Repurposing Entailment for Multi-Hop Question Answering Tasks

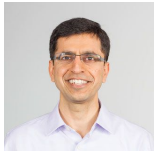
Harsh Trivedi

Heeyoung Kwon

Niranjan Balasubramanian

Tushar Khot

Ashish Sabharwal



Stony Brook University

LUNR language understanding & reasoning

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

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?

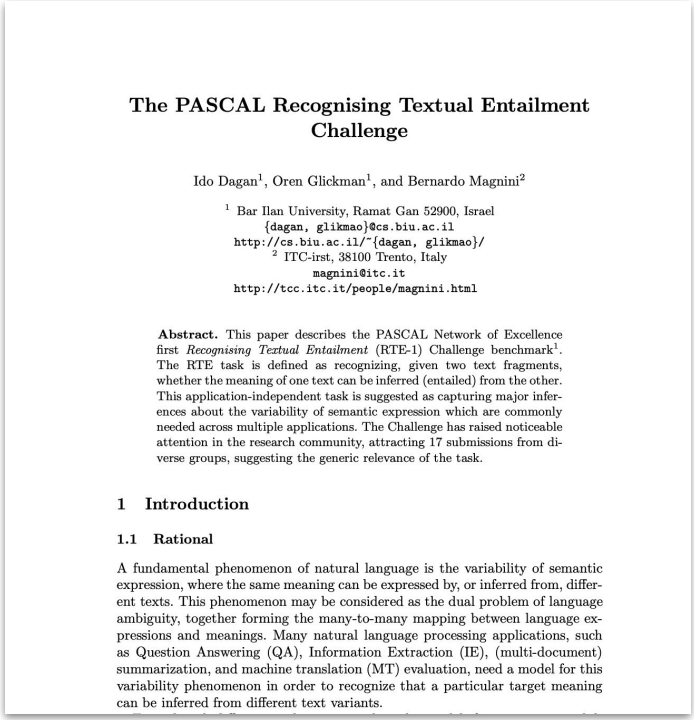
- Promise of entailment for QA
- Framing Multihop QA with entailment
- Breaking and combining pre-trained entailment models
- Improvements we get on two Multi-hop QA datasets

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 Dagan[1], Oren Glickman[1], and Bernardo Magnini[2]

[1] Bar Ilan University, Ramat Gan 52900, Israel
{dagan, glikmao}@cs.biu.ac.il
http://cs.biu.ac.il/~{dagan, glikmao}/
[2] ITC-irst, 38100 Trento, Italy
magnini@itc.it
http://tcc.itc.it/people/magnini.html

Abstract. This paper describes the PASCAL Network of Excellence first *Recognising Textual Entailment* (RTE-1) Challenge benchmark[1]. 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.

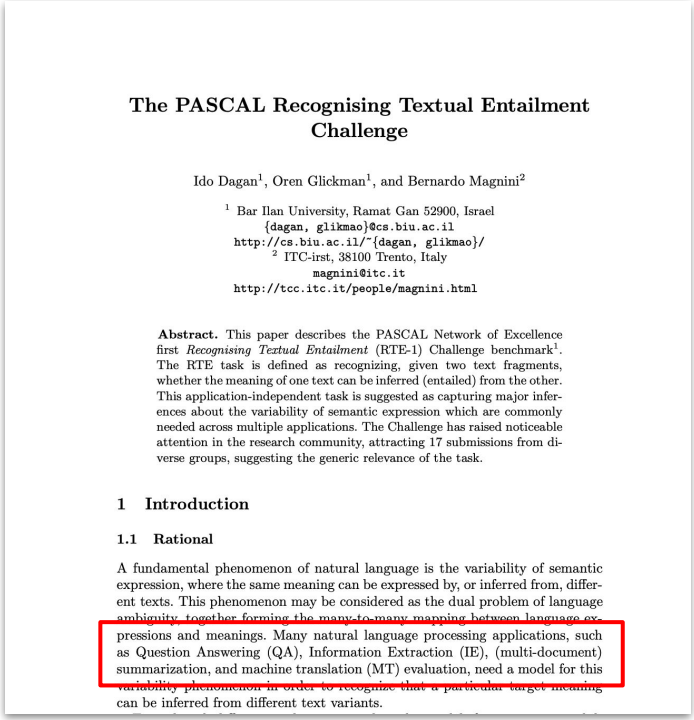
Dagan et. al, 2005

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
- …



The PASCAL Recognising Textual Entailment Challenge

Ido Dagan[1], Oren Glickman[1], and Bernardo Magnini[2]

[1] Bar Ilan University, Ramat Gan 52900, Israel
{dagan, glikmao}@cs.biu.ac.il
http://cs.biu.ac.il/~{dagan, glikmao}/
[2] ITC-irst, 38100 Trento, Italy
magnini@itc.it
http://tcc.itc.it/people/magnini.html

Abstract. This paper describes the PASCAL Network of Excellence first *Recognising Textual Entailment* (RTE-1) Challenge benchmark[1]. 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

Entailment-based Question Answering

Many pre-neural attempts

Methods for Using Textual Entailment in Open-Domain Question Answering

Sanda 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(s) 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 (Q/A) systems. In our experiments, we show that when textual entailment information is used to either filter or rank answers returned by a Q/A system, accuracy can be increased by as much as 20% overall.

better constrain the answers to individual questions (cf. (Prager et al., 2004)). While different in their approach, each of these frameworks seeks to approximate the forms of semantic inference that will allow them to identify valid textual answers to natural language questions.

Recently, the task of automatically recognizing one form of semantic inference – textual entailment – has received much attention from groups participating in the 2005 and 2006 PASCAL Recognizing Textual Entailment (RTE) Challenges (Dagan et al., 2005). [1]

As currently defined, the RTE task requires systems to determine whether, given two text fragments, the meaning of one text could be reasonably inferred, or *textually entailed*, from the meaning of the other text. We believe that systems de-

Harabagiu et al, 2006

An Entailment-Based Approach to the QA4MRE Challenge

Peter Clark[1], Phil Harrison[1], Xuchen Yao[2]

[1] Vulcan Inc, Seattle, WA 98104 ({peterc,philipha}@vulcan.com)
[2] Johns Hopkins Univ, Baltimore, Maryland 21218 (xuchen@cs.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_i. We first treat the question Q and each answer A_i independently, and find sets of sentences SQ, SA that each plausibly entail (the target of) Q or one of the A_i respectively. We then search for the closest (in the document) pair of sentences $<S_Q \in SQ, S_{Ai} \in SA>$ in these sets, and conclude that the answer A_i entailed by S_{Ai} 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 $Q+A_i$ 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 look for evidence

Clark et al, 2011

Entailment-based Question Answering for Structured Data

Bogdan Sacaleanu♣, Constantin Orasan♦, Christian Spurk♣, Shiyan Ou♦, Oscar Ferrandez♥, Milen Kouylekov♠ and Matteo Negri♠
♣LT-Lab, DFKI GmbH / Saarbrücken, Germany
♦RIILP, 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 cinemas 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.

Sacaleanu et al, 2008

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

Large Datasets

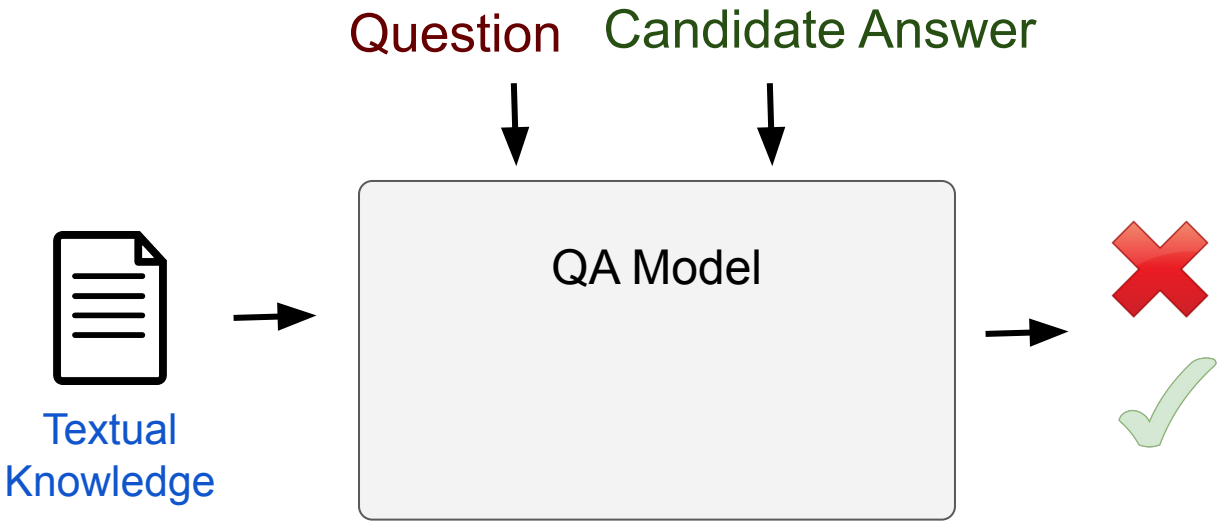
SNLI (Bowman et. al, ‘15)
MultiNLI (Williams et. al, ‘17)
SciTail (Khot et. al, ‘18)
JOCI (Zhang et. al, ‘17)
…

Neural Models

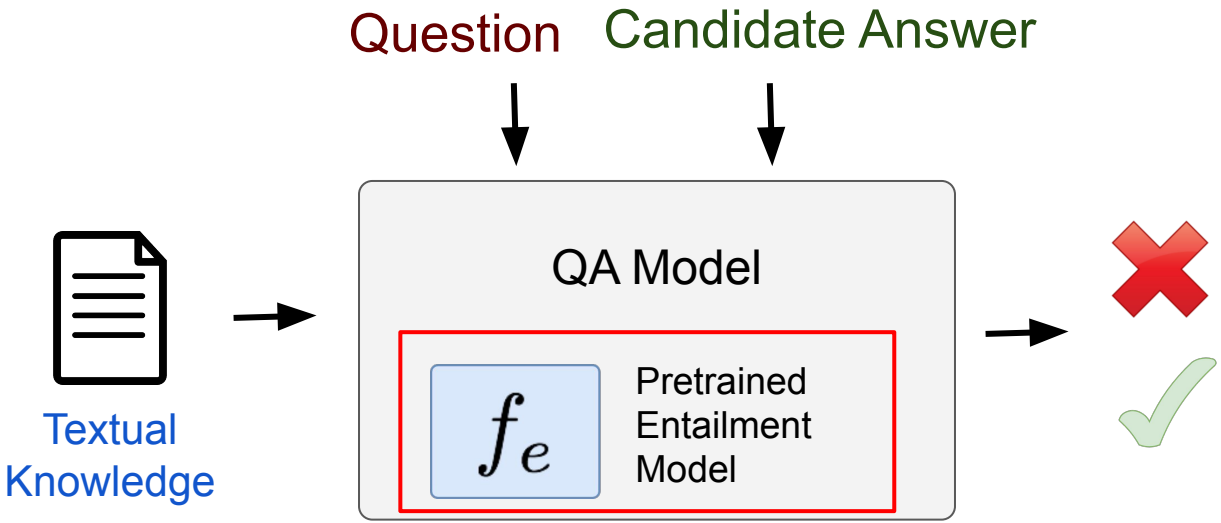
Entailment with attention (Rocktäschel et. al, ‘15)
Decomposable Attention (Parikh et. al, ‘16)
ESIM (Chen et. al, ‘16)
BiLateral Matching (Wang et. al, ‘17)
...

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

Question

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

Textual Knowledge

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

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** ?

Use trained Entailment model?

In most entailment datasets, entailment is to be decided from **single** sentence.

More Realistic

Framing QA with Entailment: ~~Simple~~ Example

Question

Q: 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** ?

More Realistic

Framing QA with Entailment: ~~Simple~~ Example

Question

Q: 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** ?

Use trained Entailment model?

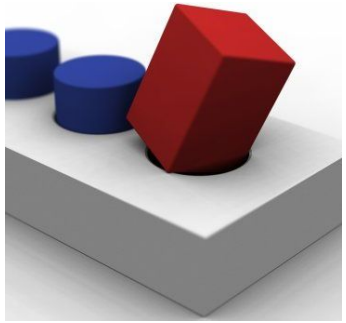
In most entailment datasets, entailment is to be decided from **single** sentence.

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

Need to aggregate information from **multiple sentences**

Question

Q: 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** ?

Multi Sentence (Hop) Reasoning Requirements

Need to be careful about distracting information.

Question

Q: 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.

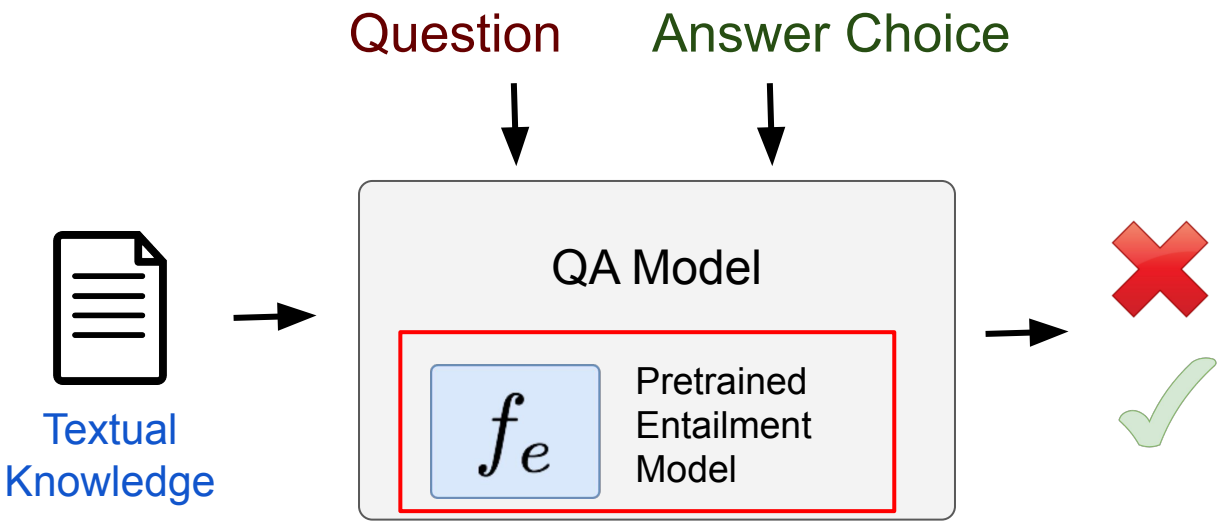
Answer Hypothesis

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

Entailment Question:

Does the information in **{P1**, **P2**, **P3**, **P4}** entail **H** ?

How to use pre-trained f_e for QA?



Idea 1: Aggregate Independent Decisions

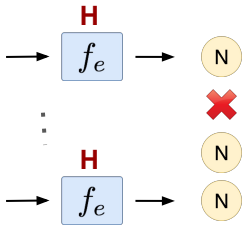
H: Facebook was launched in **Cambridge**.

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**.

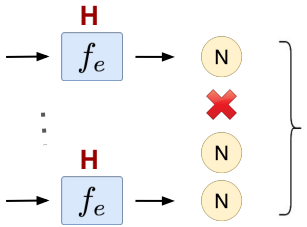
P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in **Silicon Valley**.
P3: Harvard University is at **Cambridge**, Massachusetts.
P4: Harvard is only a few miles from **Boston**.



