Repurposing Entailment for Multi-Hop Question Answering Tasks

Harsh Trivedi
Heeyoung Kwon
Niranjan Balasubramanian

Tushar Khot
Ashish Sabharwal
What is this talk about?

- Promise of entailment for QA
What is this talk about?

- Promise of entailment for QA
- Framing Multihop QA with entailment
What is this talk about?

- Promise of entailment for QA
- Framing Multihop QA with entailment
- Breaking and combining pre-trained entailment models
What is this talk about?

- Promise of entailment for QA
- Framing Multihop QA with entailment
- Breaking and combining pre-trained entailment models
- Improvements we get on two Multi-hop QA datasets
What is this talk about?

- **Promise of entailment for QA**
- Formulating Multihop QA with entailment
- Breaking and combining pre-trained entailment models
- Improvements we get on two Multi-hop QA datasets
Promise of Textual Entailment

Textual entailment has been one of the cornerstones of language processing.

The PASCAL Recognising Textual Entailment Challenge

Ido Dagan1, Oren Glickman2, and Bernardo Magnini2

1 Bar Ilan University, Ramat Gan 52900, Israel
2 ITC-irst, 38100 Trento, Italy

http://www.cs.technion.ac.il/~dagan/
http://people.irst.it/magnini.html

Abstract. This paper describes the PASCAL Network of Excellence first Recognising Textual Entailment (RTE-1) Challenge benchmark1. The RTE task is defined as recognizing, given two text fragments, whether the meaning of one text can be inferred (entailed) from the other. This application-independent task is suggested as capturing major inferences about the variability of semantic expression which are commonly needed across multiple applications. The Challenge has raised noticeable attention in the research community, attracting 17 submissions from diverse groups, suggesting the generic relevance of the task.

1 Introduction

1.1 Rational

A fundamental phenomenon of natural language is the variability of semantic expression, where the same meaning can be expressed by, or inferred from, different texts. This phenomenon may be considered as the dual problem of language ambiguity, together forming the many-to-many mapping between language expressions and meanings. Many natural language processing applications, such as Question Answering (QA), Information Extraction (IE), (multi-document) summarization, and machine translation (MT) evaluation, need a model for this variability phenomenon in order to recognize that a particular target meaning can be inferred from different text variants.

Dagan et. al, 2005
Promise of Textual Entailment

Textual entailment has been one of the cornerstones of language processing.

It was motivated for several tasks:

- Question Answering
- Information Extraction
- Document Summarization
- Machine Translation
- Text Simplification
- ...

---

Dagan et. al, 2005
Entailment-based Question Answering

Many pre-neural attempts

Methods for Using Textual Entailment in Open-Domain Question Answering

Sandra Harabagiu and Andrew Hickl
Language Computer Corporation
1701 North Collins Boulevard
Richardson, Texas 75080 USA
sanda@languagecomputer.com

Abstract

Work on the semantics of questions has argued that the relation between a question and its answer can be cast in terms of logical entailment. In this paper, we demonstrate how computational systems designed to recognize textual entailment can be used to enhance the accuracy of current open-domain automatic question answering (QA) systems. In our experiments, we show that when textual entailment information is used to either filter or rank answers returned by a QA system, accuracy can be increased by as much as 20% overall.

An Entailment-Based Approach to the QA4MRE Challenge

Peter Clark1, Phil Harrisson1, Xuchen Yao2
1 Vulcan Inc, Seattle, WA 98104 (peterclark@vulcan.com)
2 Johns Hopkins Univ, Baltimore, Maryland 21218 (xuchen@jhu.edu)

Abstract

This paper describes our entry to the 2012 QA4MRE Main Task (English dataset). The QA4MRE task poses a significant challenge as the expression of knowledge in the question and answer (in the document) typically substantially differs. Ultimately, one would need a system that can perform full machine reading — creating an internal model of the document’s meaning — to achieve high performance. Our approach is a preliminary step toward this, based on estimating the likelihood of textual entailment between sentences in the text, and the question Q and each candidate answer A. We first treat the question Q and each answer A, independently, and find sets of sentences S_Q and S_A that each plausibly entail (the target of) Q or one of the A, respectively. We then search for the closest (in the document) pair of sentences \( S_Q \) and \( S_A \) in these sets, and conclude that the answer \( A \) entailed by \( S_A \) in the closest pair is the answer. This approach assumes coherent discourse, i.e., that sentences close together are usually “talking about the same thing”, and thus conveying a single idea (namely an expression of the QA-pair).

