Knowledgeable and Adversarially-Robust Representation Learning

Mohit Bansal

(RepL4NLP, ACL 2019)

Part 1:
Adversarially-Robust Model Representations

Motivation and Topics

- Are deep learning models and their representations robust to diverse adversaries (in tasks such as QA, multi-hop reasoning, dialogue generation, and NLI)?
- How far can adversarial training go in bringing back robustness?
- What types of direct model enhancements and better evaluations are needed for robust representation learning?
- How do we ensure robustness to all types of adversaries?

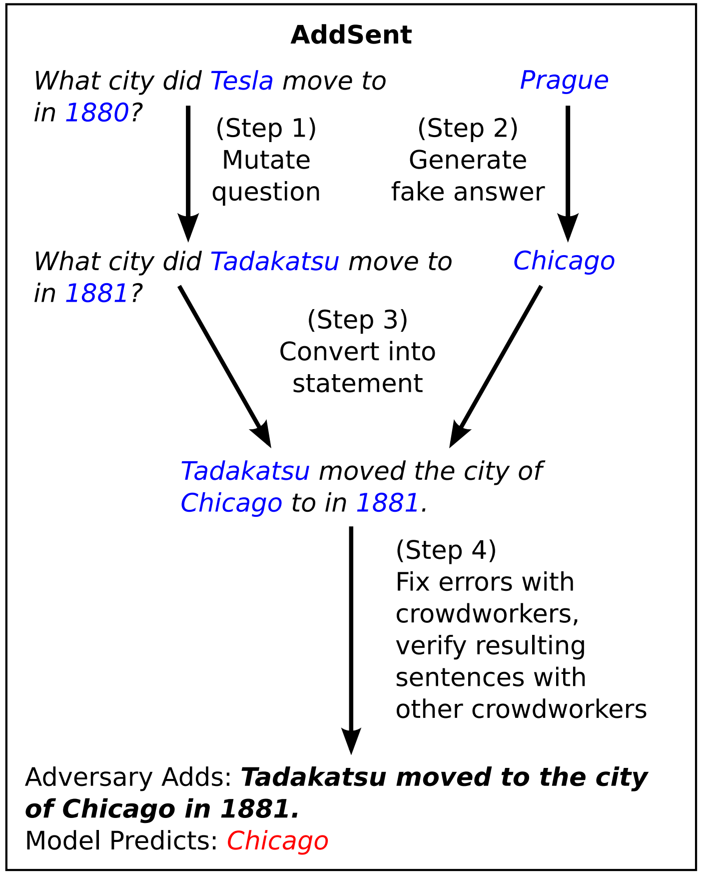
Robust Machine Comprehension Models via Adversarial Training & Model Improvements

Yicheng Wang Mohit Bansal

NAACL 2018

Robust Q&A Models: Motivation

It has been shown by Jia & Liang (2017) that many reading comprehension models trained on SQuAD lack robustness to semantics-based attacks and lose performance severely on these adversarial evaluations. Moreover, adversarial training has limited effects to bring back accuracy.



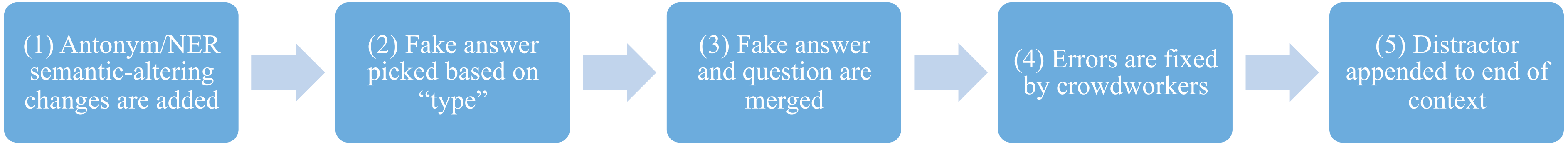
Model	Original	ADDSENT	ADDONESENT
ReasoNet-E	**81.1**	39.4	49.8
SEDT-E	80.1	35.0	46.5
BiDAF-E	80.0	34.2	46.9
Mnemonic-E	79.1	**46.2**	**55.3**
Ruminating	78.8	37.4	47.7
jNet	78.6	37.9	47.0
Mnemonic-S	78.5	**46.6**	**56.0**
ReasoNet-S	78.2	39.4	50.3
MPCM-S	77.0	40.3	50.0
SEDT-S	76.9	33.9	44.8
RaSOR	76.2	39.5	49.5
BiDAF-S	75.5	34.3	45.7
Match-E	75.4	29.4	41.8
Match-S	71.4	27.3	39.0
DCR	69.3	37.8	45.1
Logistic	50.4	23.2	30.4