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**.

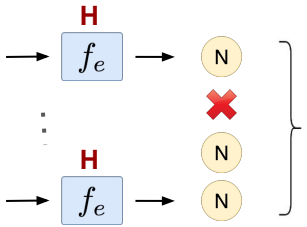


Does any entail (✓)?

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**.



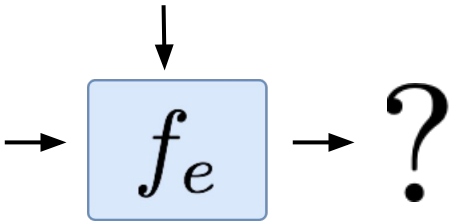
Does any entail (✓)?

Can't aggregate textual information from multiple sentences

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**.



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**.

$\rightarrow \boxed{f_e} \rightarrow ?$

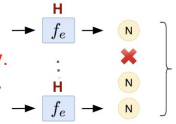
Can't avoid distracting information

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**.

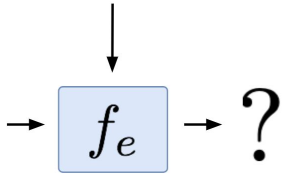


Does any entail (✓)?

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.

Challenges!

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.	Relevance Module →	0.4
Facebook headquarters was set up in Silicon Valley.		0.1
Harvard University is at Cambridge, Massachusetts.		0.4
Harvard is only a few miles from Boston.		0.1

Our proposal to address these challenges

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

Facebook was launched at Harvard University.	Relevance Module →	0.4
Facebook headquarters was set up in Silicon Valley.		0.1
Harvard University is at Cambridge, Massachusetts.		0.4
Harvard is only a few miles from Boston.		0.1

Help avoid
distracting information.

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 →

0.4
0.1
0.4
0.1

Aggregator Module

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.

→ Entails?

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 →

0.4
0.1
0.4
0.1

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 →

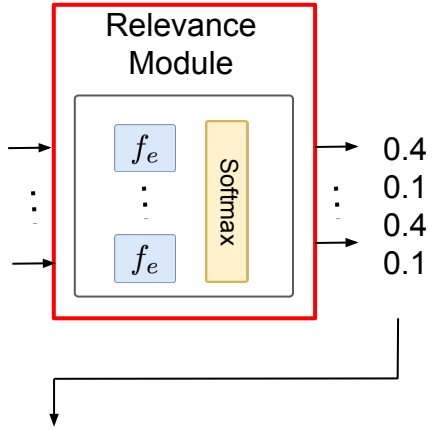
Entails?

Help aggregate **relevant information.**

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.

Entails?

⇒ Relevance module can be seen as doing entailment.

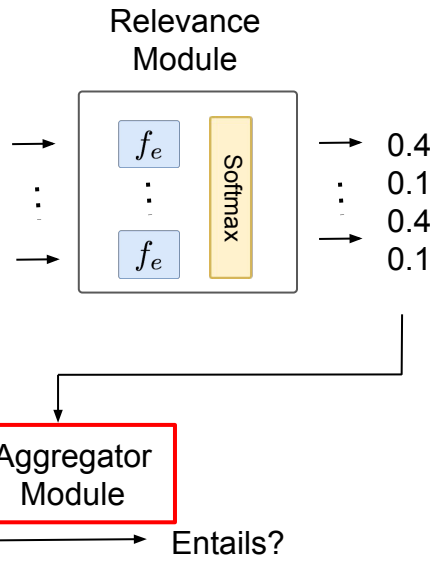
⇒ relevant $\simeq$ entails

⇒ 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.

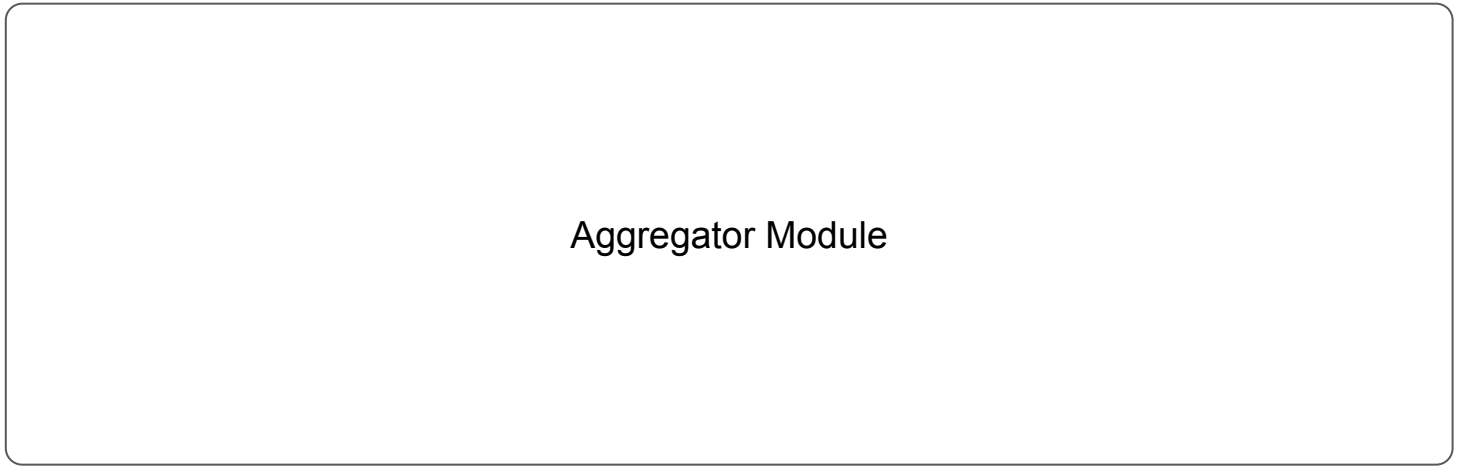
⇒ Aggregator module can also be seen as doing entailment.

⇒ 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?

Aggregator Module

Takes multiple sentences as input.

Facebook was launched at Harvard University.

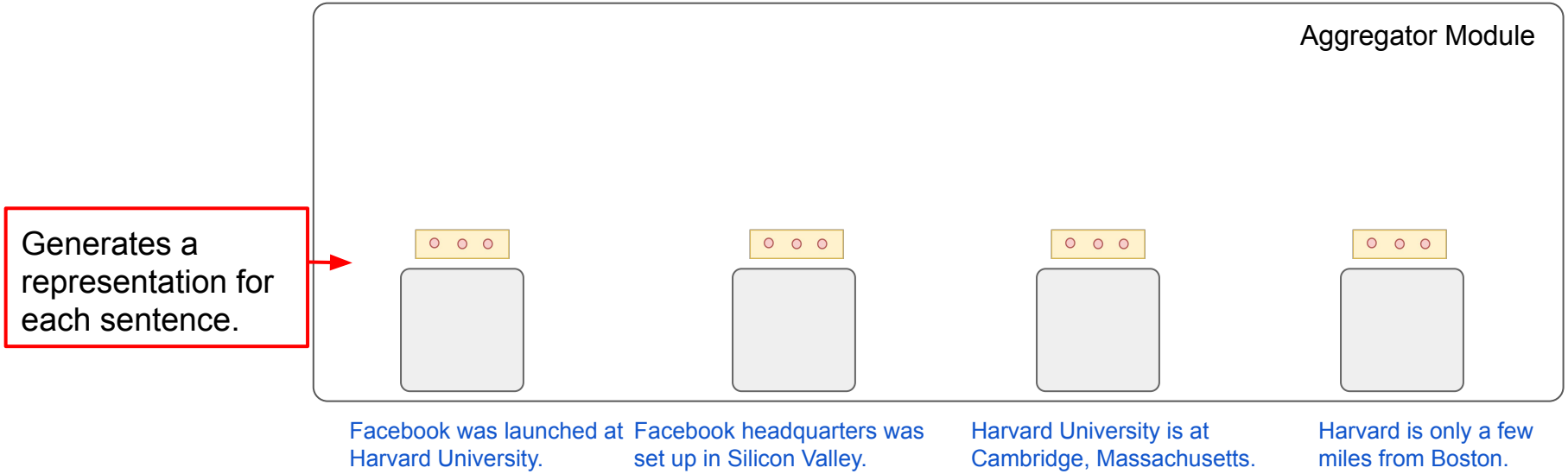
Facebook headquarters was set up in Silicon Valley.

Harvard University is at Cambridge, Massachusetts.

Harvard is only a few miles from Boston.

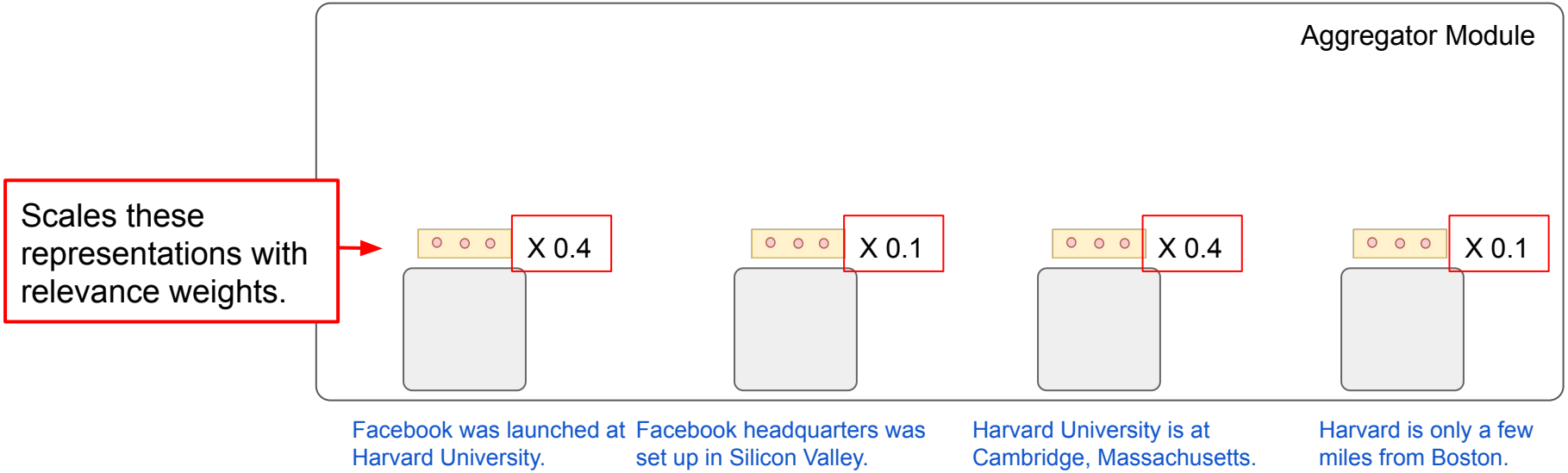
H: Facebook was launched in **Cambridge**.

What does Aggregator Module look like?



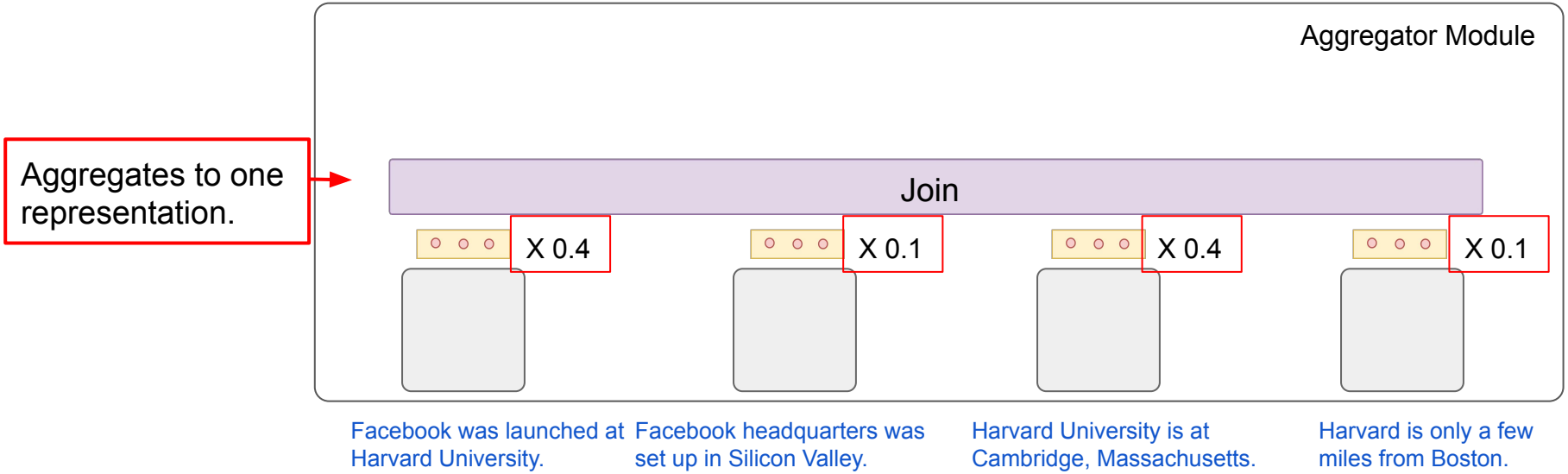
H: Facebook was launched in **Cambridge**.

What does Aggregator Module look like?



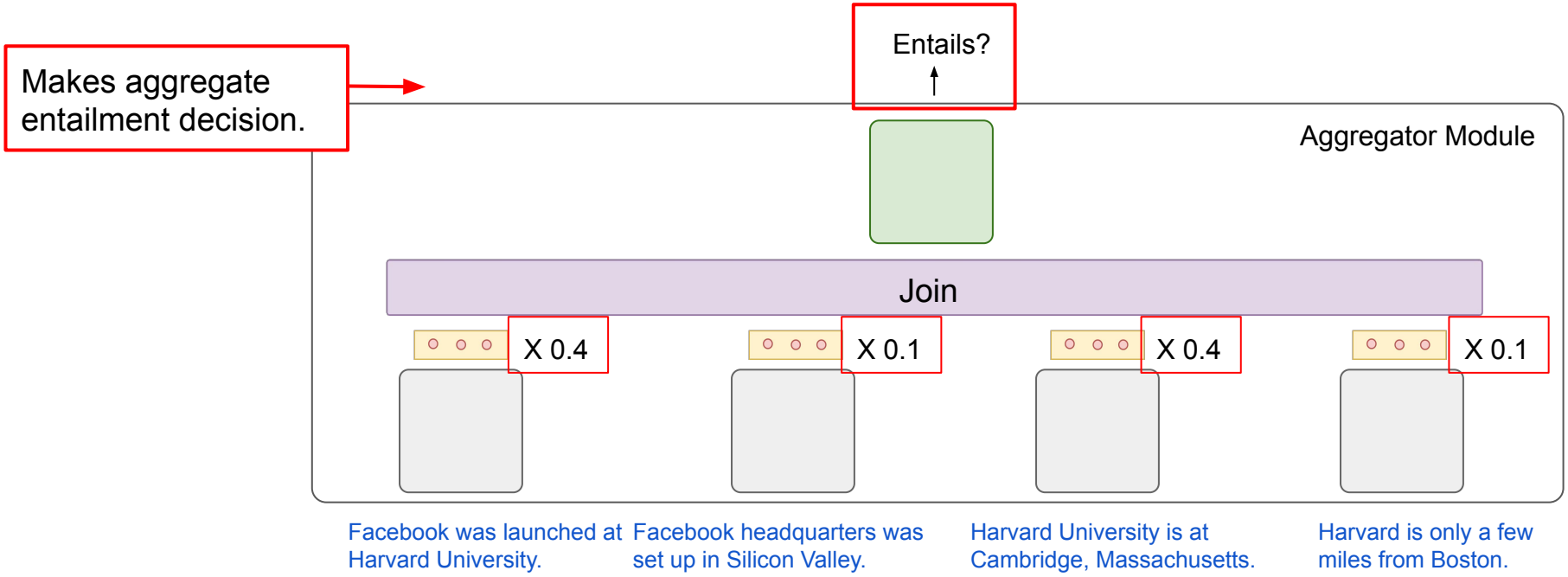
H: Facebook was launched in **Cambridge**.