In QA4MRE it is hard to “prove” entailment, as a candidate answer \( A \) may be expressed using a substantially different wording in the document, over multiple sentences, and only partially (as some aspects of \( A \) may be left implicit in the document, to be filled in by the reader). As a result, we instead estimate the likelihood of entailment (that a sentence \( S \) entails \( A \)) by looking for evidence.

Entailment-based Question Answering for Structured Data

Bogdan Sacalănu, Constantin Orasan, Christian Sparck-Jensen, Shiyuan Ou, Oscar Ferrandez, Milen Kouylekov and Matteo Negri

RIULP, University of Wolverhampton / Wolverhampton, UK
Fondazione Bruno Kessler (FBK) / Trento, Italy
University of Alicante / Alicante, Spain

Abstract

This paper describes a Question Answering system which retrieves answers from structured data regarding cinema and movies. The system represents the first prototype of a multilingual and multimodal QA system for the domain of tourism. Based on specially designed domain ontology and using Textual Entailment as a means for semantic inference, the system can be used in both monolingual and cross-language settings with slight adjustments for new input languages.

1 Introduction

Question Answering over structured data has entailment. The remainder of the paper is structured as follows: Section 2 presents the concept of entailment-based question answering; Section 3 describes our prototype which implements this concept; A brief evaluation is presented in Section 4, followed by conclusions in Section 5.

2 Entailment-based QA

Recently Textual Entailment (TE) has been proposed as a unifying framework for applied semantics (Dagan and Glickman, 2004), where the need for an explicit representation of a mapping between linguistic objects and data objects can be, at least partially, bypassed through the definition of semantic inferences at a textual level. In this framework, a text \( T \) is said to entail a hypothesis \( H \) if the meaning of \( H \) can be derived from the meaning of \( T \).
Revisiting promise of Textual Entailment for QA

Large Datasets

- SNLI (Bowman et. al, ‘15)
- MultiNLI (Williams et. al, ‘17)
- SciTail (Khot et. al, ‘18)
- JOCI (Zhang et. al, ‘17)

Neural Models

- Entailment with attention (Rocktäschel et. al, ‘15)
- Decomposable Attention (Parikh et. al, ‘16)
- ESIM (Chen et. al, ‘16)
- BiLateral Matching (Wang et. al, ‘17)
Revisiting promise of Textual Entailment for QA

Can the textual entailment models trained on these large datasets be used for Question Answering?
How to use pre-trained Entailment Model ($f_e$) for QA?
How to use pre-trained Entailment Model ($f_e$) for QA?

How to build an effective QA Model using a pre-trained Entailment Model?
What is this talk about?

- Promise of entailment for QA
- **Framing Multihop QA with entailment**
- Breaking and combining pre-trained entailment models
- Improvements we get on two Multi-hop QA datasets
Framing QA with Entailment: Simple Example

Question

Q: Where was Facebook launched?
(A) Cambridge, (B) Silicon Valley, (C) Boston
Framing QA with Entailment: Simple Example

Q: Where was Facebook launched?
(A) Cambridge, (B) Silicon Valley, (C) Boston

P: Facebook was founded in 2004 in Cambridge.
Framing QA with Entailment: Simple Example

Question

Q: Where was Facebook launched?
(A) Cambridge, (B) Silicon Valley, (C) Boston

Textual Knowledge

P: Facebook was founded in 2004 in Cambridge.

Answer Hypothesis

H: Facebook was launched in Cambridge.
Framing QA with Entailment: Simple Example

Question

Q: Where was Facebook launched?
(A) Cambridge, (B) Silicon Valley, (C) Boston

Textual Knowledge

P: Facebook was founded in 2004 in Cambridge.