	Training data	
Test data	Original	Augmented
Original	75.8	75.1
ADDSENT	34.8	70.4
ADDSENTMOD	34.3	39.2

[Jia and Liang, EMNLP 2017]

Improved Adversarial Training

- **AddSent** (Jia and Liang, 2017) is a five-step process that generates distractors which are syntactically similar to the question but semantically different:



For more effective adversarial training, we make changes to step (5) and step (2) to make the generated adversaries more diverse and hard-to-overfit (**AddSentDiverse**):

- **Random Distractor Placement:** To prevent the trained model from over-fitting the adversary by ignoring the last sentence, we *<u>randomly insert the sentence into the paragraph</u>*.

- **Dynamic Fake Answer Generation**: To prevent the trained model from having any bias toward a specific set of 'fake answers', we *<u>dynamically generate a fake answer that has the same 'type' as the real answer</u>*.

- Propose the addition of **synonymy/antonymy lexical semantic features** using WordNet to enhance a model's overall capabilities in detecting semantics-altering perturbations (which effectively complements adversarial training; improves adv-eval performance by an average of 36.5%).

[Wang and Bansal, NAACL 2018]

Improved Adversarial Training

Question: Who originally proposed the Alaska Permanent Fund?

Answer: Governor Keith Miller

(1) Add semantics-altering perturbations to the question

Answer Type: NER-Person

Who originally proposed the Idaho Temporary Investment?

SQuAD Answer Set

Ariel Sharon

(2) Dynamically generate fake answer of the same type

(3) Combine into Statement

Ariel Sharon originally proposed the Idaho Temporary Investment.

