Using Natural Language Relations between Answer Choices for Machine Comprehension

Rajkumar Pujari and Dan Goldwasser

June 5, 2019

PURDUE UNIVERSITY
Department of Computer Science
COLLEGE OF SCIENCE

Minneapolis, MN
NAACL HLT 2019
June 2-7, 2019

Intuition

Intuition

When humans perform Reading Comprehension, we answer all the given questions consistently.

But, when we test Machine Comprehension, most computational settings consider each question or each choice in isolation.

Intuition

Intuition

When humans perform Reading Comprehension, we answer all the given questions consistently.

But, when we test Machine Comprehension, most computational settings consider each question or each choice in isolation.

Example

1. *When were the eggs added to the pan to make the omelette?*
    - When they turned on the stove
    - When the pan was the right temperature ✪
2. *Why did they use stove to cook omelette?*
    - They didn't use the stove but a microwave
    - Because they needed to heat up the pan ✪

Source: SemEval 2018 Task-11 dataset ([Ostermann et al. 2018])

Intuition (contd.)

- Similarly, in settings where multiple choices could be correct, we could use the relationships between choices.

Intuition (contd.)

- Similarly, in settings where multiple choices could be correct, we could use the relationships between choices.

Example

- *How can the military benefit from the existence of the CIA?*
  1. They can use them as they wish
  2. The agency is keenly attentive to the military's strategic and tactical requirements ✪
  3. The CIA knows what intelligence the military requires and has the resources to obtain that intelligence ✪
- c_3 entails $c_2 \implies$ flip c_2 from wrong to correct.

Source: MultiRC dataset ([Khashabi et al. 2018])

Abstract

1. We propose a method to leverage entailment and contradiction relations between the answer choices to improve machine comprehension.

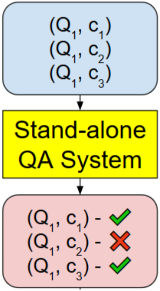
Abstract

1. We propose a method to leverage entailment and contradiction relations between the answer choices to improve machine comprehension.
2. We first perform Question Answering (QA) and "weakly-supervised" Natural Language Inference (NLI) relation detection separately. Then, we use the NLI relations to re-evaluate the answers.

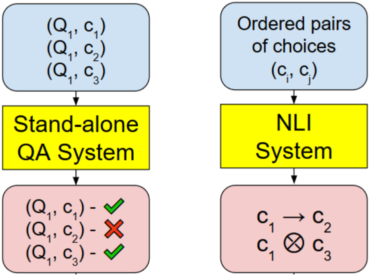
Abstract

1. We propose a method to leverage entailment and contradiction relations between the answer choices to improve machine comprehension.
2. We first perform Question Answering (QA) and "weakly-supervised" Natural Language Inference (NLI) relation detection separately. Then, we use the NLI relations to re-evaluate the answers.
3. We also propose a multitask learning model that learns both the tasks jointly.

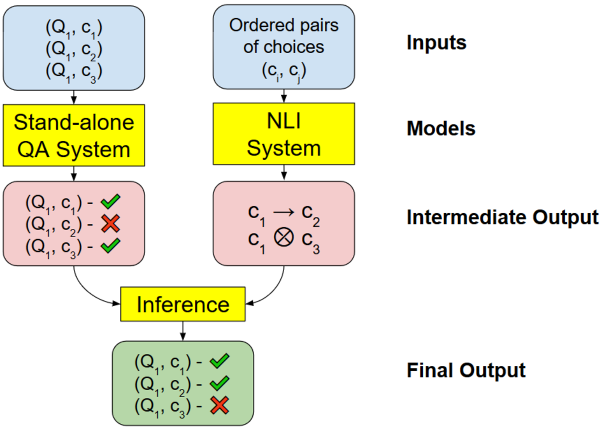
Approach



Approach


(Q_1, c_1)
(Q_1, c_2)
(Q_1, c_3)
Stand-alone QA System
(Q_1, c_1) -
(Q_1, c_2) -
(Q_1, c_3) -
Ordered pairs of choices
(c_i, c_j)
NLI System
$c_1 \rightarrow c_2$
$c_1 \otimes c_3$

Approach


(Q1, c1)
(Q1, c2)
(Q1, c3)
Ordered pairs of choices (ci, cj)
Inputs
Stand-alone QA System
NLI System
Models
(Q1, c1) -
(Q1, c2) -
(Q1, c3) -
c1 → c2
c1 ⊗ c3
Intermediate Output
Inference
(Q1, c1) -
(Q1, c2) -
(Q1, c3) -
Final Output

Stand-alone QA System

- We use the TriAN-single model proposed by [Wang et al. 2018] for SemEval-2018 task-11 as our stand-alone QA system.

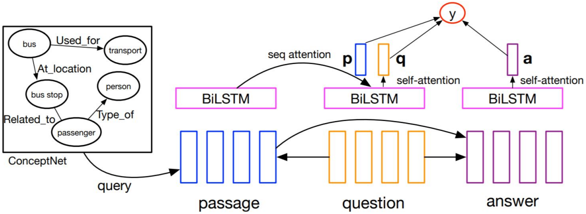


Figure: TriAN model architecture (figure adopted from [Wang et al. 2018])

NLI System

- Our NLI system was inspired from decomposable-attention model proposed by [Parikh et al. 2016]

NLI System

- Our NLI system was inspired from decomposable-attention model proposed by [Parikh et al. 2016]
- Issue: Choices are often short phrases. NLI relations among them exist only in the context of the given question.