Answer Hypothesis

H: Facebook was launched in Cambridge.

Entailment Question:

Does the information in P entail H?
Framing QA with Entailment: Simple Example

Question: Where was Facebook launched?
(A) Cambridge, (B) Silicon Valley, (C) Boston

Textual Knowledge: Facebook was founded in 2004 in Cambridge.

Answer Hypothesis: Facebook was launched in Cambridge.

Entailment Question: Does the information in P entail H?

Use trained Entailment model? 😊

In most entailment datasets, entailment is to be decided from single sentence.
Framing QA with Entailment: Simple Example

Q: Where was Facebook launched?
(A) Cambridge, (B) Silicon Valley, (C) Boston

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.

Entailment Question: Does the information in \{P1, P2, P3, P4\} entail H?
Question: Where was Facebook launched? (A) Cambridge, (B) Silicon Valley, (C) Boston

Textual Knowledge:

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.

Answer Hypothesis: H: Facebook was launched in Cambridge.

Entailment Question: Does the information in \{P1, P2, P3, P4\} entail H?
Mismatch: Textual Entailment & Question Answering

How Textual Entailment models are trained
Reason over **Single** Sentence

Mismatch

How we want to use them for Question Answering
Reason over **Multiple** Sentences
Multi Sentence (Hop) Reasoning Requirements
## Multi Sentence (Hop) Reasoning Requirements

**Question:** Where was Facebook launched?  
(A) Cambridge, (B) Silicon Valley, (C) Boston

**Textual Knowledge:**  
- **P1:** Facebook was launched at Harvard University.  
- **P2:** Facebook headquarters was set up in Silicon Valley.  
- **P3:** Harvard University is at Cambridge, Massachusetts.  
- **P4:** Harvard is only a few miles from Boston.

**Answer Hypothesis:** Facebook was launched in **Cambridge**.

**Entailment Question:** Does the information in **{P1, P2, P3, P4}** entail **H**?
Multi Sentence (Hop) Reasoning Requirements

Question: Where was Facebook launched?
(A) Cambridge, (B) Silicon Valley, (C) Boston

Textual Knowledge:
P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.

Answer Hypothesis: Facebook was launched in Cambridge.

Entailment Question: Does the information in \{P1, P2, P3, P4\} entail H?

Need to be careful about distracting information.
How to use pre-trained $f_e$ for QA?

Textual Knowledge $\rightarrow$ Pretrained Entailment Model $\rightarrow$ QA Model $\rightarrow$ Answer Choice

Question $\rightarrow$ X $\rightarrow$ Correct
Idea 1: Aggregate Independent Decisions

H: Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.
Idea 1: Aggregate Independent Decisions

**H**: Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.
Idea 1: Aggregate Independent Decisions

H: Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.

Does any entail (✓)?
Idea 1: Aggregate Independent Decisions

**H**: Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.

Can’t aggregate textual information from multiple sentences.
Facebook was launched at Harvard University. Facebook headquarters was set up in Silicon Valley. Harvard University is at Cambridge, Massachusetts. Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
Idea 2: Concatenate Premises

**H:** Facebook was launched in **Cambridge**.

Facebook was launched at Harvard University. Facebook headquarters was set up in **Silicon Valley**. Harvard University is at **Cambridge**, Massachusetts. Harvard is only a few miles from **Boston**.

Can’t avoid **distracting information**
Challenges of using \( f_e \) directly

**Aggregate Independent Decisions**

- **H**: Facebook was launched in **Cambridge**.
- **P1**: Facebook was launched at Harvard University.
- **P2**: Facebook headquarters was set up in **Silicon Valley**.
- **P3**: Harvard University is at **Cambridge**, Massachusetts.
- **P4**: Harvard is only a few miles from **Boston**.

Does any entail (✓)?

**Concatenate Premises**

- **H**: Facebook was launched in **Cambridge**.

Facebook was launched at Harvard University. Facebook headquarters was set up in **Silicon Valley**. Harvard University is at **Cambridge**, Massachusetts. Harvard is only a few miles from **Boston**.