What does Aggregator Module look like?



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

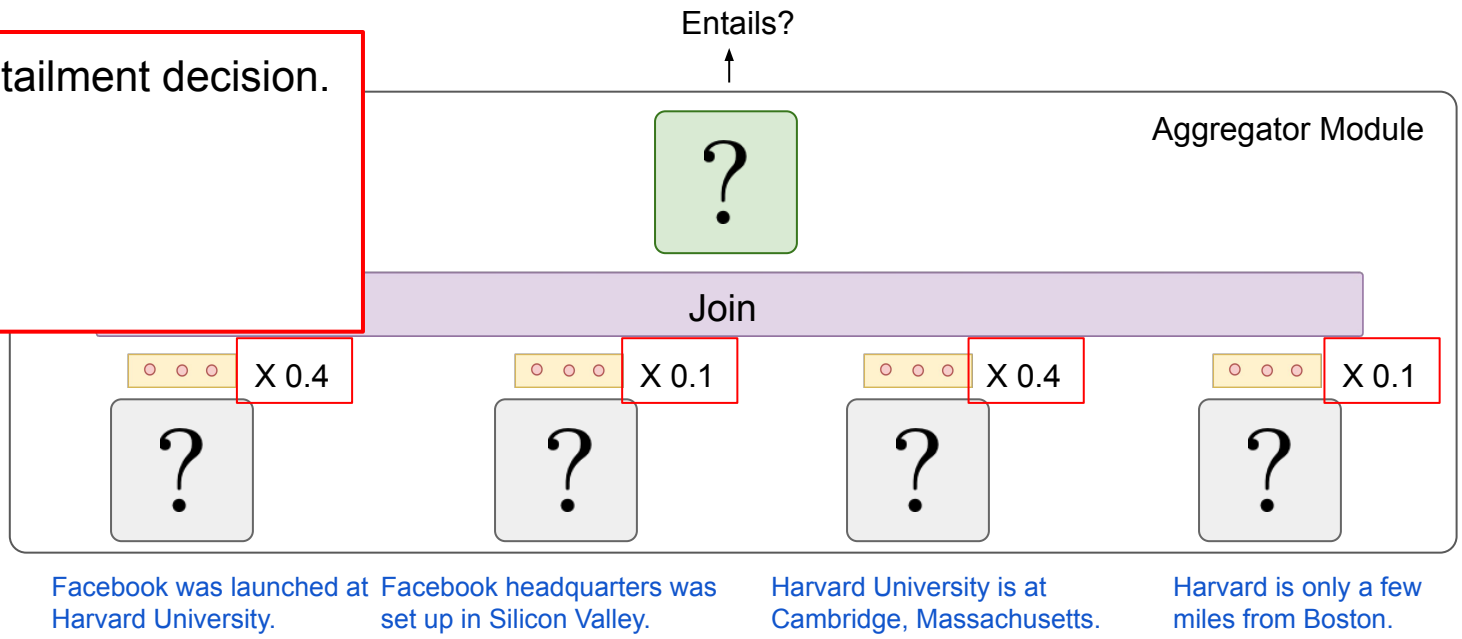
What does Aggregator Module look like?



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

What does Aggregator Module look like?

Aggregator ⇒ entailment decision.

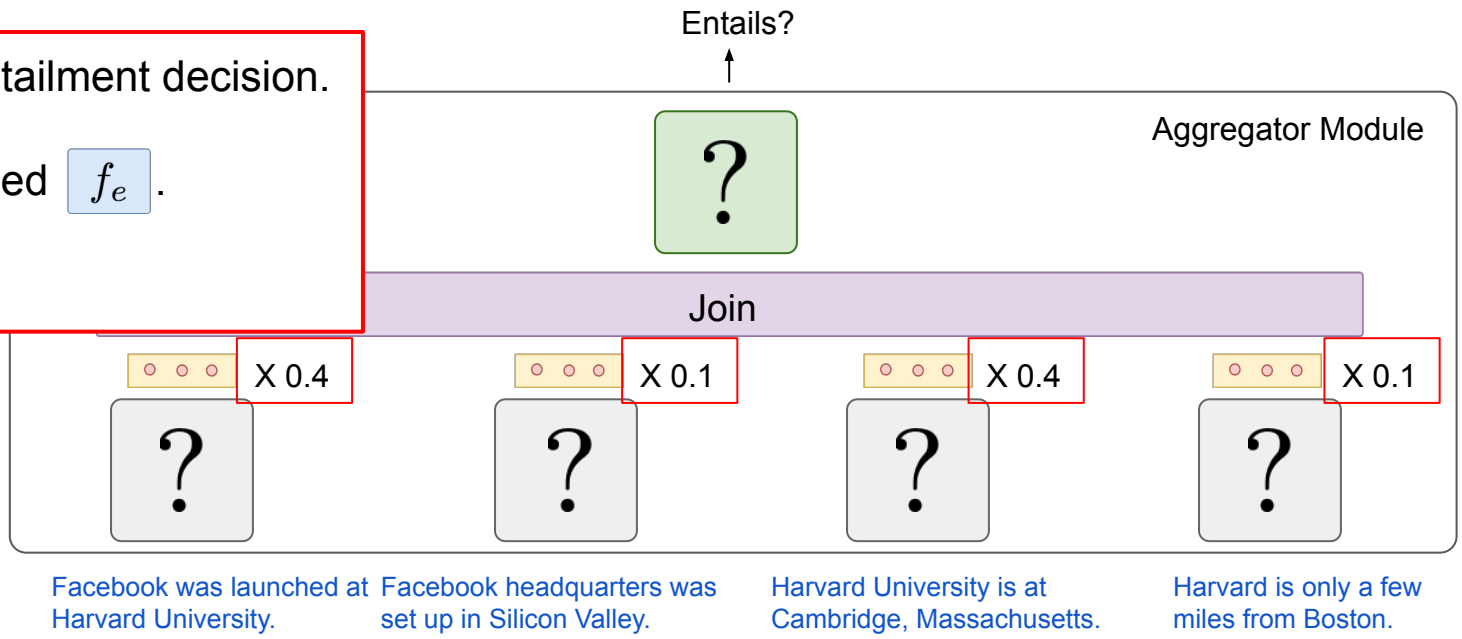


H: Facebook was launched in **Cambridge**.

What does Aggregator Module look like?

Aggregator ⇒ entailment decision.

We have pretrained f_e.



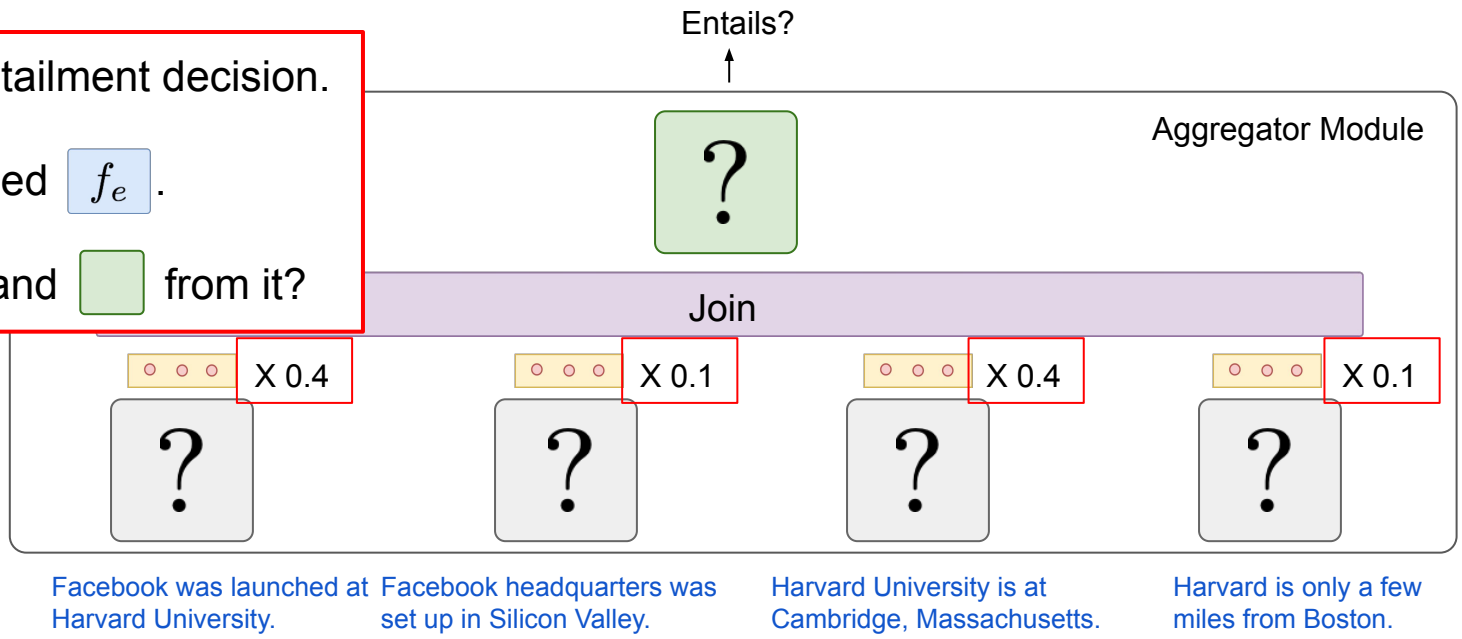
H: Facebook was launched in **Cambridge**.

What does Aggregator Module look like?

Aggregator ⇒ entailment decision.

We have pretrained f_e.

How to get [ ] and [ ] from it?

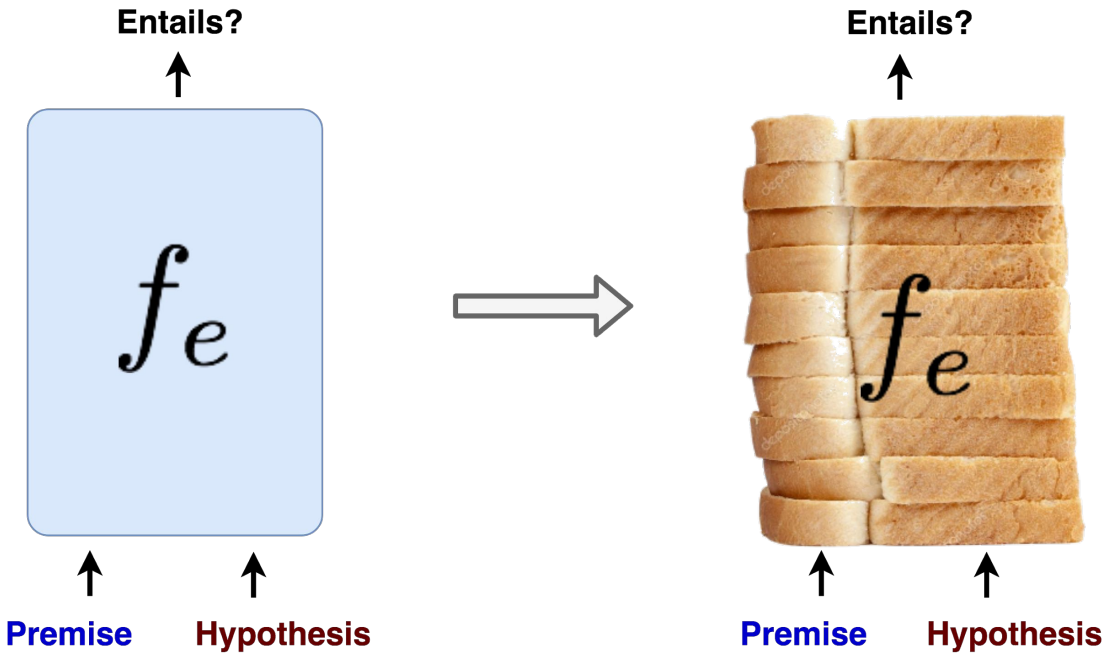


H: Facebook was launched in **Cambridge**.

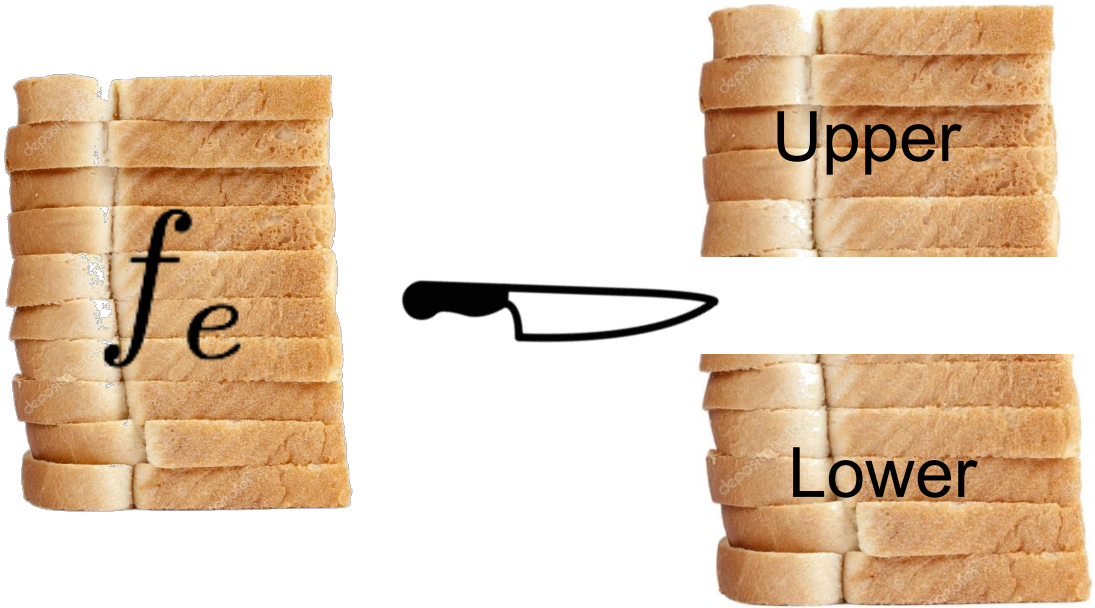
What is this talk about?

