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

Zhuosheng Zhang
zhangzs@sjtu.edu.cn
Shanghai Jiao Tong University, China

2020/05/25 (MDT)
Outline

❖ Introductions to Machine Reading Comprehension (MRC)
❖ Development of Contextualized Language Model (CLM)
❖ Technical Methods
❖ Technical Highlights
❖ Trends and Discussions
❖ Conclusions
Outline

❖ Introductions to Machine Reading Comprehension (MRC)
❖ Development of Contextualized Language Model (CLM)
❖ Technical Methods
❖ Technical Highlights
❖ Trends and Discussions
❖ Conclusions
There are two categories of branches in 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
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
- 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** |
| ( @entity5 ) – a bus carrying members of @entity5 unit overturned at an @entity7 military base Sunday, killing 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. 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 @entity22. e-mail to a friend bus carrying @entity5 unit overturned at _____ military base @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 trade 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. The Silk Road became less important because _____, |
| **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. What do robots that resemble humans attempt to 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 5-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. How many more points did the Dolphins score compared to the Bills by the game’s end? |
| **Question** |
| How many more points did the Dolphins score compared to the Bills by the game’s end? |
| **Answer** |
| 28 |

A full collection of the latest datasets can be found in the Appendix in our survey paper.
The study of MRC has experienced two significant peaks, namely,

- the burst of deep neural networks, especially attention-based models;
- the evolution of CLMs.

The Boom of MRC researches
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 CLMs
- **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.

Related paper:

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

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

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

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.
Outline

❖ Introductions to Machine Reading Comprehension (MRC)
❖ Development of Contextualized Language Model (CLM)
❖ Technical Methods
❖ Technical Highlights
❖ Trends and Discussions
❖ Conclusions
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
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

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  |
What is CLM?

Revisit the definitions of the recent contextualized encoders:

- ELMo: Deep contextualized word representations
- BERT: Pre-training of deep bidirectional transformers for language understanding

The focus is **contextualized** representation from language models, in terms of

- the evolution of language representation architectures, and
- the actual usages of these models nowadays

Common practice:

- fine-tuning the model using task-specific data,
- pre-training is neither the necessary nor the core element.

Pre-training and fine-tuning are just the manners we use the models.

Therefore, we call these pre-trained models **contextualized language models** (CLMs) in our work.
The core is the evolution of CLM training objectives: \textit{n-gram, masked LM, permutation LM}, etc.

The standard and common objective: \textit{n-gram LM}.

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

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

Training objective:

\[
\max_{\theta} \sum_{w} \log p_{\theta}(w),
\]
The Evolution of CLM Training Objectives

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

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

A bidirectional form:

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

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

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

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

$$\Rightarrow \text{Autoregressive n-gram LM!} \quad \mathbb{E}_{z \in Z_L} \sum_{k=c+1}^{L} \log p_\theta(w_{zk} | w_{z1:k-1}).$$