Example

What do human children learn by playing games and sports?

1. Learn about the world ✪
2. Learn to cheat

NLI System

- Our NLI system was inspired from decomposable-attention model proposed by [Parikh et al. 2016]
- Issue: Choices are often short phrases. NLI relations among them exist only in the context of the given question.

Example

What do human children learn by playing games and sports?

1. Learn about the world ✪
2. Learn to cheat

- Resolution: We modified the architecture proposed in [Parikh et al. 2016] to accommodate the question-choice pairs as opposed to sentence pairs in the original model.

Inference

- We enforce consistency between the QA answers and the NLI relations at inference time.

Inference

- We enforce consistency between the QA answers and the NLI relations at inference time.
- The answers and the relations are scored by the confidence scores from the QA and the NLI systems.

Inference

- We enforce consistency between the QA answers and the NLI relations at inference time.
- The answers and the relations are scored by the confidence scores from the QA and the NLI systems.
- We used the following rules to enforce consistency:
  1. c_i is true & c_i entails $c_j \implies c_j$ is true.
  2. c_i is true & c_i contradicts $c_j \implies c_j$ is false.

Inference

- We enforce consistency between the QA answers and the NLI relations at inference time.
- The answers and the relations are scored by the confidence scores from the QA and the NLI systems.
- We used the following rules to enforce consistency:
  1. c_i is true & c_i entails $c_j \implies c_j$ is true.
  2. c_i is true & c_i contradicts $c_j \implies c_j$ is false.
- We used Deep Relational Learning (DRaiL) framework proposed by [Zhang et al. 2016] for inference

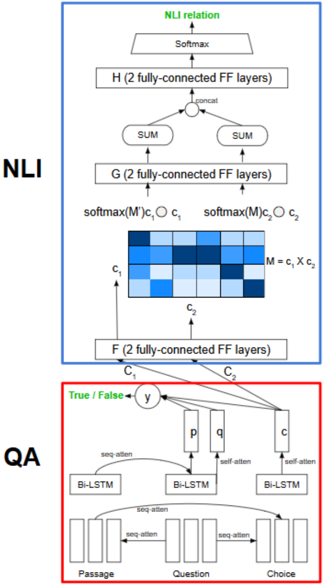
Self-Training

- We devised a self-training protocol to adopt the NLI system to the Machine Comprehension datasets (weak-supervision)

Self-Training

- We devised a self-training protocol to adopt the NLI system to the Machine Comprehension datasets (weak-supervision)
- If the "SNLI-trained" NLI model predicted entailment with a confidence above a threshold and the gold labels of the ordered choice pair were true-true, the relation was labeled entailment, and similarly we generate data for `contradiction`

Joint Model



The design of our joint model is motivated by the two objectives:

1. To leverage the benefit of multitask learning
2. To obtain a better representation for the question-choice pair for NLI detection

MultiRC Results

Method	EM_0	EM_1
Stand-alone QA	18.15	52.99
QA + NLI_{SNLI}	19.41	56.13
QA + $NLI_{MultiRC}$	21.62	55.72
Joint Model	20.36	57.08
Human	56.56	83.84

Table: Summary of results on MultiRC dataset. EM_0 is the percentage of questions for which all the choices are correct. EM_1 is the the percentage of questions for which at most one choice is wrong.

SemEval 2018 Results

Model	Dev	Test
Stand-alone QA	83.20%	80.80%
Joint Model	85.40%	82.10%

Table: Accuracy of various models on SemEval'18 task-11 dataset

Error Analysis

- Identification of NLI relations is far from perfect.
- NLI system returns `entailment` when there is a high lexical overlap
- NLI system returns `contradiction` upon the presence of a strong negation word such as *not*.

Summary

- We proposed a framework to use entailment and contradiction relations to improve Machine Comprehension

Summary

- We proposed a framework to use entailment and contradiction relations to improve Machine Comprehension
- Self-training results suggest the presence of other subtle relationships among choices.

Summary

- We proposed a framework to use entailment and contradiction relations to improve Machine Comprehension
- Self-training results suggest the presence of other subtle relationships among choices.
- Consider:
  1. I went shopping this extended weekend
  2. I ate a lot of junk food recently

Summary

- We proposed a framework to use entailment and contradiction relations to improve Machine Comprehension
- Self-training results suggest the presence of other subtle relationships among choices.
- Consider:

  1. I went shopping this extended weekend
  2. I ate a lot of junk food recently

  Text: *I snack when I shop*

Thank you!

Questions?

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth.
2018.
Looking beyond the surface: A challenge set for reading comprehension over multiple sentences.
In *NAACL*.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal.
2018.
Semeval-2018 task 11: Machine comprehension using commonsense knowledge.
In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit.
2016.
A decomposable attention model for natural language inference.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November. Association for Computational Linguistics.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Liu Jingming.
2018.
Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension.
CoRR, abs/1803.00191.

Xiao Zhang, Maria Leonor Pacheco, Chang Li, and Dan Goldwasser.
2016.
Introducing DRAIL - a step towards declarative deep relational learning.
In *Proceedings of the Workshop on Structured Prediction for NLP@EMNLP 2016, Austin, TX, USA, November 5, 2016*, pages 54–62.