- 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

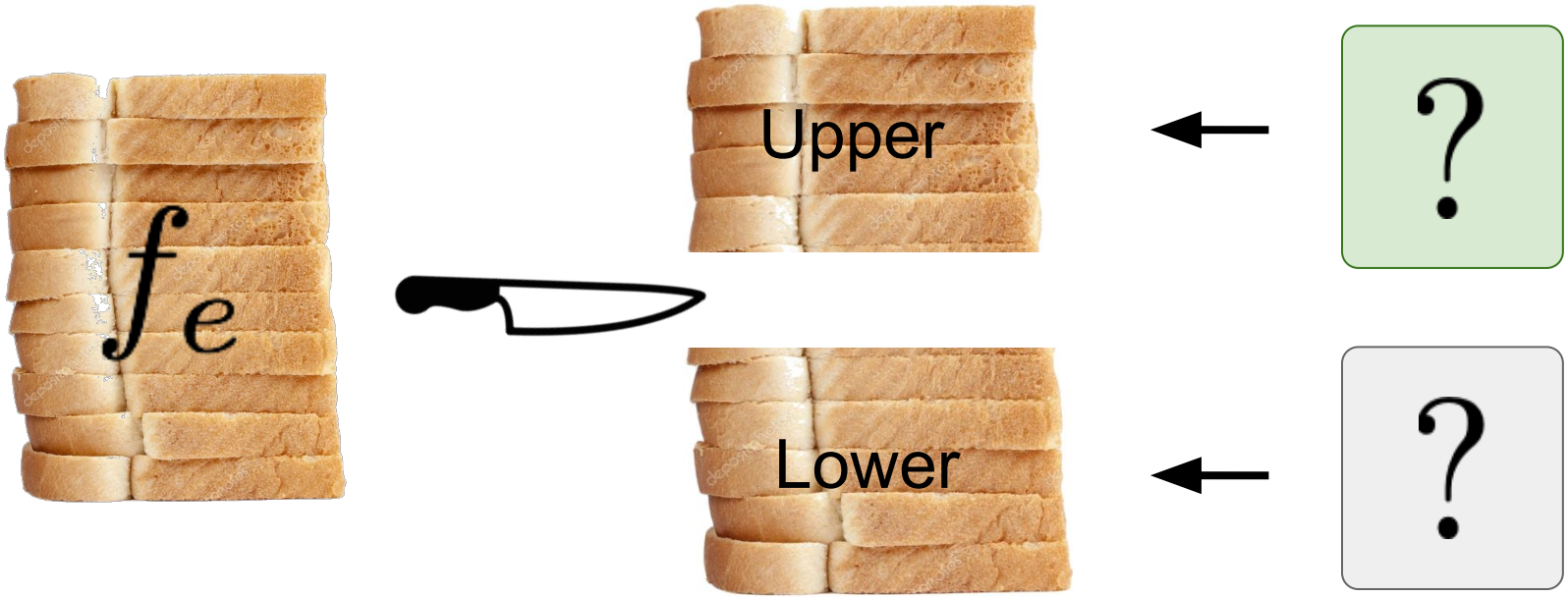
Neural Models are **Multi**-Layered



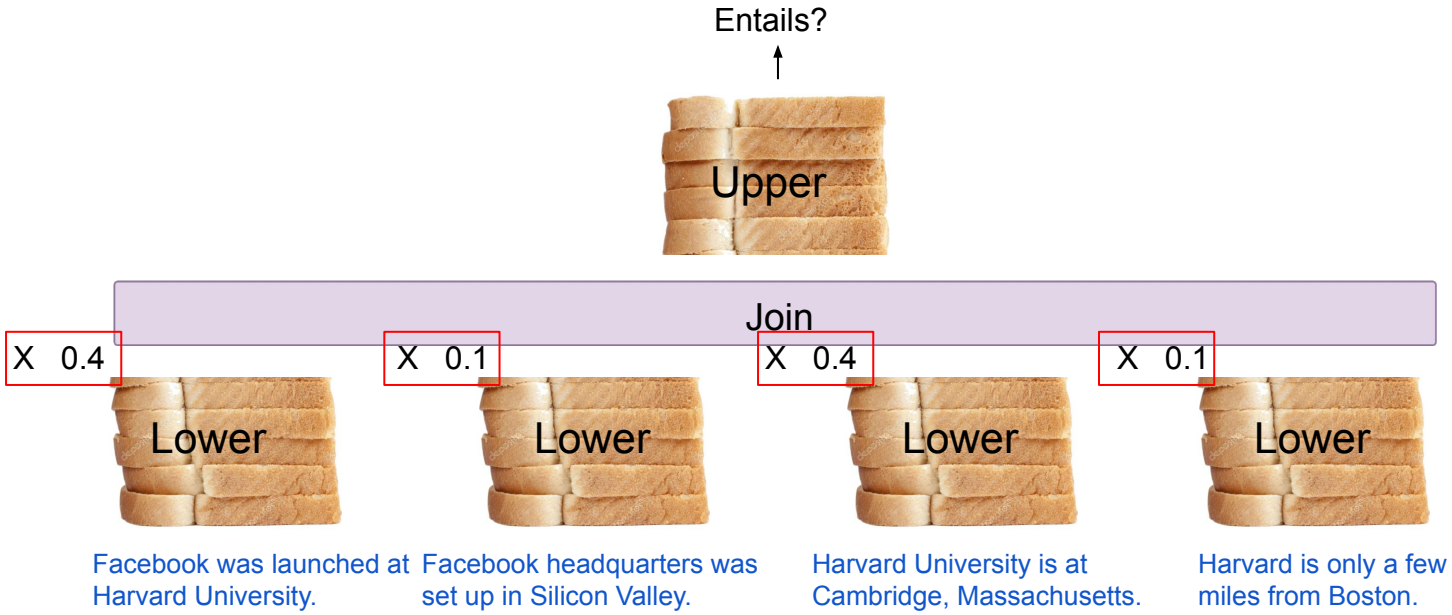
Entailment model (f_e) can be cut into parts



Entailment model (f_e) can be cut into parts

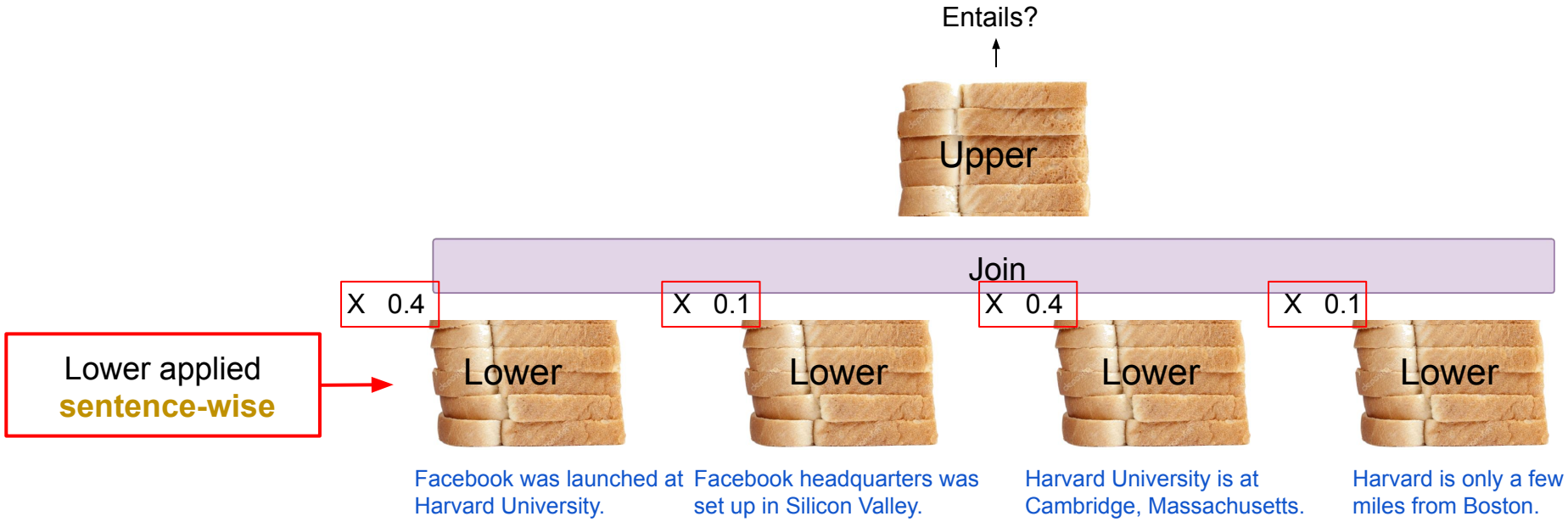


Aggregator Module: Incorporate weights (Join)



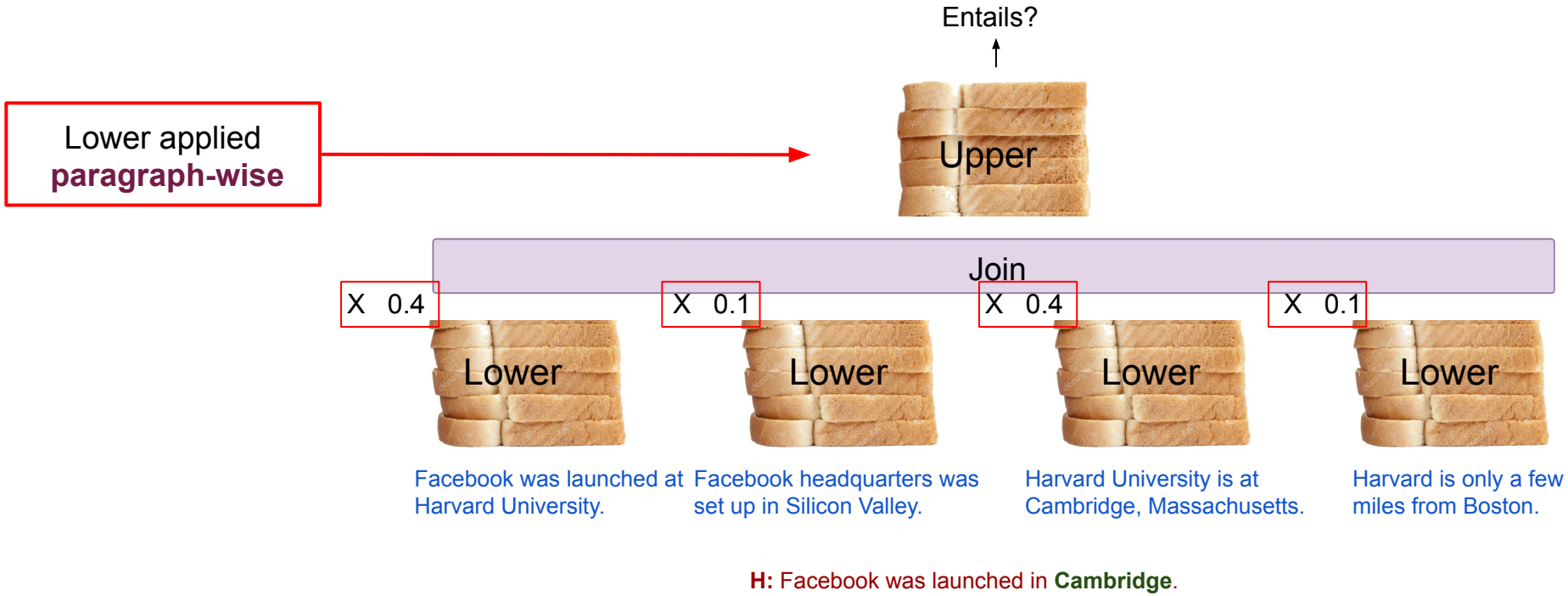
H: Facebook was launched in **Cambridge**.

Aggregator Module: Incorporate weights (Join)



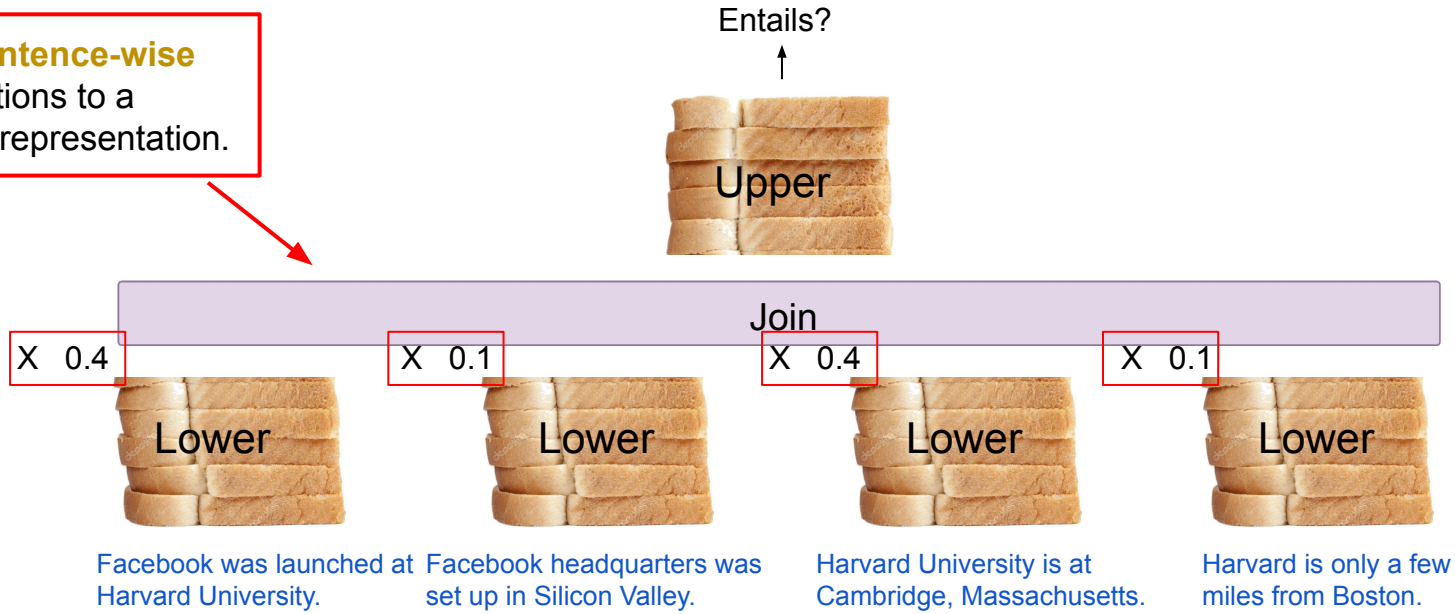
H: Facebook was launched in **Cambridge**.

Aggregator Module: Incorporate weights (Join)



Aggregator Module: Incorporate weights (Join)

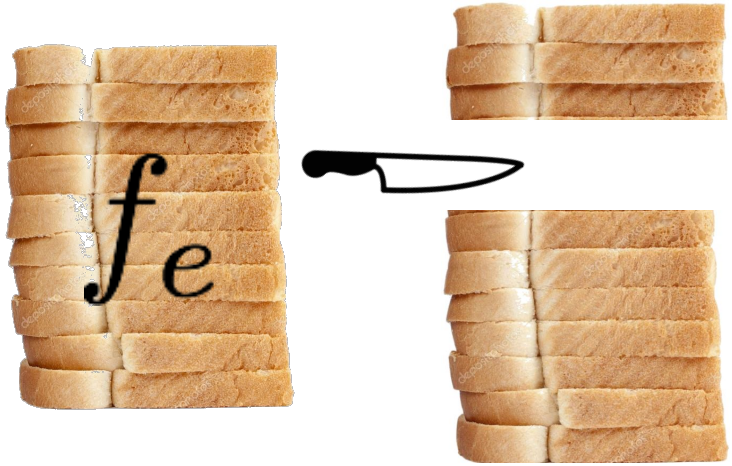
Join **scaled** **sentence-wise** representations to a **paragraph-wise** representation.