Can’t **aggregate information** from multiple sentences.

Can be confused by **distracting** information.
Our proposal to address these challenges
Our proposal to address these challenges

H: Facebook was launched in Cambridge.

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

Relevance Module

<table>
<thead>
<tr>
<th>Statement</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook was launched in Cambridge.</td>
<td>0.4</td>
</tr>
<tr>
<td>Facebook was launched at Harvard University.</td>
<td>0.4</td>
</tr>
<tr>
<td>Facebook headquarters was set up in Silicon Valley.</td>
<td>0.1</td>
</tr>
<tr>
<td>Harvard University is at Cambridge, Massachusetts.</td>
<td>0.1</td>
</tr>
<tr>
<td>Harvard is only a few miles from Boston.</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Our proposal to address these challenges

H: Facebook was launched in Cambridge.

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

Relevance
Module

Help avoid distracting information.
Our proposal to address these challenges

\( H: \) Facebook was launched in \textbf{Cambridge}.

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

\( H: \) Facebook was launched in \textbf{Cambridge}.

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.
Our proposal to address these challenges

H: Facebook was launched in Cambridge.
Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.
Our proposal to address these challenges

\( H: \) Facebook was launched in Cambridge.

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

\( H: \) Facebook was launched in Cambridge.

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

\( \Rightarrow \) Relevance module can be seen as doing entailment.

\( \Rightarrow \) relevant \( \sim \) entails

\( \Rightarrow \) So use pre-trained \( f_e \) directly.
Our proposal to address these challenges

H: Facebook was launched in Cambridge.
Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

⇒ Aggregator module can also be seen as doing entailment.
⇒ So we can leverage $f_e$.
⇒ But key differences from $f_e$:
  • takes many sentences
  • takes sentence weights
What does Aggregator Module look like?
What does Aggregator Module look like?

Takes multiple sentences as input.

Aggregator Module

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
What does Aggregator Module look like?

Generates a representation for each sentence.

H: Facebook was launched in Cambridge.
What does Aggregator Module look like?

Scales these representations with relevance weights.

Facebook was launched at Harvard University. Facebook headquarters was set up in Silicon Valley. Harvard University is at Cambridge, Massachusetts. Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
What does Aggregator Module look like?

Aggregates to one representation.

Join

X 0.4  X 0.1  X 0.1  X 0.4

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
What does Aggregator Module look like?

Makes aggregate entailment decision.

Entails?

Join

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
What does Aggregator Module look like?

Aggregator ⇒ entailment decision.

- Facebook was launched at Harvard University.
- Facebook headquarters was set up in Silicon Valley.
- Harvard University is at Cambridge, Massachusetts.
- Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
What does Aggregator Module look like?

Aggregator ⇒ entailment decision.

We have pretrained $f_e$.

H: Facebook was launched in Cambridge.
What does Aggregator Module look like?

Aggregator $\Rightarrow$ entailment decision.

We have pretrained $f_e$.

How to get $\square$ and $\square$ from it?

Entails?

Join

Facebook was launched at Harvard University.

Facebook headquarters was set up in Silicon Valley.

Harvard University is at Cambridge, Massachusetts.

Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
What is this talk about?

- Promise of entailment for QA
- Formulating Multihop QA with entailment
- **Breaking and combining pre-trained entailment models**
- Improvements we get on two Multi-hop QA datasets
Neural Models are Multi-Layered
Entailment model ($f_e$) can be cut into parts
Entailment model ($f_e$) can be cut into parts
Aggregator Module: Incorporate weights (Join)

Entails?

H: Facebook was launched in Cambridge.

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.
Aggregator Module: Incorporate weights (Join)

Lower

Facebook was launched at Harvard University.

Lower

Facebook headquarters was set up in Silicon Valley.

Lower

Harvard University is at Cambridge, Massachusetts.

Lower

Harvard is only a few miles from Boston.

Entails?

H: Facebook was launched in Cambridge.
Aggregator Module: Incorporate weights (Join)

Upper

Entails?

Join

Lower applied paragraph-wise

Facebook was launched at Harvard University.