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

Page 15
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

Transformer takes token positions in a sentence as inputs

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

MPNet: Masked LM + Premuted LM
Architectures of CLMs

- RNN: GRU/LSTM
- Transformer
- Transformer-XL
Derivative of CLMs

Masking Strategy

Knowledge Injection

Training Objective

Model Optimization.
## Performance of CLM derivatives

| Method               | SQuAD1.1 |      |      |      | SQuAD2.0 |      |      |      | RACE |      |
|----------------------|----------|------|------|------|----------|------|------|------|------|------|
|                      | Dev      | ↑ Dev| Test | ↑ Test| Dev      | ↑ Dev| Test | ↑ Test| Acc  | ↑ Acc|
| ELMo                 | 85.6     | -    | 85.8 | -    | -        | -    | -    | -    | -    | -    |
| GPTv1                | -        | -    | -    | -    | -        | -    | -    | -    | 59.0 | -    |
| BERTbase             | 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  | -    | -    | -        | -    | -    | -    | -    | -    |
| BERTlarge            | 91.1*    | 5.5  | 91.8*| 6    | 81.9     | 5.1  | 83.0 | -    | 72.0†| -    |
| SemBERTlarge         | -        | -    | -    | -    | 83.6     | 6.8  | 85.2 | 2.2  | -    | -    |
| SG-Net               | -        | -    | -    | -    | 88.3     | 11.5 | 87.9 | 4.9  | 74.2 | 15.2 |
| SpanBERTlarge        | -        | -    | 94.6 | 8.8  | -        | -    | 88.7 | 5.7  | -    | -    |
| StructBERTlarge      | 92.0     | 6.4  | -    | -    | -        | -    | -    | -    | -    | -    |
| RoBERTAlarge         | 94.6     | 9.0  | -    | -    | 89.4     | 12.6 | 89.8 | 6.8  | 83.2 | 24.2 |
| ALBERTxxlarge        | 94.8     | 9.2  | -    | -    | 90.2     | 13.4 | 90.9 | 7.9  | 86.5 | 27.5 |
| XLNetlarge           | 94.5     | 8.9  | 95.1*| 9.3  | 88.8     | 12    | 89.1*| 6.1  | 81.8 | 22.8 |
| UniLM                | -        | -    | -    | -    | 83.4     | 6.6  | -    | -    | -    | -    |
| ELECTRAlarge         | 94.9     | 9.3  | -    | -    | 90.6     | 13.8 | 91.4 | 8.4  | -    | -    |
| Megatron-LM3.9B      | 95.5     | 9.9  | -    | -    | 91.2     | 14.4 | -    | -    | 89.5 | 30.5 |
| T511B                | 95.6     | 10.0 | -    | -    | -        | -    | -    | -    | -    | -    |
Correlations Between MRC and CLM

MRC and CLM are complementary to each other.

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

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

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

|          | NLU | GLUE | SQuAD1.1 | MRC | SQuAD2.0 | RACE |
|----------|-----|------|----------|-----|----------|------|
| ELMo     | ✓   | ×    | ✓        | ×   | ✓        | ×    |
| GPT\(_v_1\) | ✓   | ✓    | ×        | ×   | ✓        | ✓    |
| BERT     | ×   | ✓    | ✓        | ✓   | ✓        | ×    |
| RoBERTa  | ×   | ✓    | ✓        | ✓   | ✓        | ✓    |
| ALBERT   | ×   | ✓    | ✓        | ✓   | ✓        | ✓    |
| XLNet    | ×   | ✓    | ✓        | ✓   | ✓        | ✓    |
| ELECTRA  | ×   | ✓    | ✓        | ✓   | ✓        | ×    |
Outline

❖ Introductions to Machine Reading Comprehension (MRC)
❖ Development of Contextualized Language Model (CLM)
❖ Technical Methods
❖ Technical Highlights
❖ Trends and Discussions
❖ Conclusions
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 CLM 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 CLMs for Fine-tuning**

  **Encoder:** CLM;  **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

Best single model in CMRC 2017 shared task

| Model               | CMRC-2017  |
|---------------------|------------|
|                     | Valid | 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  | Test  | CFT  |
|------------------|-----|-------|------|
|                  | Valid | Test | 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 †                       | 79.2         | 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 1977 merger 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] ARG0 [led] VERB [the university] ARG1 through [the Great Depression and World War II] 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
SemBERT: Semantics-aware