(5) Randomly insert into context

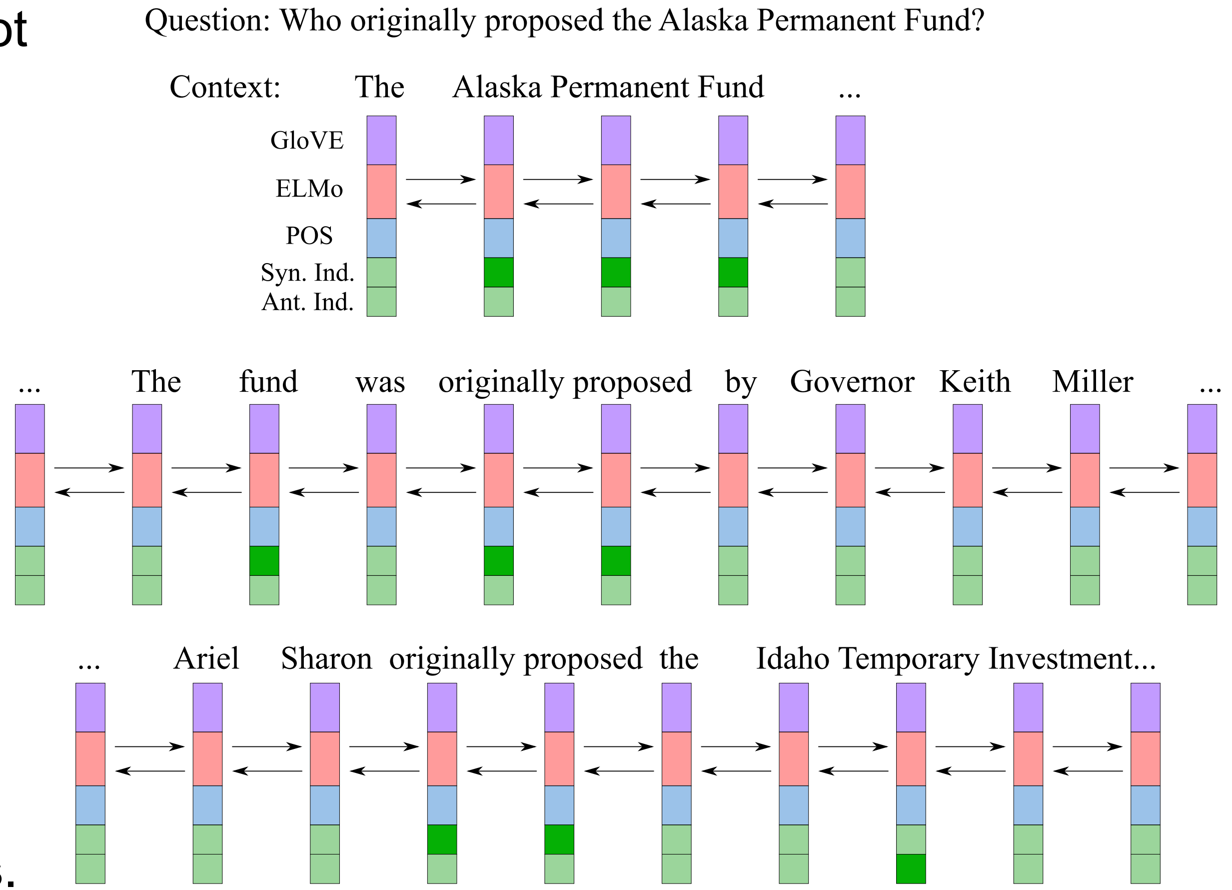
The Alaska Permanent Fund is a constitutionally authorized appropriation of oil revenues, established by voters in 1976 to manage a surplus in state petroleum revenues from oil, largely in anticipation of the recently constructed Trans-Alaska Pipeline System. The fund was originally proposed by Governor Keith Miller on the eve of the 1969 Prudhoe Bay lease sale, out of fear that the legislature would spend the entire proceeds of the sale (which amounted to $900 million) at once. Ariel Sharon originally proposed the Idaho Temporary Investment. It was later championed by Governor Jay Hammond and Kenai state representative Hugh Malone. It has served as an attractive political prospect ever since, diverting revenues which would normally be deposited into the general fund.

[Wang and Bansal, NAACL 2018]

Model Enhancements

WordNet Model Enhancements: Models cannot be fully resilient to semantics-based attacks with only adversarial training, because its inputs are bad at capturing named-entities & antonyms:

- Models use word embeddings trained on hypothesis: 'words that occur in similar contexts have similar meanings'; this is not true for antonyms & NERs.
- We add two indicator features for the existence of synonyms and antonyms in the other input (context or query).
- Synonym indicators effective at distinguishing named entity neighbors from actual synonyms.
- Antonym indicators effective at finding subtle yes crucial opposite meanings.



[Wang and Bansal, NAACL 2018]

Results

Setup/Results Summary: Our experiments were done on the BiDAF + Self-Attention + ELMo (BSAE) (Peters et al., 2018) model:

- We see that adversarial training with one type of adversary does not generalize to other, similar adversaries.
- We see that inserting distractors in the middle, while not biased, performs poorly compared to random insertion.
- We see that using a fixed set of fake answers causes the model to overfit on those fake answers, and hurts overall robustness.
- We see that the addition of lexical WordNet features is only effective when used jointly with adversarial training (because the model now has the capacity to understand +utilize the adversarial training data's tricky information). It also prevents the decrease in regular task performance during adversarial training.