Facebook headquarters was set up in Silicon Valley.

Harvard University is at Cambridge, Massachusetts.

Harvard is only a few miles from Boston.

H: Facebook was launched in Cambridge.
Aggregator Module: Incorporate weights (Join)

Join scaled sentence-wise representations to a paragraph-wise representation.

Entails?

H: Facebook was launched in Cambridge.

- Facebook was launched at Harvard University.
- Facebook headquarters was set up in Silicon Valley.
- Harvard University is at Cambridge, Massachusetts.
- Harvard is only a few miles from Boston.
Cut and incorporate weights at multiple layers

... because different layers in neural stack capture different types of information
Full Aggregator Module

Incorporate **sentence** relevance weights at **multiple** levels layers.

Entails?

**Concatenate**

**Feedforward**

![Diagram showing the process of concatenating and feeding through different layers with sentence relevance weights.]
Multee (MUX) Model

Repurposed pretrained single-sentence entailment model for multi-sentence QA

H: Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard ...
P2: Facebook headquarters was set up ...
P3: Harvard University is at Cambridge, ...
P4: Harvard is only a few miles from Bos ...

Multee: MUlti Layer aggregation of TExtual Entailment Representations
Multee ( ): Choices: \( f_e \) and Join layers?

**H:** Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard ...
P2: Facebook headquarters was set up ...
P3: Harvard University is at Cambridge, ...
P4: Harvard is only a few miles from Bos ...

```
H: Facebook was launched in Cambridge.
```

Multee: Multi Layer aggregation of Textual Entailment Representations
Multee ((png)) Choices: $f_e$ and Join layers?

ESIM Stack (Chen et al, 16)
How to define Join operation?

Different Layers have different output/s and so need different Join Operations

ESIM Stack (Chen et al, 16)
How to define Join operation?

Different Layers have different output/s and so need different Join Operations.

ESIM Stack (Chen et al, 16)
Join Operations - Cross Attention Layer (CA)

Outputs sentence-wise attention matrices:

\[ H \times P_i \]
Join Operations - **Cross Attention Layer (CA)**

Expects paragraph attention matrix

\[ H \times \left( \sum_i P_i \right) \]
Join Operations - **Cross Attention Layer (CA)**

\[ \alpha_1 \times A \]

**P1:** "Alice is struggling with homework"

**P2:** "Alice went to Bob during class break"

**P3:** "Alice was helped."

**H:** "Bob helped Alice during with her homework during the class break."
Join Operations - **Cross Attention Layer (CA)**

\[ \alpha_1 \times B \]

Concatenate and Normalize Horizontally

\[ \alpha_1 \times \]
\[ \alpha_2 \times \]
\[ \alpha_3 \times \]

- P1: "Alice is struggling with homework"
- P2: "Alice went to Bob during class break"
- P3: "Alice was helped."

H: "Bob helped Alice during with her homework during the class break."
Join Operations - Cross Attention Layer (CA)

\[ \alpha_1 \times C \]

Concatenate and Normalize Horizontally

- \( \alpha_1 \times \)
- \( \alpha_2 \times \)
- \( \alpha_3 \times \)

Terms:
- \( \alpha_1 \times \) for each sentence:
  - P1: "Alice is struggling with homework"
  - P2: "Alice went to Bob during class break"
  - P3: "Alice was helped."

- \( C \) for the context:

H: "Bob helped Alice during with her homework during the class break."
Join Operations - Cross Attention Layer (CA)

Renormalize horizontally

Concatenate and Normalize Horizontally

\( \alpha_1 \times \)

\( \alpha_2 \times \)

\( \alpha_3 \times \)

P1: "Alice is struggling with homework"

P2: "Alice went to Bob during class break"

P3: "Alice was helped."

H: "Bob helped Alice during with her homework during the class break."
What is this talk about?

- Promise of entailment for QA
- Framing Multihop QA with entailment
- Breaking and combining pre-trained entailment models
- Improvements we get on two Multi-hop QA datasets
Evaluation

1. Does Multee (ertoire) help over direct usages of Entailment models (Idea 1 & 2)?
2. How Multee (ertoire) compares with other QA models?
To test this we build Mutlee:

- With ESIM entailment model.
- Pretrained on SNLI + MultiNLI entailment datasets.

