RACE: Large-scale ReAding Comprehension Dataset From Examinations

Guokun Lai* Qizhe Xie* Hanxiao Liu Yiming Yang
Eduard Hovy

Carnegie Mellon University, Language Technologies Institute

Outline

Carnegie Mellon University

Introduction

Carnegie Mellon University

- Machine comprehension evaluates the ability to answer questions related to a document
- Several large-scale datasets have been proposed: CNN/Daily Mail, SQuAD, Who did What, Children's Book Test, NewsQA, MS MARCO, etc
  - Questions are relatively easy: Candidate answers are extracted from the context
  - Answers and questions are noisy: crowd-sourced or automatically-generated
  - Topic coverages are biased
- RACE dataset:
  - Collected from exams to evaluate human students' reading comprehension ability
  - Designed by human experts
  - Substantially more difficult
  - Ensured quality and broad topic coverage

Related Work

Carnegie Mellon University

- MCTest:
  - Similar question and candidate answer forms with RACE
  - High quality, small scale
- Cloze-style datasets:
  - CNN/Daily Mail (Hermann et al., 2015), Children's Book Test (Hill et al., 2015), Who Did What (Onishi et al., 2016)
  - Best model's performance is close to human's performance
- Datasets with span-based answers
  - SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2016), MS MARCO (Nguyen et al., 2016), TrivialQA (Joshi et al., 2017), QUASAR (Dhingra et al., 2017)
  - More challenging due to a larger answer space, but still limits the possible types of questions
- Datasets from Examinations
  - AI2 Elementary School Science Questions dataset (Khashabi et al., 2016), NTCIR QA Lab (Shibuki et al., 2014), etc
  - Insufficient data to train deep learning models

Outline

Carnegie Mellon University

RACE: Collection Methodology

Carnegie Mellon University

- Collect raw data from three free websites in China that have English examinations designed for middle school and high school students
  - Two difficulty levels: RACE-M from middle school exams and RACE-H from high school exams
  - RACE-M has a smaller vocabulary and has fewer reasoning questions
- Data filtering: remove incorrect format, not self-contained, duplication
  - Before cleaning: 137,918 passages and 519,878 questions
  - After cleaning: 27,933 passages and 97,687 questions (4/5 removed)
- Use OCR to recognize answers that are stored as images

Question Type Analysis

Carnegie Mellon University

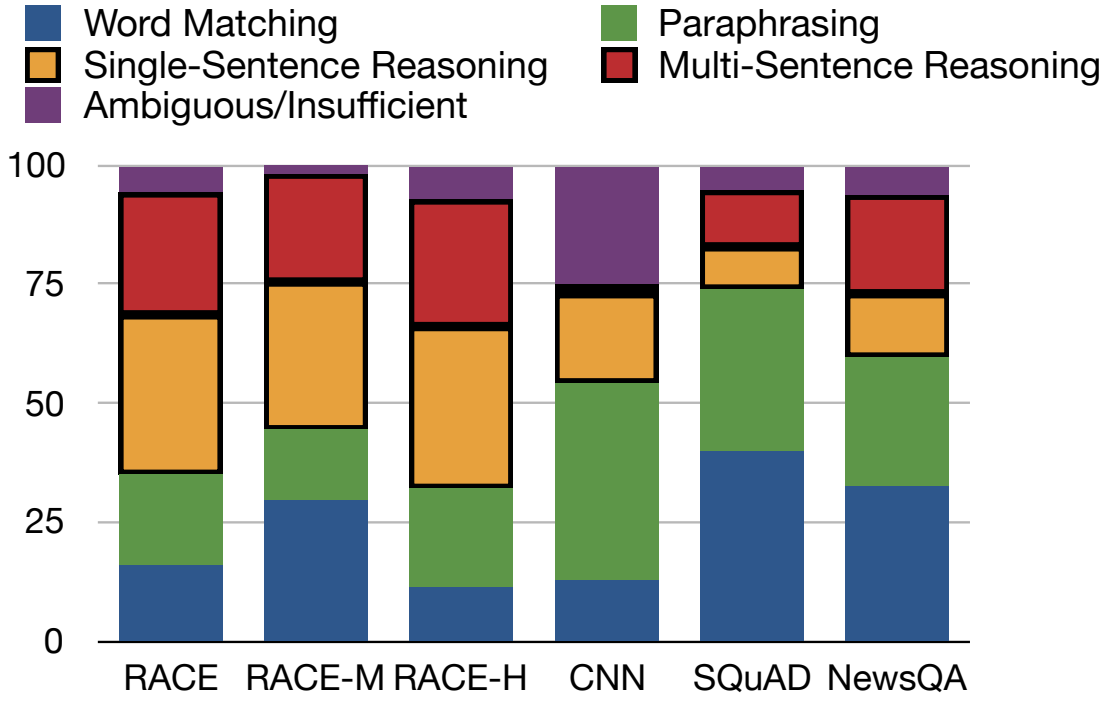

Figure 1: Question type statistics. RACE has 59.2% reasoning questions while SQuAD has 20.5% reasoning questions