[Wang and Bansal, NAACL 2018]

Results

Adversarial Training Results:

Training	SQuAD-Dev	AddSent	AddSent Prepend	AddSent Random	AddSent Mod	Average
Original	**84.65**	42.45	41.46	40.48	41.96	50.20
AddSent	83.76	**79.55**	51.96	59.03	46.85	64.23
AddSentDiverse	83.49	76.95	**77.45**	**76.02**	**77.06**	**78.19**

Random Distractor Placement Results:

Training	AddSent	AddSentPrepend	Average
InsFirst	60.22	**79.81**	70.02
InsLast	**79.54**	51.96	65.75
InsMid	74.74	74.33	74.54
InsRandom	76.33	77.38	**76.85**

[Wang and Bansal, NAACL 2018]

Results

Dynamic Fake Answer Generation Results:

Training	AddSentPrepend	AddSentMod
Fixed-FakeAns	77.37	73.65
Dynamic-FakeAns	**77.45**	**77.06**

Model Enhancement Results:

Model/Training	SQuAD-Dev	AddSent
BSAE/Reg.	**84.65**	42.45
BSAE/Adv.	83.49	76.95
BSAE+SA/Reg.	84.62	44.60
BSAE+SA/Adv.	84.49	**78.91**

[Wang and Bansal, NAACL 2018]

Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

Yichen Jiang

Mohit Bansal

ACL 2019

Data/code available at
https://github.com/jiangycTarheel/Adversarial-MultiHopQA

Single-Hop QA

[Rajpurkar et al., 2016]

Question

"Which NFL team represented the AFC at Super Bowl 50?"

Single-Hop QA

[Rajpurkar et al., 2016]

Question

"Which NFL team represented the AFC at Super Bowl 50?"

Context

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers …

[Jiang and Bansal, ACL 2019]

Single-Hop QA

[Rajpurkar et al., 2016]

Question

"Which NFL team represented the AFC at Super Bowl 50?"

Context

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers …

Answer

"Denver Broncos"

[Jiang and Bansal, ACL 2019]

Multi-Hop QA

[Yang et al., 2018]

Question

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

[Jiang and Bansal, ACL 2019]

Multi-Hop QA

[Yang et al., 2018]

Question

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

Kasper Schmeichel $\xrightarrow{son\_of}$??? $\xrightarrow{voted\_as}$???

[Jiang and Bansal, ACL 2019]

Multi-Hop QA

[Yang et al., 2018]

Question

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

Context

Kasper Schmeichel is a Danish professional footballer ... He is the son of former Manchester United and Danish international goalkeeper Peter Schmeichel.

Kasper Schmeichel $\xrightarrow{son\_of}$??? $\xrightarrow{voted\_as}$???

[Jiang and Bansal, ACL 2019]

Multi-Hop QA

[Yang et al., 2018]

Question

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

Context

Kasper Schmeichel is a Danish professional footballer ... He is the son of former Manchester United and Danish international goalkeeper Peter Schmeichel.

Kasper Schmeichel $\xrightarrow{son\_of}$ Peter Schmeichel $\xrightarrow{voted\_as}$???

Multi-Hop QA

[Yang et al., 2018]

Question

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

Context

Kasper Schmeichel is a Danish professional footballer ... He is the son of former Manchester United and Danish international goalkeeper Peter Schmeichel.

Peter Bolesław Schmeichel is a Danish former professional footballer … was voted the IFFHS World's Best Goalkeeper in 1992 …

Kasper Schmeichel $\xrightarrow{son\_of}$ Peter Schmeichel $\xrightarrow{voted\_as}$???

[Jiang and Bansal, ACL 2019]

Multi-Hop QA

[Yang et al., 2018]

Question

“What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?”