Evaluation Multihop-QA Datasets

- OpenBookQA
- MultiRC
Main Results - OpenBookQA

1. Does Multee ( ) help over direct usages of Entailment models (Idea 1 & 2)?

Train / Finetune
1. SNLI + MultiNLI
2. OpenBookQA

Accuracy

- Idea 1: 47.6
- Idea 2: 42.6
- Multee: 55.8
- Our Model: 55.8
2. How Muitee (ophage) compares with other QA models?

![Bar chart showing comparison between Muitee and other QA models.](image)

- Muitee Match: 50.2
- Muitee KER: 51.4

Source: Mihaylov et al., ‘18
2. How Multee (하신 모델) compares with other QA models?

- **Data Specific Baselines**
  - Question Match: 50.2, 51.4
  - KER

- **Open AI Transformer**
  - OFT: 52, 52.8
  - OFT (ensemble)

---

1. **Large Scale LM**
2. **OpenBookQA**

- Training / Finetune
  - Mihaylov et al, ‘18
  - Radford et al, ‘18; Sun et. al ‘18
Main Results - OpenBookQA

2. How Multee (📚) compares with other QA models?

- Data Specific Baselines
  - Question Match
    - OpenBookQA: 50.2
    - KER: 51.4
  - KER

- Open AI Transformer
  - single
    - 52
  - ensemble
    - 52.8

- NAACL’19
  - Reading Strategies
    - single
      - 55.2
    - ensemble
      - 55.8

Train / Finetune

- Mihaylov et al, ‘18
- Radford et al, ‘18; Sun et. al ‘18
- Sun et. al ‘19
Main Results - OpenBookQA

2. How Multee (صدق) compares with other QA models?

- **Data Specific Baselines**
  - Question Match: 50.2
  - KER: 51.4

- **Open AI Transformer**
  - Single: 52
  - Ensemble: 52.8

- **Reading Strategies**
  - Single: 55.2
  - Ensemble: 55.8

- **NAACL’19**
  - Our Model: 55.8

- **Multee**

**Train / Fine-tune**
- **OpenBookQA**
  - Mihaylov et al, ‘18
  - Radford et al, ‘18; Sun et. al ‘18
  - Sun et. al ‘19

- **Large Scale QA (RACE)**
  - 1. Large Scale LM
  - 2. Large Scale QA (RACE)
  - 3. OpenBookQA
Main Results - MultiRC

2. How Multee ( _) compares with other QA models?

Train / Finetune

- Data Specific Baselines
  - Logistic Regression: 66.5
  - IR: 64.3

- Open AI Transformer
  - single: 69.3
  - ensemble: 70.3

- MultiRC
  1. Large Scale LM
  2. MultiRC

- SNLI + MultiNLI
  - single: 71.5
  - ensemble: 73.1

- Multee
  - Our Model: 71.7

Readers Strategies

- NAACL’19
  - single: 75
  - ensemble: 75

- MultiRC
  1. Large Scale LM
  2. Large Scale QA (RACE)
  3. MultiRC

- SNLI + MultiNLI
  1. SNLI + MultiNLI
  2. MultiRC

Khashabi et al, ‘18
Radford et al, ‘18; Sun et. al ‘18
Sun et. al ‘19
Ablations - Is Relevance Module really helpful?

**H**: Facebook was launched in **Cambridge**.

P1: Facebook was launched at Harvard ...
P2: Facebook headquarters was set up ...
P3: Harvard University is at Cambridge, ...
P4: Harvard is only a few miles from Bos ...

![Relevance Module Diagram]

Entails? **Cambridge**
Ablations - Is **Relevance Module** really helpful?

**OpenBookQA**
- Without Relevance Weight: 50.6
- With Relevance Weights: 55.8

**MultiRC**
- Without Relevance Weight: 70.3
- With Relevance Weights: 71.7

Full Model
Ablations - Is **Multi**-Layer aggregation really helpful?