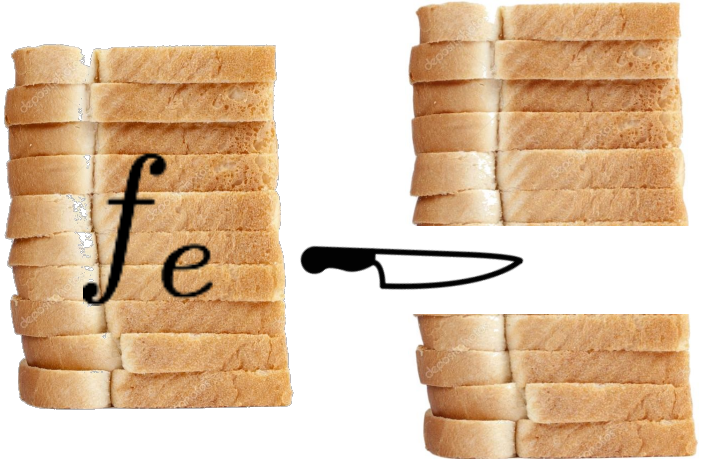


H: Facebook was launched in **Cambridge**.

Cut and incorporate weights at multiple layers

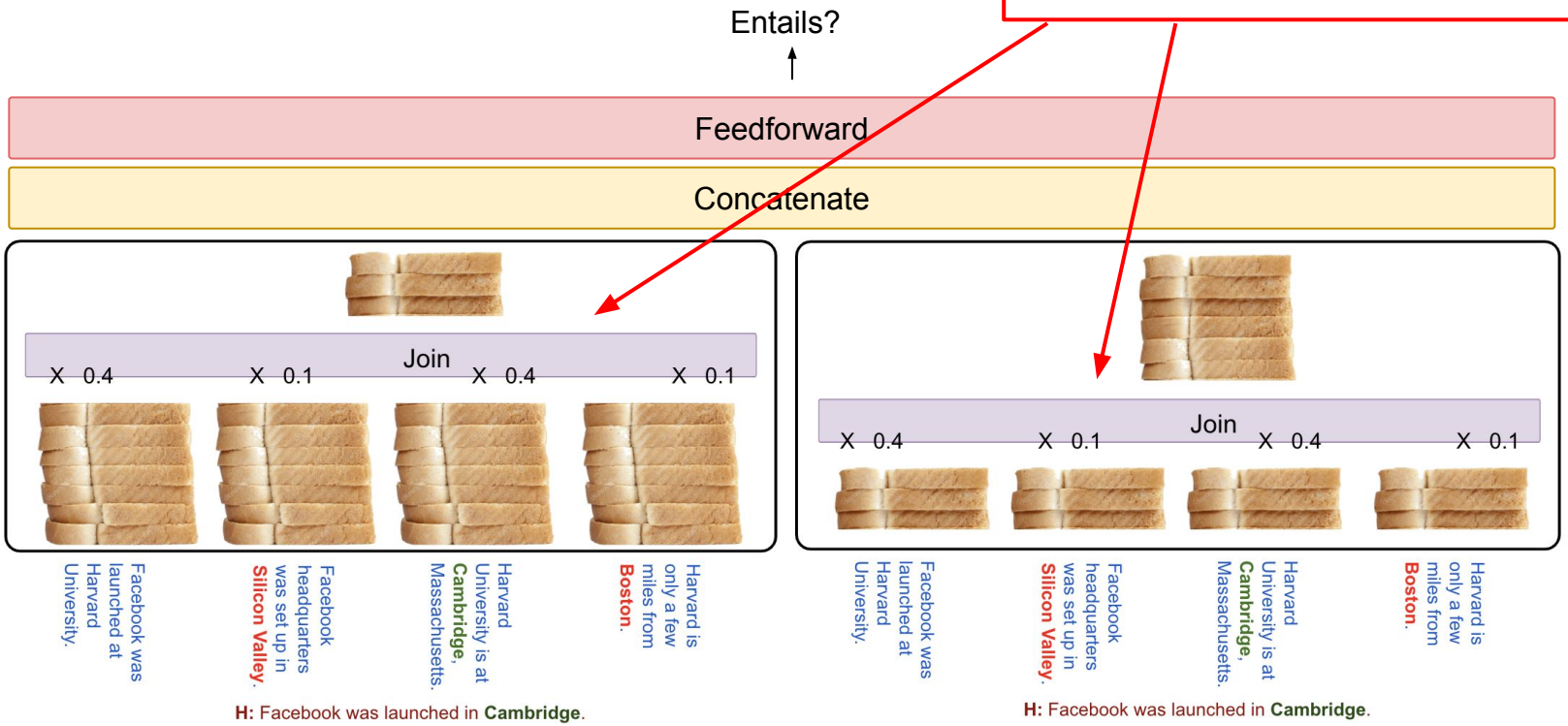
... because different layers in neural stack capture different types of information





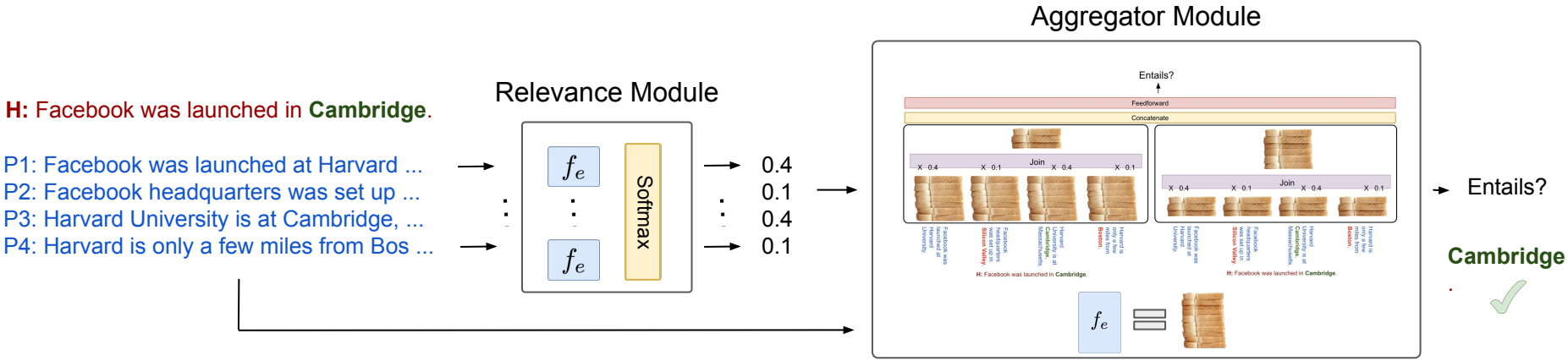
Full Aggregator Module

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



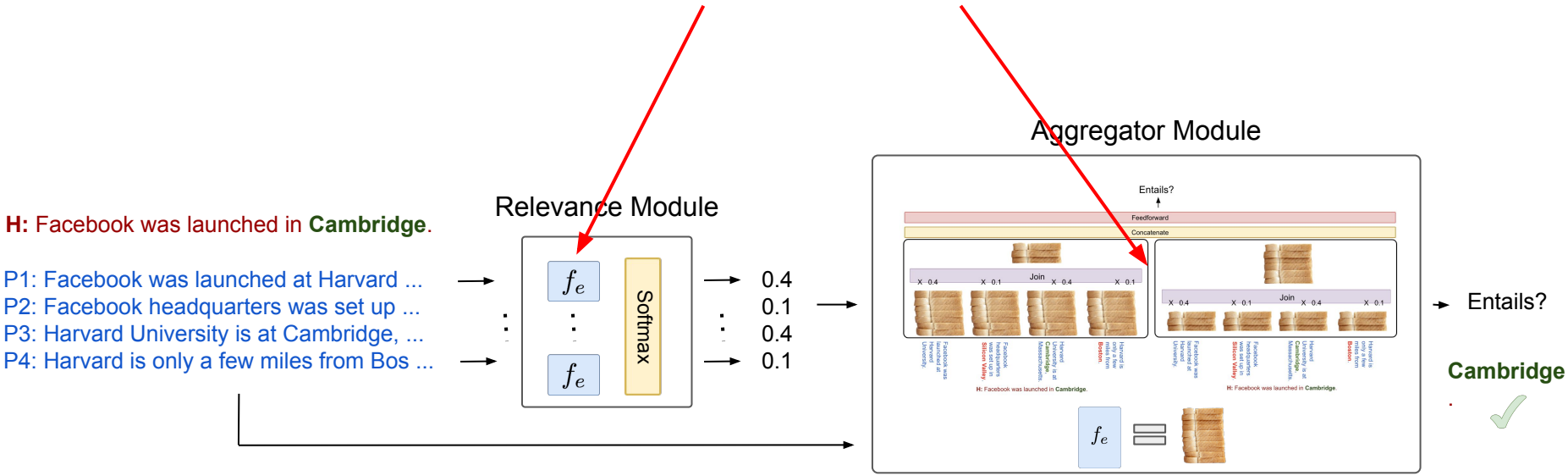
Multee ( ) Model

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



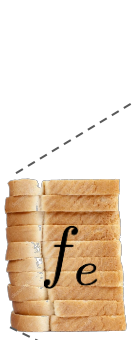
Multee: **MU**lti **L**ayer aggregation of **TE**xtual **E**ntailment Representations

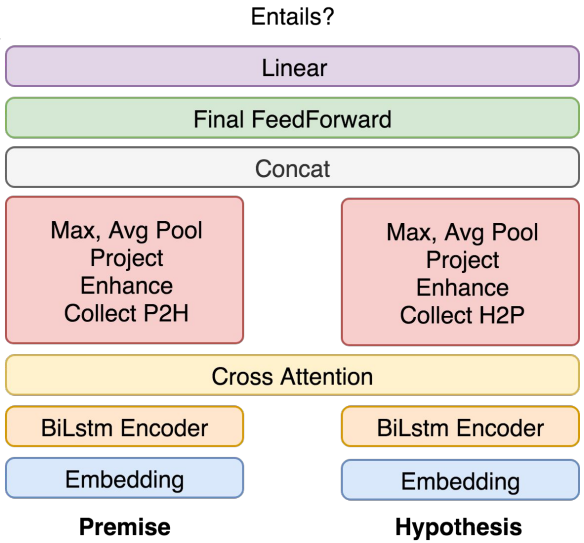
Multee ( ) Choices: f_e and **Join** layers?



Multee: **MU**lti **L**ayer aggregation of **TE**xtual **E**ntailment Representations

Multee ( ) Choices: f_e and **Join** layers?





Final Layer (FL)

Cross Attention Layer (CA)

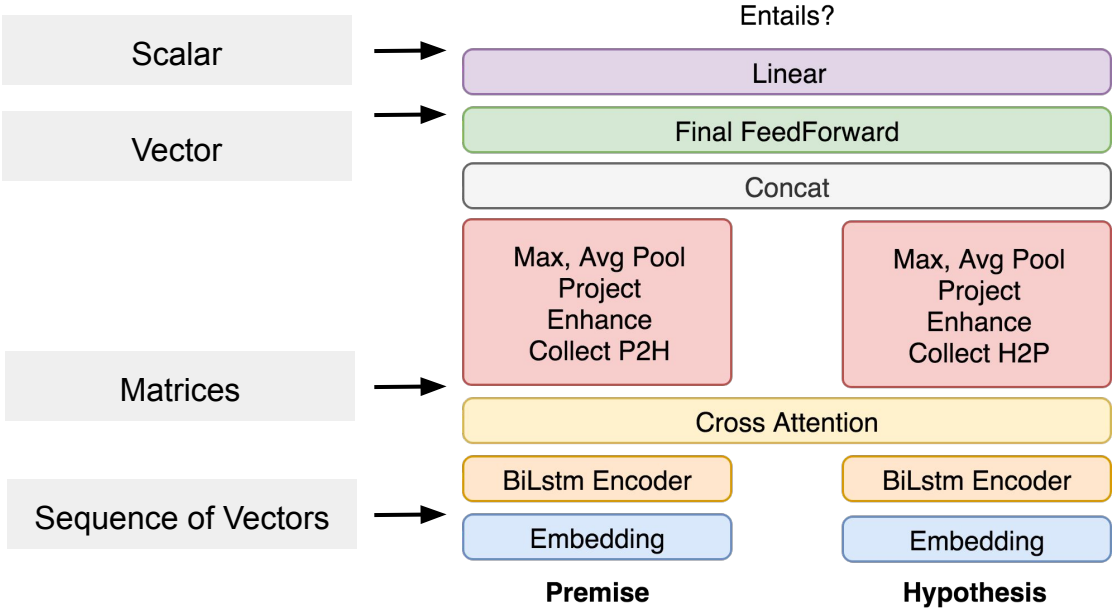
ESIM Stack (Chen et al, 16)

f_e

Join Layers

How to define Join operation?