| Method | Classification | Natural Language Inference | Semantic Similarity | Score |
|--------|----------------|-----------------------------|---------------------|-------|
|        | COLA (acc)     | SST-2 (acc)                 | MNLI (m/mm) (acc)   | (F1)  |
|        |                |                             |                     | STS-B |
| ALBERT | 69.1           | 97.1                        | 91.391.0            | 90.2  |
|        | 88.2           |                             |                     |       |
| RoBERTa | 67.8           | 96.7                        | 90.850.2            | 98.9  |
|        | 88.2           |                             |                     |       |
| XLNet  | 67.8           | 96.8                        | 92.098.9            | 93.6  |
|        | 89.6           |                             |                     |       |
| BiLSTM+ELMo+Attn | 36.0            | 90.4                        | 76.476.1            | 79.9  |
|        |                |                             |                     | 56.8  |
| GPT    | 45.4           | 91.3                        | 82.181.4            | 88.1  |
|        |                |                             |                     | 56.0  |
| GPT and STILTs | 47.2          | 93.1                        | 80.880.6            | 87.2  |
|        |                |                             |                     | 69.1  |
| MT-DNN | 61.5           | 95.6                        | 86.286.0            | 75.5  |
|        |                |                             |                     | 90.0  |
| BERT    | 52.1           | 95.5                        | 84.685.4            | 66.4  |
|        |                |                             |                     | 88.9  |
| BERTLARGE | 60.5          | 94.9                        | 86.585.9            | 70.1  |
|        |                |                             |                     | 89.3  |

GLUE 实验结果*

| Model | EM | F1  |
|-------|----|-----|
| #1 BERT + DAE + AoA | 85.9 | 88.6 |
| #2 BERT + NAC | 85.2 | 87.9 |
| #3 BERT + NAC + SST | 85.2 | 87.7 |
| U-Net (Sun et al., 2018) | 69.2 | 75.6 |
| BERT + ELMo + Semilar (Her et al., 2018) | 71.5 | 74.2 |

GLUE 实验结果*

| Model | Dev | Test |
|-------|-----|------|
| BERT | 90.8 | 90.7 |
| BERTLARGE | 91.7 | 91.6 |
| SemBERT | 92.3 | 91.6 |

SQuAD: The best among all the published work.
https://nlp.stanford.edu/projects/squad/

SQuAD2.0: The best among all the published work.
GLUE: substantial gains over all the tasks.
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
  - $P_i$: ancestor node set for each $i$th word; $M$: SDOI mask
    $$M[i, j] = \begin{cases} 
    1, & \text{if } j \in P_i \text{ or } j = i \\
    0, & \text{otherwise.}
    \end{cases}$$

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 | F1 | Test EM | 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 |
| BERT†_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  |
|-------|--------|--------|-------|
| Turkers | 85.1 | 69.4 | 73.3 |
| Ceiling | 95.4 | 94.2 | 94.5 |

**Leaderboard**

| Model | RACE-M | RACE-H | RACE  |
|-------|--------|--------|-------|
| DCMN  | 77.6   | 70.1   | 72.3  |
| BERT†_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** |

**Leaderboard**

1. XLNet + DAF + Verifier (ensemble) PingAN Omni-Sentic
2. XLNet + SG-Net Verifier (ensemble) Shanghai Jiao Tong University & CloudWalk
3. XLNet + SG-Net Verifier (single model) Shanghai Jiao Tong University & CloudWalk
4. BERT + DAE + AoA (ensemble) Joint Laboratory of HIT and iFLYTEK Research
5. RoBERTa (single model) Facebook AI
6. BERT + ConvLSTM + MTL + Verifier (ensemble) Layer 6 AI
7. BERT + N-Gram Masking + Synthetic Self-Training (ensemble) Google AI Language
8. XLNet (single model) Shanghai Jiao Tong University
9. SG-Net (ensemble) Shanghai Jiao Tong University
10. SemBERT (ensemble) Shanghai Jiao Tong University
11. 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.

- **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.