H: Facebook was launched in **Cambridge**.

P1: Facebook was launched at Harvard...
P2: Facebook headquarters was set up...
P3: Harvard University is at Cambridge, ...
P4: Harvard is only a few miles from Bos ...

Relevance Module

\[ f_e \rightarrow 0.4 \rightarrow \text{Softmax} \rightarrow 0.1 \rightarrow \ldots \rightarrow \ldots \rightarrow 0.4 \rightarrow 0.1 \rightarrow f_e \]

Entails?

Cambridge
Ablations - Is Multi-Layer aggregation really helpful?

**OpenBookQA**

- **Cross Attention (CA)**: 45.8
- **Final Layer (FL)**: 51
- **CA + FL**: 55.8

**MultiRC**

- **Cross Attention (CA)**: 71.1
- **Final Layer (FL)**: 71.5
- **CA + FL**: 71.7

Both graphs show the performance metrics (Accuracy for OpenBookQA and F1m for MultiRC) across different model configurations.
Is Multee (_eof) still relevant in context of BERT (_face)?
Is Multee ( _) still relevant in context of BERT ( _)?

In theory, Multee ( _) is (mostly) Model agnostic!

```
Entails?
```

```
Linear
Final FeedForward
Concat
```

```
Max, Avg Pool
Project
Enhance
Collect P2H
```

```
Max, Avg Pool
Project
Enhance
Collect H2P
```

```
Cross Attention
BiLstm Encoder
Embedding
```

Premise

Hypothesis

\[ f_e \]
Conclusion

- Neural entailment models can be effective for QA
  - Multee with ESIM gives strong results on two Multi-Hop QA datasets.

- We repurposed entailment model for QA
  - Breaking and recombining pre-trained models can get us from TE to QA.
  - Models trained for short inputs → Uses in tasks with long inputs?

- Code is available at: http://github.com/stonyBrookNLP/multee/
Extra Slides:
Framing QA with Entailment: Simple Example

Question: Where was Facebook launched?
(A) Cambridge, (B) Silicon Valley, (C) Boston

Textual Knowledge:
- P1: Facebook was launched at Harvard University.
- P2: Facebook headquarters was set up in Silicon Valley.
- P3: Harvard University is at Cambridge, Massachusetts.
- P4: Harvard is only a few miles from Boston.

Answer Hypothesis: Facebook was launched in Cambridge.

Entailment Question: Does the information in \{P1, P2, P3, P4\} entail H?
Is Multee (☒) still relevant in context of BERT (🤖)?

<table>
<thead>
<tr>
<th></th>
<th>BERT</th>
<th>Multee</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenBookQA</td>
<td>63.8</td>
<td>55.8</td>
</tr>
</tbody>
</table>

- **Official Leaderboard**
- Super Glue Benchmark (not directly comparable)

<table>
<thead>
<tr>
<th></th>
<th>BERT</th>
<th>Multee</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiRC</td>
<td>66.2</td>
<td>69.9</td>
</tr>
</tbody>
</table>

New Results; not in paper
Trade offs of Joining at different layers.

Independent Decisions

Closer from training

Joint Decisions

Farther from training
What does Aggregator do?

1. Scale Down Information with relevance weights.
2. Check Entailment from Scaled-Down Information

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

Original Information

Facebook was launched at Harvard University.
Facebook headquarters was set up in Silicon Valley.
Harvard University is at Cambridge, Massachusetts.
Harvard is only a few miles from Boston.

Relevance Weights

X 0.4
X 0.1
X 0.4
X 0.1

Scaled-Down Information

More amenable to work with $f_e$, because it was trained with limited information.

H: Facebook was launched in Cambridge.

Entails?
Main Results - MultiRC

1. Does Multee (🔗) help over direct usages of Entailment models (Idea 1 & 2)?

![Graph showing comparison between Multee and direct usages of Entailment models.](image)

- **Train / Finetune**
  1. SNLI + MultiNLI
  2. MultiRC

- **F1m**
  - Idea 1 & 2: 70.3, 70.7
  - Multee: 71.7

- **Our Model**