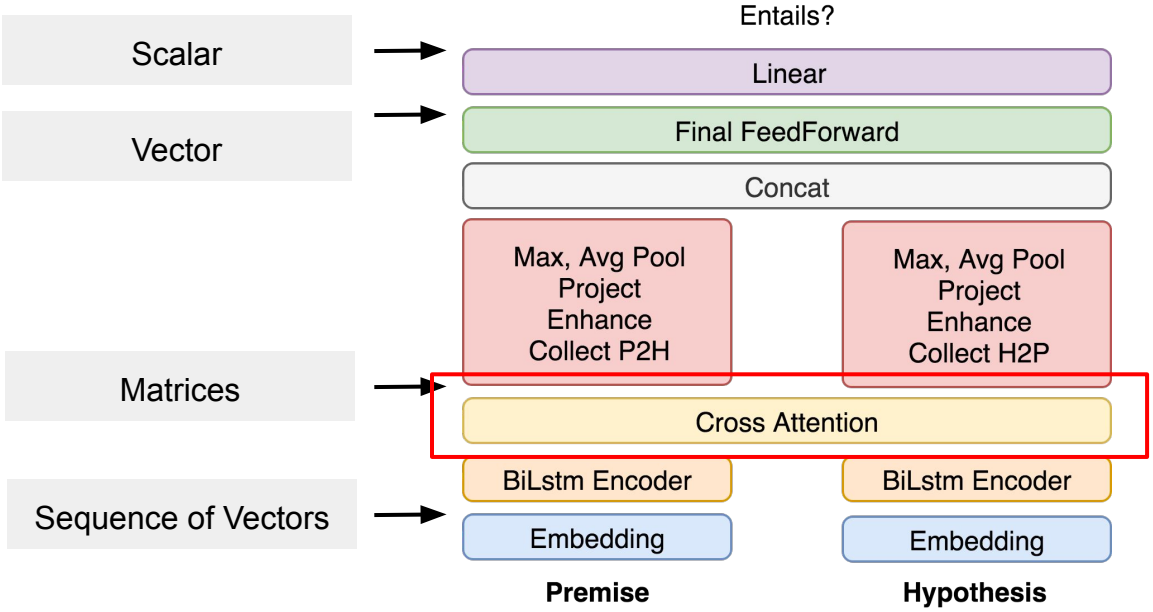
Different Layers have different output/**s** and so need different Join Operations



ESIM Stack (Chen et al, 16)

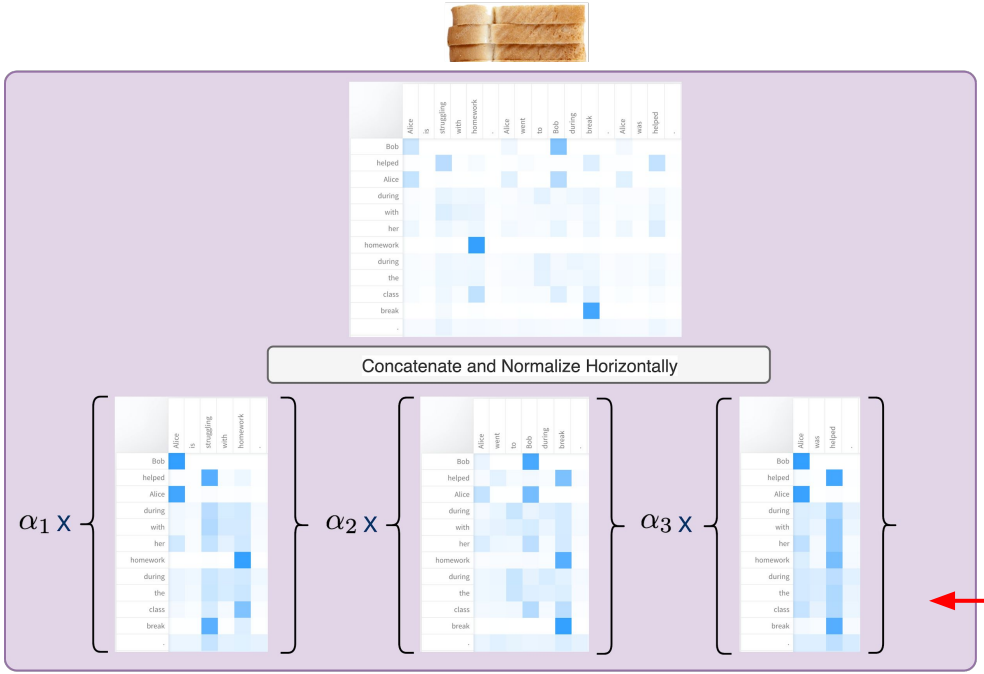
How to define Join operation?

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



ESIM Stack (Chen et al, 16)

Join Operations - **C**ross **A**ttention **L**ayer **(CA)**



Outputs sentence wise attention matrices

$$H \times P_i$$

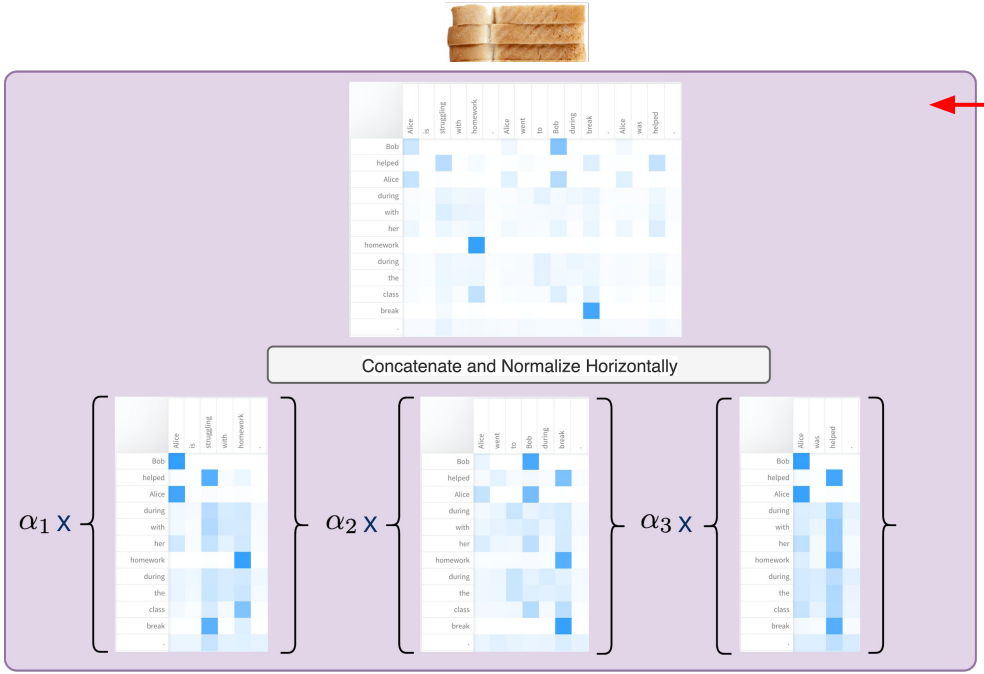
P1: "Alice is struggling with homework"

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 - **C**ross **A**ttention **L**ayer **(CA)**



Expects paragraph attention matrix

$$H \times (\sum_i P_i)$$

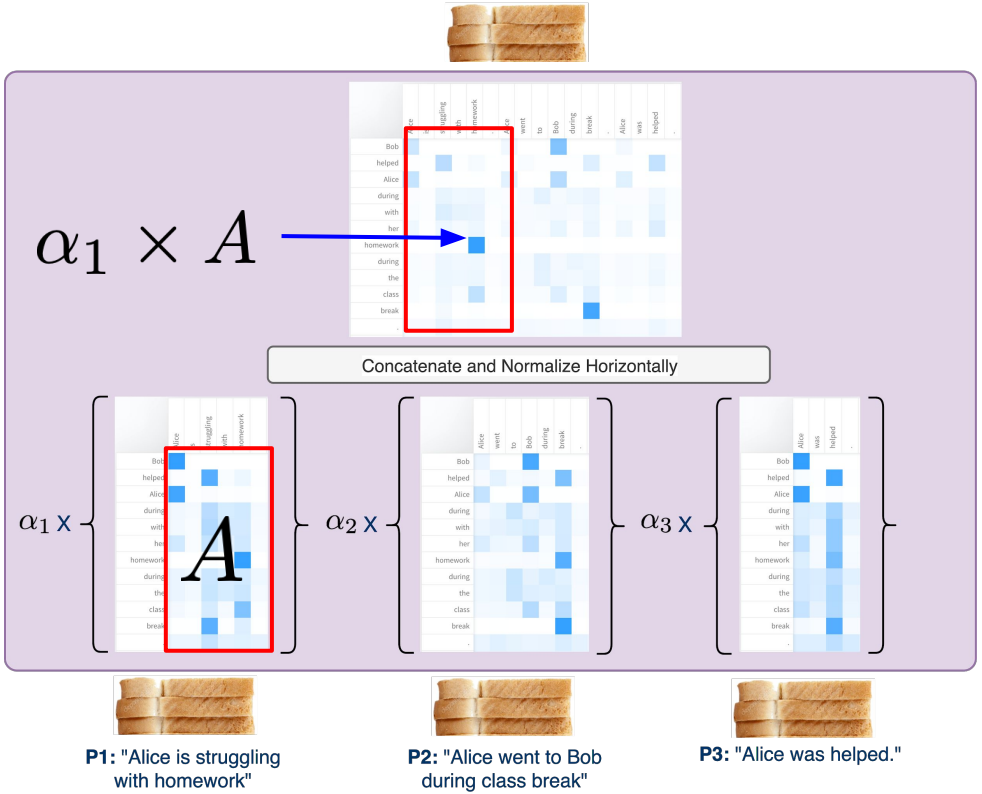
P1: "Alice is struggling with homework"

P2: "Alice went to Bob during class break"

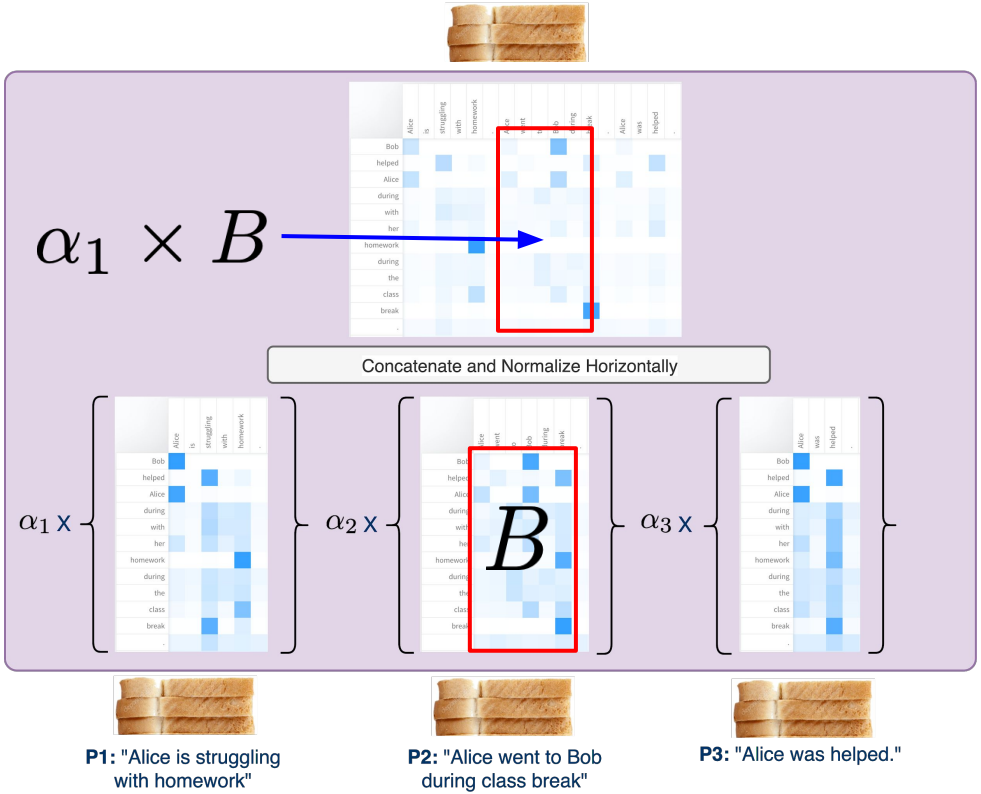
P3: "Alice was helped."

H: "Bob helped Alice during with her homework during the class break."

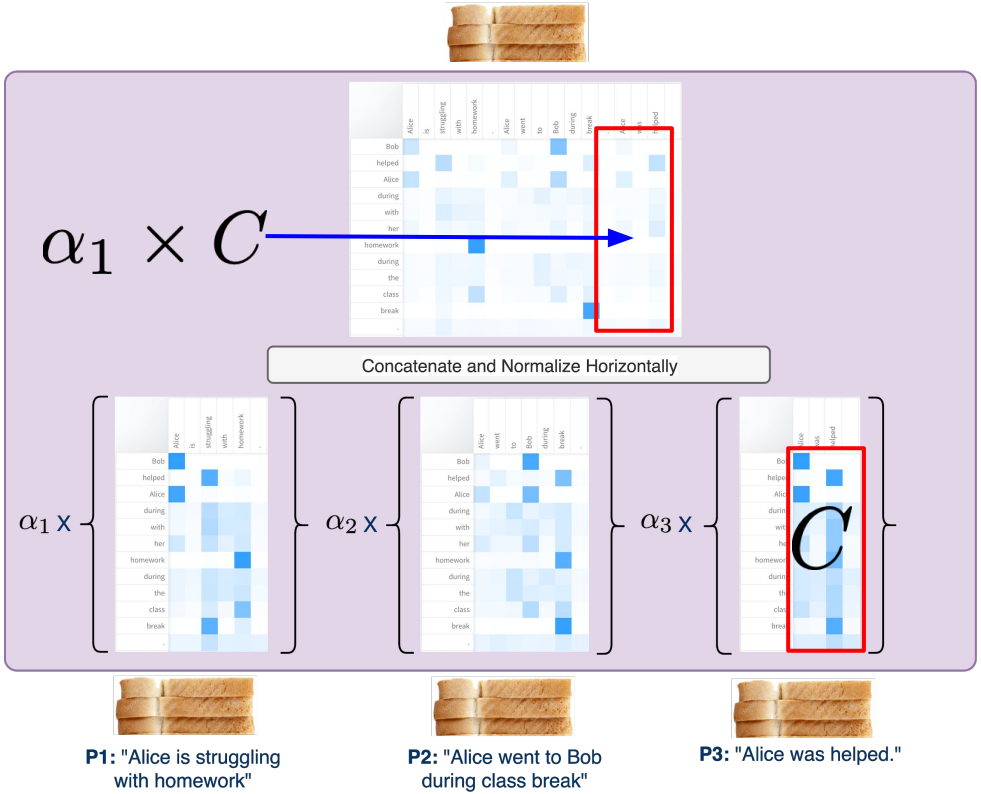
Join Operations - **C**ross **A**ttention **L**ayer **(CA)**



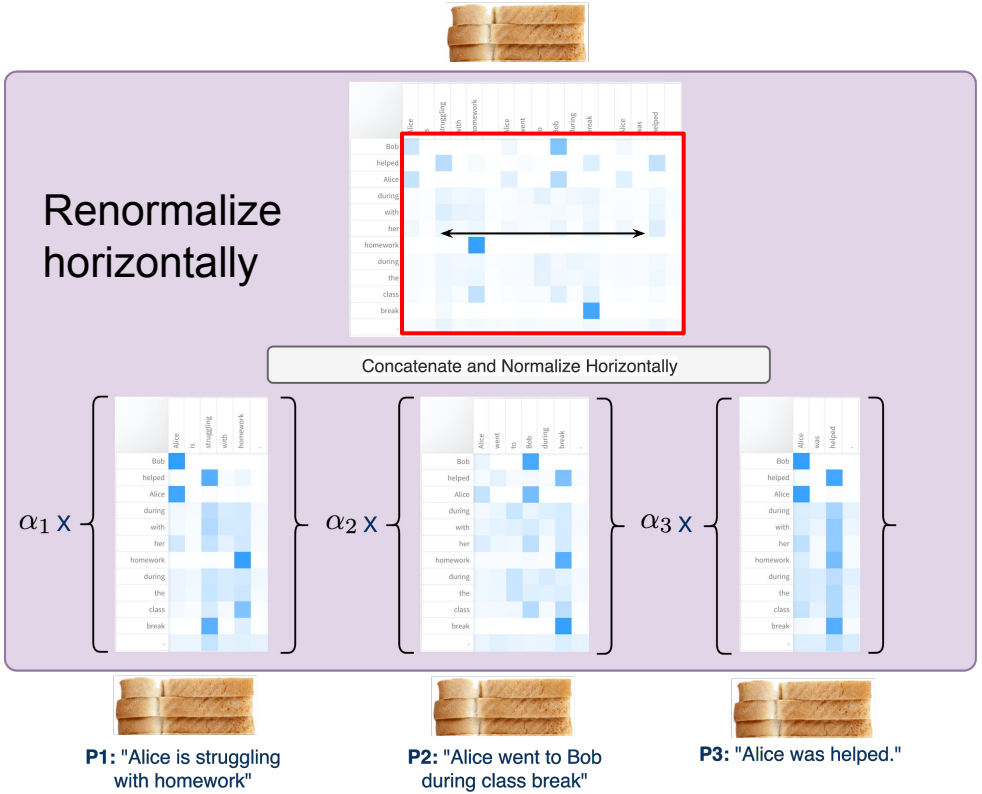
Join Operations - **C**ross **A**ttention **L**ayer **(CA)**



Join Operations - **C**ross **A**ttention **L**ayer **(CA)**



Join Operations - **C**ross **A**ttention **L**ayer **(CA)**



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 ( ) help over direct usages of Entailment models (Idea 1 & 2)?
2. How Multee ( ) compares with other QA models?