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

- **Answer Pointer:**
  - Pointer Network for span prediction (start and end positions):
    \[ p(a|H^r) = p(a_s|H^r)p(a_e|a_s, H^r). \]
  - 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 al., 2019)
Decoder (our work: answer verifier)

- Retro-Reader

Zhuosheng Zhang, Junjie Yang, Hai Zhao (2020). Retrospective Reader for Machine Reading Comprehension. Arxiv 2001.09694

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)

\[ \text{score}_{has} = 0.03, \text{score}_{na} = 1.73, \lambda = -0.98 \]
Outline

❖ Introductions to Machine Reading Comprehension (MRC)
❖ Development of Contextualized Language Model (CLM)
❖ Technical Methods
❖ Technical Highlights
❖ Trends and Discussions
❖ Conclusions
CLMs 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 (See 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 |
| 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 |

| 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          | -             | -             | -    |
| GPT1.0  | 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_xlarge | -   | 157GB | 235M   | 94.8          | -             | 90.2          | 90.9          | 86.5 |
| ELECTRA_large | 33B  | -    | 335M   | 94.9          | -             | 90.6          | 91.4          | -    |

- Knowledge from large-scale corpora
- Deep architectures
Decline of Matching Attention

(a) sequence-aware interaction patterns

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

| Model                                      | Matching | M | H | RACE |
|--------------------------------------------|----------|---|---|------|
| Human Ceiling Performance (Lai et al. 2017)|          | 95.4 | 94.2 | 94.5 |
| Amazon Mechanical Turk (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)                  | $M_{P,A}$ | 62.9 | 57.4 | 59.0 |
| RSM (Sun et al. 2019b)                     | $M_{P,Q,A}$ | 69.2 | 61.5 | 63.8 |
| DCMN (Zhang et al. 2019a)                  | $M_{P,Q,A}$ | 77.6 | 70.1 | 72.3 |
| OCN (Ran et al. 2019a)                     | $M_{P,Q,A}$ | 76.7 | 69.6 | 71.7 |
| BERTlarge (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)            | $[M_{P,Q,A}; M_{QA,P}]$ | 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 CLMs 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 CLMs 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.
Outline

❖ Introductions to Machine Reading Comprehension (MRC)
❖ Development of Contextualized Language Model (CLM)
❖ Technical Methods
❖ Technical Highlights
❖ Trends and Discussions
❖ Conclusions
Interpretability of Human-parity Performance

- What kind of knowledge or reading comprehension skills the systems have grasped?
- For CLM 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
Complex Reasoning

- The progress from match-based “reading” to deep “comprehension”
- Require intelligent behavior and reasoning, instead of shallow pattern matching.
  - Multi-hop QA
  - Open-domain QA
  - Conversational Reasoning
  - Commonsense QA
  - Table QA
  - ……

- Technical trend: **Graph Neural Network** (GNN)
  - Injecting extra commonsense from knowledge graphs
  - Modeling entity relationships
  - Graph-attention can be considered as a particular case of self-attention as that used in CLMs.
Most current MRC systems are based on the hypothesis of **given passages** as reference.

Real-world MRC applications: the reference documents, are always **lengthy and detail-riddled**.

Recent LM based models work slowly or even unable to process long texts.

Potential Solution:

- Selecting relevant information
- Knowledge compression
- Hardcore: Training encoders that can handle long documents, using more resources……
Other Trends and Challenges

- Some languages do not have high-quality MRC datasets.
  - transferring the well-trained English MRC models through domain adaptation
  - training semi-supervised or multilingual MRC systems

- 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.

- Deeper But Efficient Network
  - Training small but effective models
  - Rapid and accurate reading comprehension solving ability for real-world deployment
Outline

❖ Introductions to Machine Reading Comprehension (MRC)
❖ Development of Contextualized Language Model (CLM)
❖ Technical Methods
❖ Technical Highlights
❖ Trends and Discussions
❖ Conclusions
Conclusion

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

Paper Link: https://arxiv.org/abs/2005.06249
Codes: https://github.com/cooelf/AwesomeMRC
Slides: http://bcmi.sjtu.edu.cn/~zhangzs/slides/mrc_seminar.pdf
Thank You!

Homepage: http://bcmi.sjtu.edu.cn/~zhangzs
E-mail: zhangzs@sjtu.edu.cn