- Word matching: exact match
- Paraphrasing: paraphrase or entailment
- Single-sentence reasoning: incomplete information or conceptual overlap
- Multi-sentence reasoning: synthesizing information from multiple sentences
- Insufficient/Ambiguous: no answer or the answer is not unique

Subdividing Reasoning Question Types

Carnegie Mellon University

- Detail reasoning: details of the passage
- Whole-picture reasoning: comprehension of the entire story
- **Passage summarization** (Not introduced before): summarization of the passage
- **Attitude analysis** (Not introduced before): opinions/attitudes of the author towards something
- World knowledge: external knowledge such as simple arithmetics

A Sample Passage from RACE-M

Carnegie Mellon University

Passage: Do you love holidays but hate gaining weight? You are not alone. Holidays are times for celebrating. Many people are worried about their weight. With proper planning, though, it is possible to keep normal weight during the holidays. The idea is to enjoy the holidays but not to eat too much. You don't have to turn away from the foods that you enjoy.
Here are some tips for preventing weight gain and maintaining physical fitness:
Don't skip meals. Before you leave home, have a small, low-fat meal or snack. This may help to avoid getting too excited before delicious foods.
Control the amount of food. Use a small plate that may encourage you to "load up". You should be most comfortable eating an amount of food about the size of your fist.
Begin with soup and fruit or vegetables. Fill up beforehand on water-based soup and raw fruit or vegetables, or drink a large glass of water before you eat to help you to feel full.
Avoid high-fat foods. Dishes that look oily or creamy may have large amount of fat. Choose lean meat. Fill your plate with salad and green vegetables. Use lemon juice instead of creamy food.

Stick to physical activity. Don't let exercise take a break during the holidays. A 20-minute walk helps to burn off extra calories.

1): Which of the following statements is WRONG according to the passage? (Question type: detail reasoning)
A.You should never eat delicious foods.
B.Drinking some water or soup before eating helps you to eat less.
C.Holidays are happy days but they may bring you weight problems.
D.Physical exercise can reduce the chance of putting on weight.

2): Which of the following can NOT help people to lose weight according to the passage? (Question type: detail reasoning)
A.Eating lean meat.
B.Creamy food.
C.Eating raw fruit or vegetables.
D.Physical exercise.

3): Many people can't control their weight during the holidays mainly because they _. (Question type: paraphrasing)
A.can't help eating too much
B.take part in too many parties
C.enjoy delicious foods sometimes
D.can't help turning away from foods.

4): If the passage appeared in a newspaper, which section is the most suitable one? (Question type: whole-picture reasoning)
A.Holidays and Festivals section
B.Health and Fitness section
C.Fashion section
D.Student Times Club section

5): What is the best title of the passage? (Question type: summarization)
A.How to avoid holiday feasting.
B.Do's and don'ts for keeping slim and fit.
C.How to avoid weight gain over holidays.
D.Wonderful holidays, boring experiences.

Outline

Carnegie Mellon University

Human and Model Performance

Carnegie Mellon University

	RACE-M	RACE-H	RACE	CNN	DM	CBT-N	CBT-C	WDW
Random	24.6	25.0	24.9	0.06	0.06	10.6	10.2	32.0
Sliding Window	37.3	30.4	32.2	24.8	30.8	16.8	19.6	48.0
Stanford AR	**44.2**	43.0	43.3	73.6	76.6	–	–	64.0
Gated Attention Reader	43.7	**44.2**	**44.1**	**77.9**	**80.9**	**70.1**	**67.3**	**71.2**
Turkers	85.1	69.4	73.3	–	–	–	–	–
Human Ceiling Performance	95.4	94.2	94.5	–	–	81.6	81.6	84

Table 1: Model and human performance on several datasets. DM, CBT and WDW denote Daily Mail, Children's Book Test and Who-did-What

- Baselines:
  - Sliding Window: A TF-IDF based matching algorithm
  - Stanford AR and Gated Attention Reader: state-of-the-art neural models
- RACE has higher human ceiling performance, which shows our data is quite clean
- RACE is harder for current models, resulting in a significant gap to be improved