Evaluation

1. Does Multee ( ) help over direct usages of Entailment models (Idea 1 & 2)?
2. How Multee ( ) compares with other QA models?

To test this we build Multee:

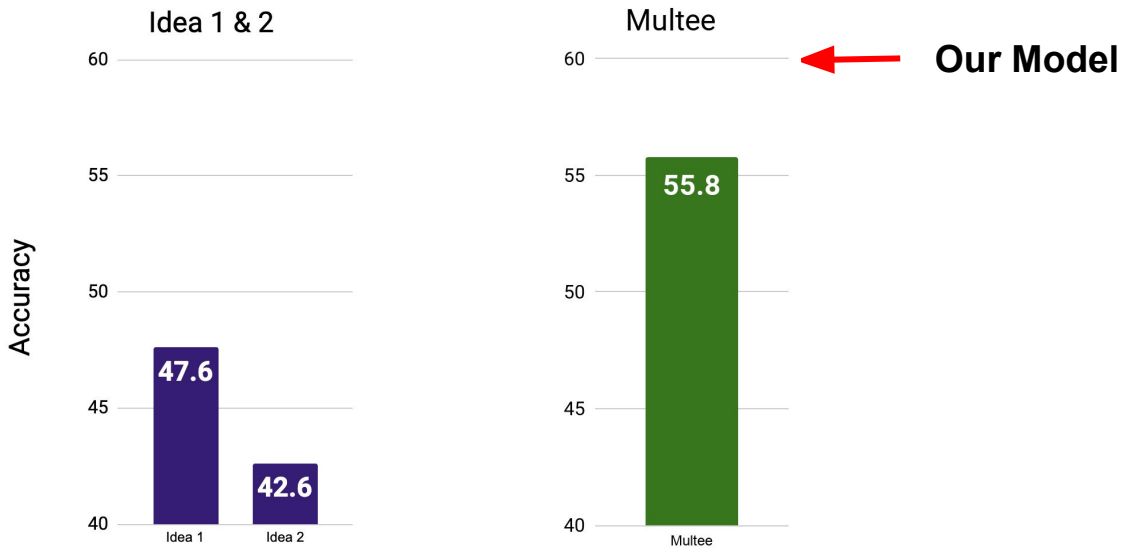
- With **ESIM** entailment model.
- Pretrained on **SNLI + MultiNLI** entailment datasets.

Evaluation Multihop-QA Datasets

- OpenBookQA
- MultiRC

Main Results - OpenBookQA

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



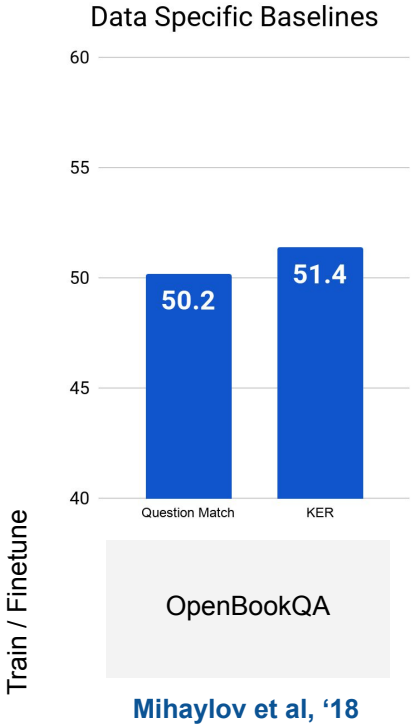
Train / Finetune

1. SNLI + MultiNLI
2. OpenBookQA

1. SNLI + MultiNLI
2. OpenBookQA

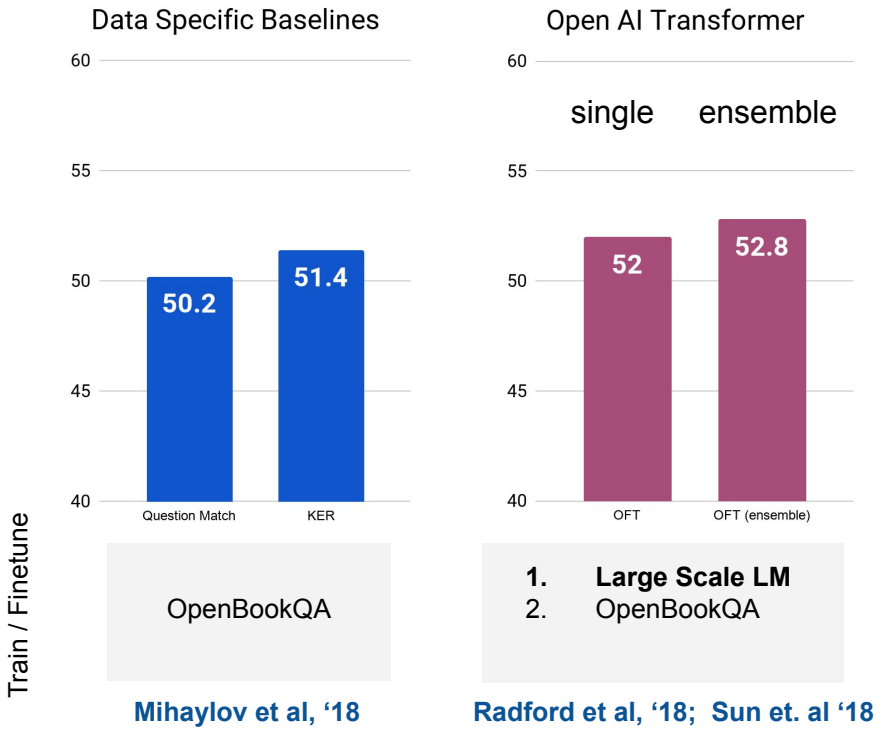
Main Results - OpenBookQA

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



Main Results - OpenBookQA

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

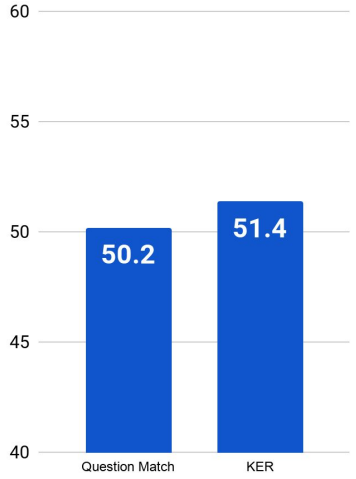


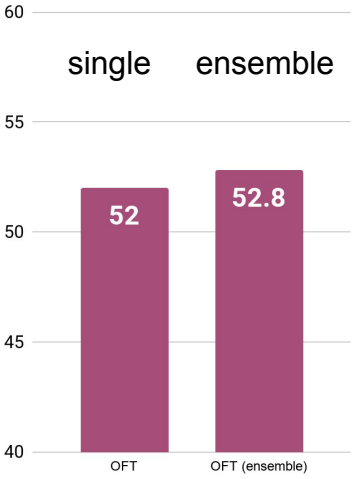
Main Results - OpenBookQA

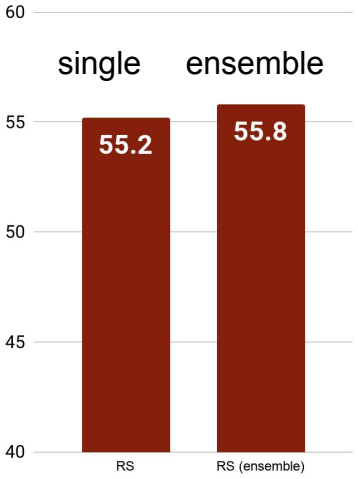
2. How Multee ( ) compares with other QA models?

NAACL'19

Data Specific Baselines | Open AI Transformer | Reading Strategies







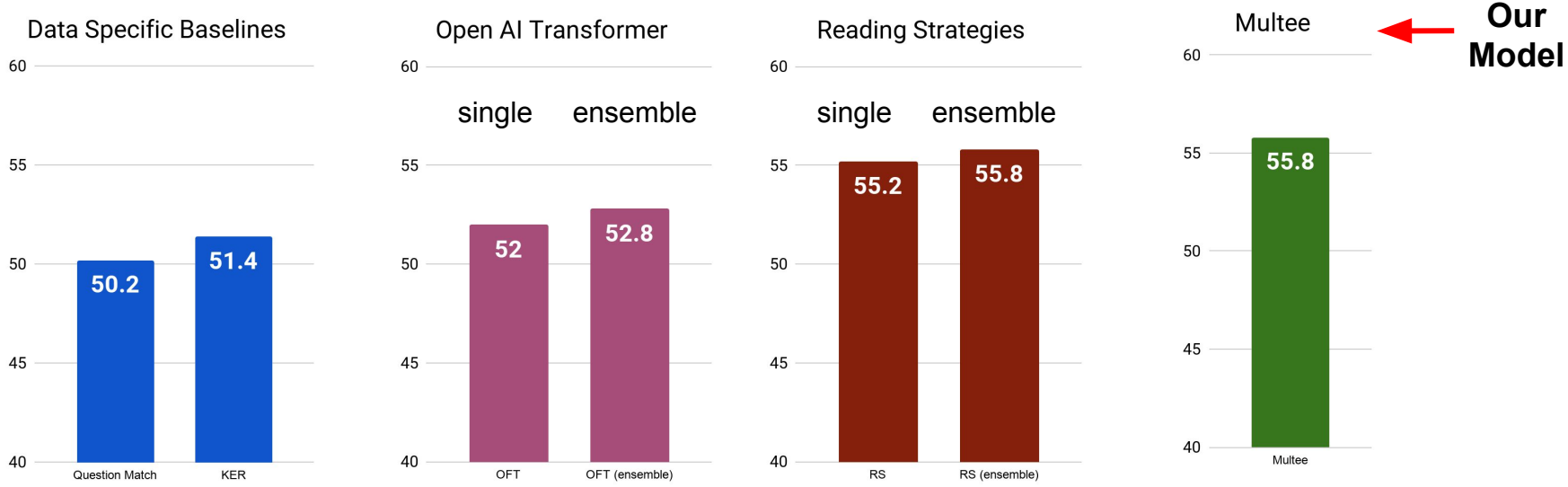
Train / Finetune

OpenBookQA	1. **Large Scale LM** 2. OpenBookQA	1. **Large Scale LM** 2. **Large Scale QA (RACE)** 3. OpenBookQA
Mihaylov et al, '18	**Radford et al, '18; Sun et. al '18**	**Sun et. al '19**

Main Results - OpenBookQA

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

NAACL'19



Train / Finetune

Data Specific Baselines	Open AI Transformer	Reading Strategies	Multee
OpenBookQA	1. **Large Scale LM** 2. OpenBookQA	1. **Large Scale LM** 2. **Large Scale QA (RACE)** 3. OpenBookQA	1. SNLI + MultiNLI 2. OpenBookQA
Mihaylov et al, '18	**Radford et al, '18; Sun et. al '18**	**Sun et. al '19**	

Main Results - MultiRC

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

NAACL'19

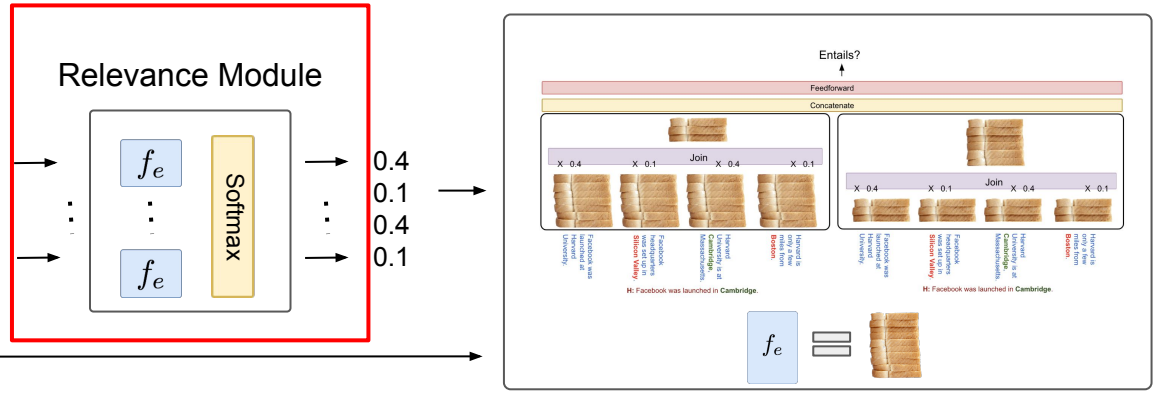
Train / Finetune

Data Specific Baselines	Open AI Transformer	Reading Strategies	Multee
MultiRC	1. **Large Scale LM** 2. MultiRC	1. **Large Scale LM** 2. **Large Scale QA (RACE)** 3. MultiRC	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**.

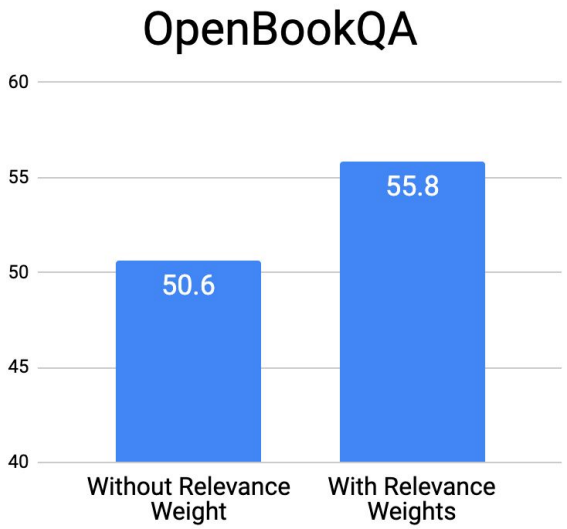
P1: Facebook was launched at Harvard ...
P2: Facebook headquarters was set up ...
P3: Harvard University is at Cambridge, ...
P4: Harvard is only a few miles from Bos ...



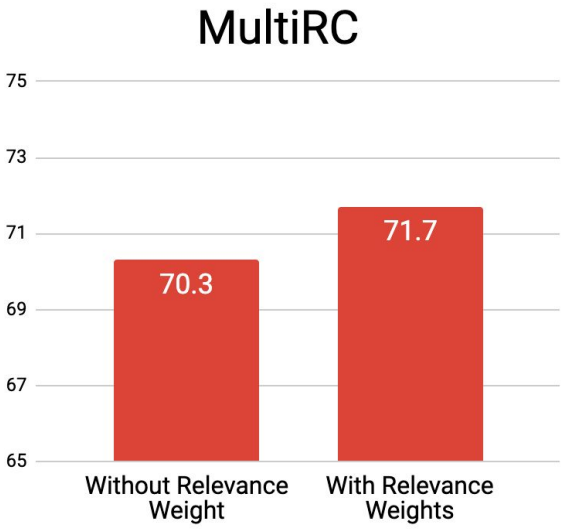
Entails?

