniform masks	32.3	24.2
REALM with random span masks	35.3	26.1
30× stale MIPS	28.7	15.1

Multilingual, Multimodal, Multitask

- Multitask
  - Training with various types of MRC corpus
- Multilingual/Cross-lingual
  - Languages other than English are not well-addressed due to the lack of data
- Multimodal Semantic Grounding
  - jointly modeling diverse modalities will be potential research interests
  - beneficial for real-world applications, e.g., online shopping and E-commerce customer support

[1] MRQA: Workshop on Machine Reading for Question Answering
[2] Cui, Yiming, et al. Cross-Lingual Machine Reading Comprehension. EMNLP 2019.
[3] Anthony Ferritto, Sara Rosenthal, Mihaela Bornea, Kazi Hasan, Rishav Chakravarti, Salim Roukos, Radu Florian, Avirup Sil. A Multilingual Reading Comprehension System for more than 100 Languages. COLING 2020 (Demos).
[4] Hao Tan, Mohit Bansal. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP 2020.

Conclusion

- MRC boosts the progress from language processing to understanding
- The rapid improvement of MRC systems greatly benefits from the progress of PrLMs
- The theme of MRC is gradually moving from shallow text matching to cognitive reasoning

Our Survey Papers:

[1] Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond

Paper Link: https://arxiv.org/abs/2005.06249

[2] Advances in Multi-turn Dialogue Comprehension: A Survey

Paper Link: https://arxiv.org/abs/2103.03125

Our codes are publicly available at: https://github.com/cooelf

Q & A

Thank You !