Questions Type Analysis

Carnegie Mellon University

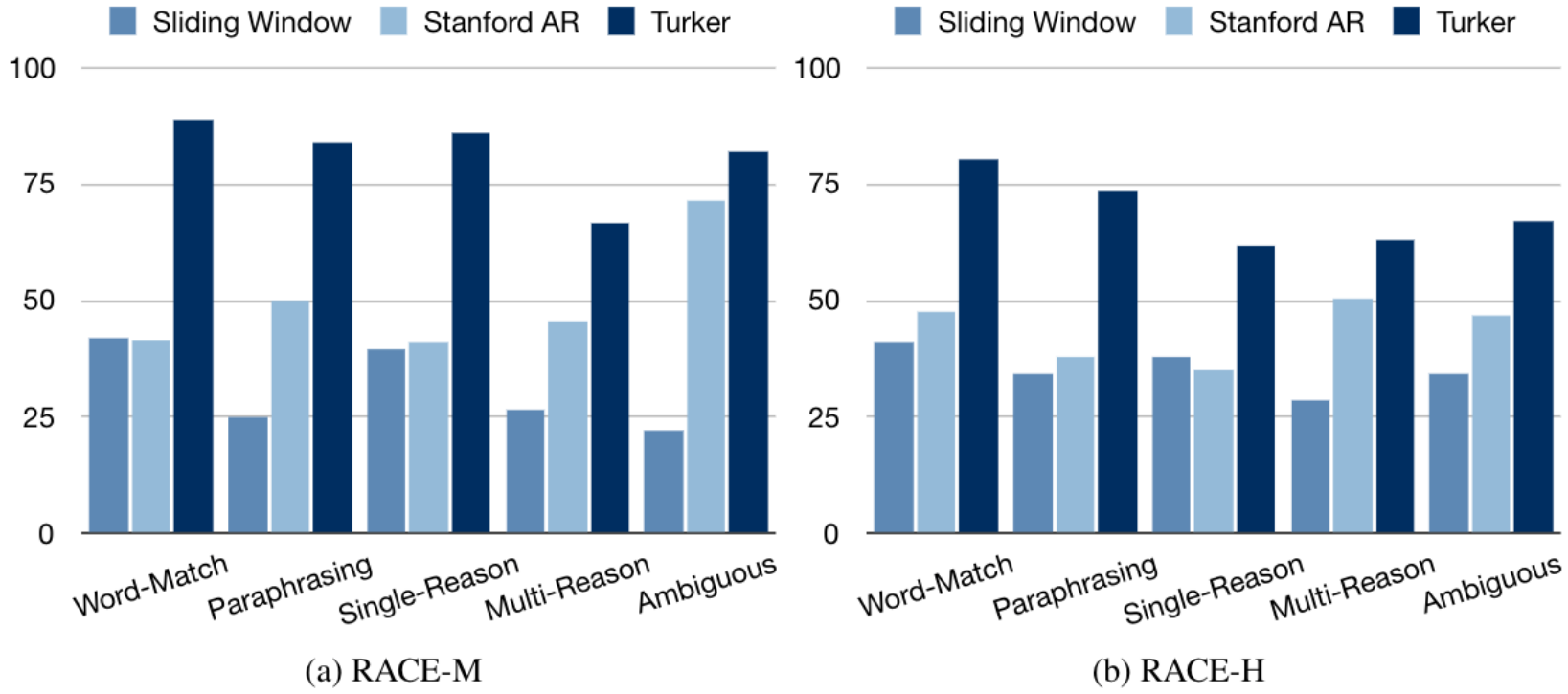

Figure 2: Accuracy of human and model on each question type

- Turkers and Sliding window are good at matching questions.
- Surprisingly, Stanford AR does not have a better performance on the matching questions

Possible Reasons

Carnegie Mellon University

- The proportion of matching question is smaller. Data-driven model will be trained to focus more on reasoning behavior
- Jia and Liang show models are easily attacked by adversarial examples. There are many adversarial examples designed by human in RACE
  - A simple example:
    Passage: Look! Here's a pencil box, it's orange, it's my pencil box, it's on the desk. Look! This is a pen, it's black. And this is an eraser, it's blue and white. They're both in the pencil box. This is a ruler, it's red, it's on the pencil box. That is a ruler, too. It's yellow. It's in the drawer. Where's my math book? Ah, it's there, under the sofa.
    Questions:
    1. The pencil box is _.
    A. yellow B. white C. blue **D. orange**
    2. Where is my English book?
    A. Under the sofa. B. On the desk. **C. Sorry, I don't know.**
    D. On the sofa.

Conclusion

Carnegie Mellon University

- RACE is Designed by human experts: Collected from exams to test human's reading comprehension ability
- Substantially more difficult than existing datasets: RACE has 59.2% reasoning questions while SQuAD has 20.5% reasoning questions
- A rich type of questions and broad coverage in various article domains
- Large-scale: nearly 28,000 passages and 100,000 questions
- Significant gap between state-of-the-art models (44%) and the ceiling human performance (95%)
- We hope this dataset will stimulate the development of more advanced machine comprehension models.