Cambridge

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

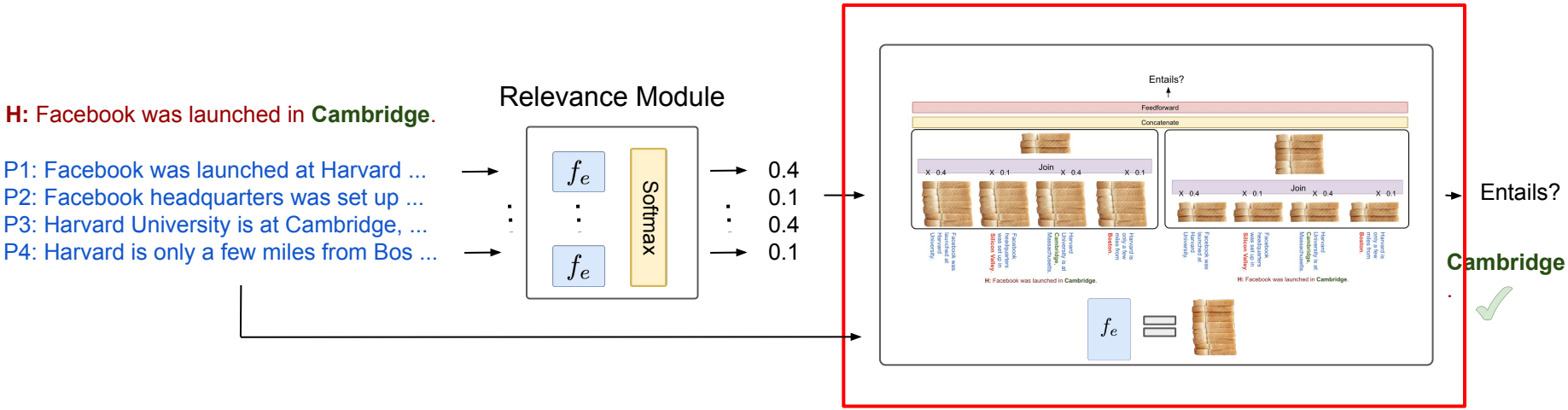


Full Model

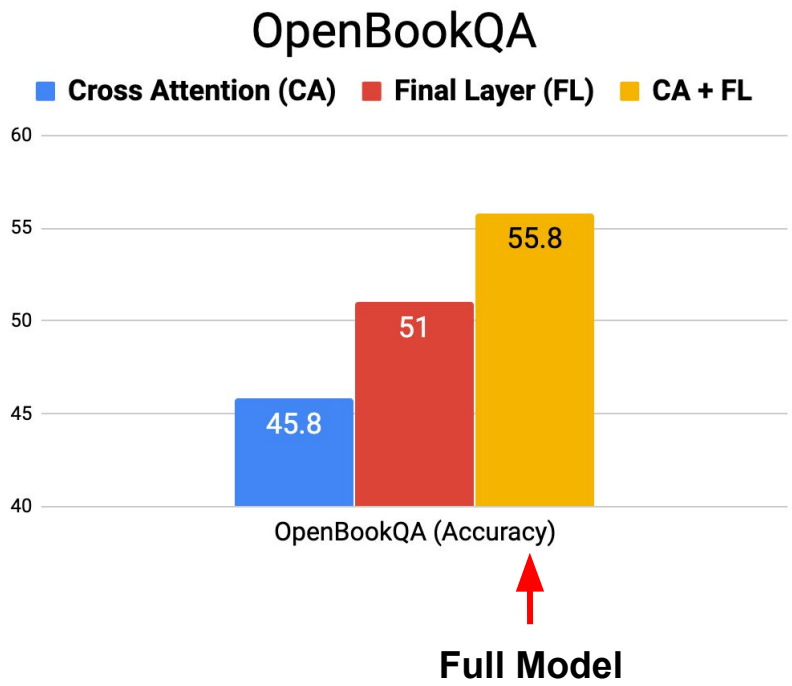


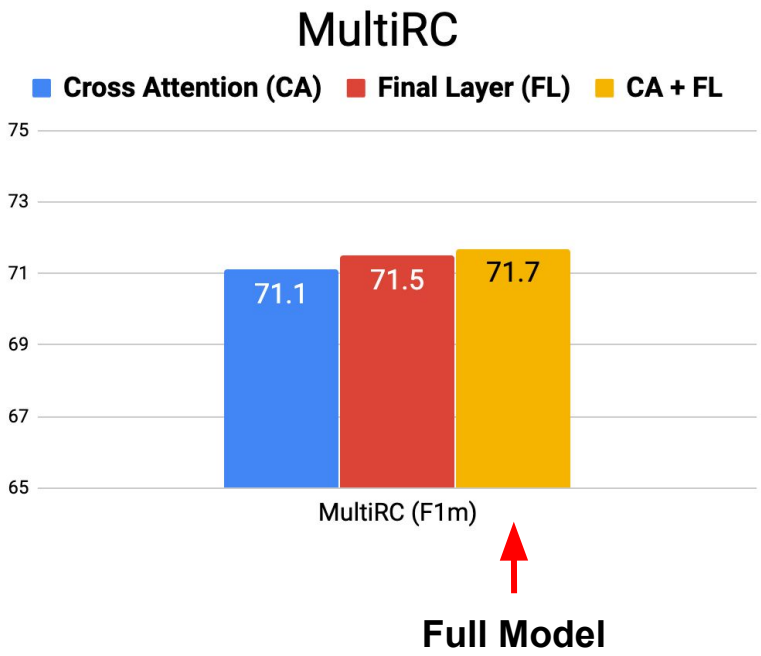
Full Model

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



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

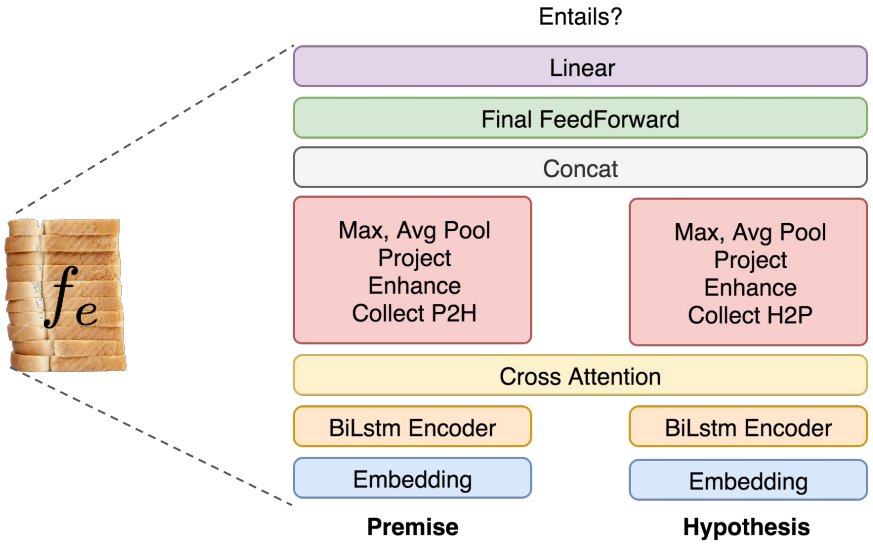


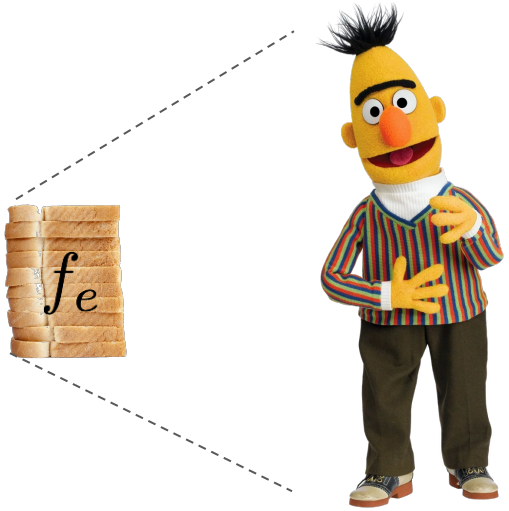


Is Multee ( ) still relevant in context of BERT ( )?

Is Multee ( ) still relevant in context of BERT ( )?

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



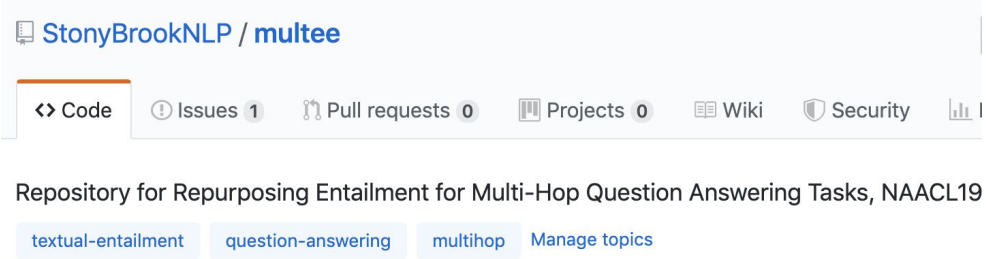


Conclusion

- 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:

More Realistic

Framing QA with Entailment: ~~Simple~~ Example

Question

Q: 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

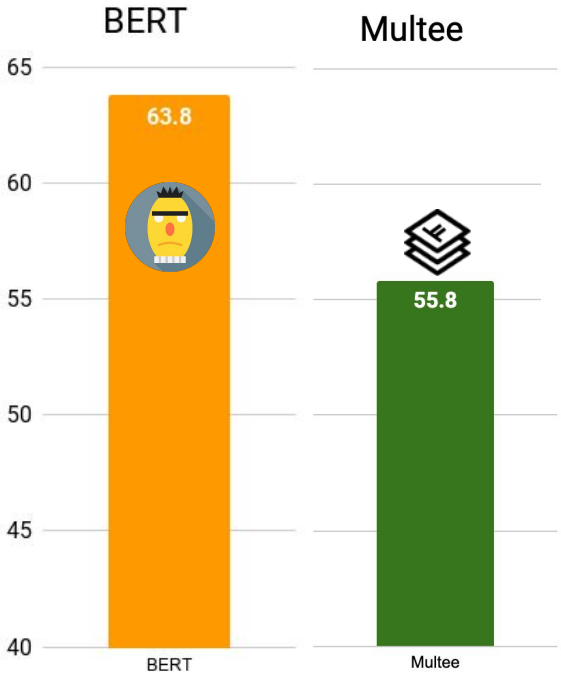
H: Facebook was launched in **Cambridge.**

Entailment Question:

Does the information in **{P1**, **P2**, **P3**, **P4}** entail **H** ?

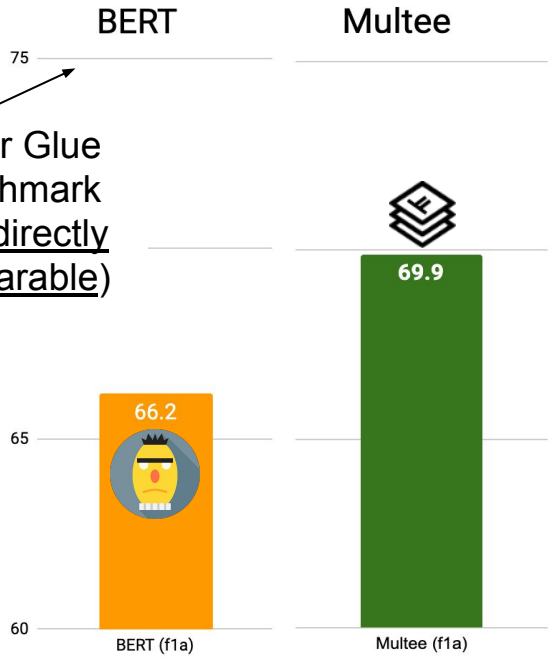
Is Multee ( ) still relevant in context of BERT ( )?

Official Leaderboard



OpenBookQA

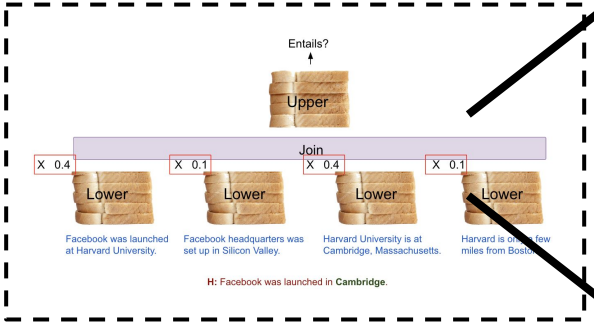
Super Glue Benchmark (not directly comparable)



MultiRC

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

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.

X 0.4
X 0.1
X 0.4
X 0.1

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**.

Entails ?

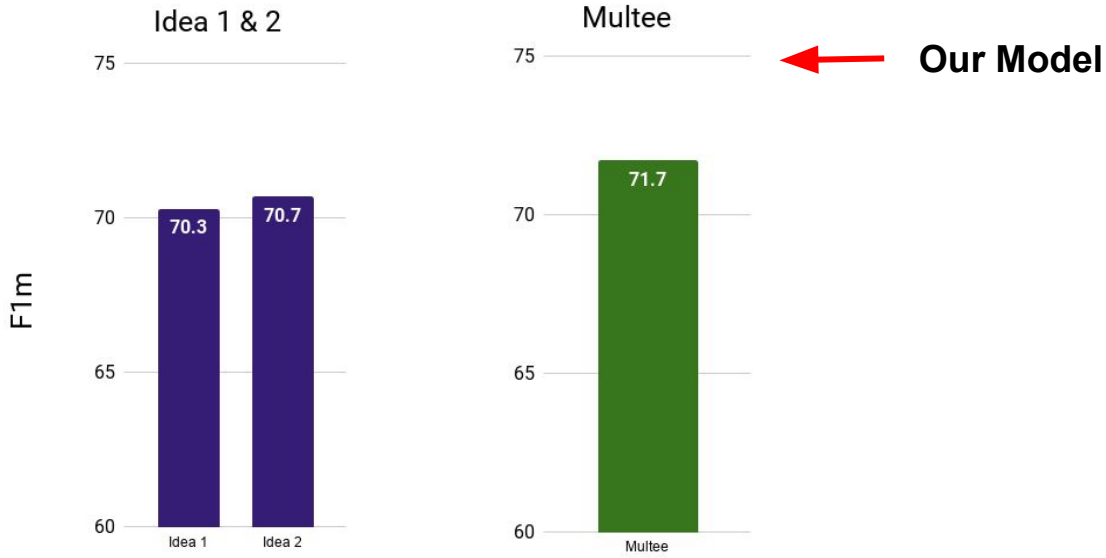
Original Information **X** Relevance Weights → Scaled-Down Information

Scaling down reduces distracting information

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

Main Results - MultiRC

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



Train / Finetune

1. SNLI + MultiNLI
2. MultiRC

1. SNLI + MultiNLI
2. MultiRC