Context

Kasper Schmeichel is a Danish professional footballer ... He is the son of former Manchester United and Danish international goalkeeper Peter Schmeichel.

Peter Bolesław Schmeichel is a Danish former professional footballer … was voted the IFFHS World's Best Goalkeeper in 1992 …

Kasper Schmeichel $\xrightarrow{son\_of}$ Peter Schmeichel $\xrightarrow{voted\_as}$ World’s Best Goalkeeper

Bridge Entity

[Jiang and Bansal, ACL 2019]

Is ***compositional reasoning*** necessary to answer these multi-hop questions?

[Jiang and Bansal, ACL 2019]

Is ***compositional reasoning*** necessary to answer these multi-hop questions?

Reasoning Chain:

Kasper Schmeichel	$\xrightarrow{son\_of}$	Peter Schmeichel	$\xrightarrow{voted\_as}$	World's Best Goalkeeper
Question Entity		*Bridge Entity*		*Answer*

[Jiang and Bansal, ACL 2019]

Is *compositional reasoning* necessary to answer these multi-hop questions?

Not always!

[Jiang and Bansal, ACL 2019]

Reasoning Shortcut

Question

"What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?"

Reasoning Chain:

Kasper Schmeichel $\xrightarrow{son\_of}$ Peter Schmeichel $\xrightarrow{voted\_as}$ World's Best Goalkeeper

Question Entity — *Bridge Entity* — *Answer*

Reasoning Shortcut:

[Placeholder] $\xrightarrow{voted\_as}$ World's Best Goalkeeper

Answer

[Jiang and Bansal, ACL 2019]

Reasoning Shortcut

Question

"What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?"

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

[Jiang and Bansal, ACL 2019]

Reasoning Shortcut

Question

"What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?"

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

Edson Arantes do Nascimento is a retired Brazilian professional footballer. In 1999, he was **voted** World Player of the Century by **IFFHS**. [Missing: 1992]

Reasoning Shortcut

Question

“What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?”

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

Edson Arantes do Nascimento is a retired Brazilian professional footballer. In 1999, he was **voted** World Player of the Century by **IFFHS**. [Missing: 1992]

Kasper Hvidt is a Danish retired handball goalkeeper, .. also **voted** as Goalkeeper of the Year March 20, 2009, [Missing: 1992, IFFHS]

[Jiang and Bansal, ACL 2019]

Reasoning Shortcut

Question

"What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?"

The answer can be directly inferred by word-matching the documents to the question !!!

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

Edson Arantes do Nascimento is a retired Brazilian professional footballer. In 1999, he was **voted** World Player of the Century by **IFFHS**. [Missing: 1992]

Kasper Hvidt is a Danish retired handball goalkeeper, .. also **voted** as Goalkeeper of the Year March 20, 2009, [Missing: 1992, IFFHS]

[Jiang and Bansal, ACL 2019]

How to eliminate this reasoning shortcut from the data to **ENFORCE** compositional reasoning?

[Jiang and Bansal, ACL 2019]

How to eliminate this reasoning shortcut from the data to **ENFORCE** compositional reasoning?

Building **adversarial documents** as better distractors

[Jiang and Bansal, ACL 2019]

Adversarial Document

Question

"What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?"

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

Adversarial Document

R. Bolesław Kelly is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Defender **in 1992** and 1993.

[Jiang and Bansal, ACL 2019]

Adversarial Document

Question

"What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?"

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

Adversarial Document

R. Bolesław Kelly is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Defender **in 1992** and 1993.

Adversarial Document

Question

"What was the father of Kasper Schmeichel **voted to be by the IFFHS in 1992**?"

Context

Peter Bolesław Schmeichel is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Goalkeeper **in 1992** and 1993.

Adversarial Document

R. Bolesław Kelly is a Danish former professional footballer .., and was **voted** the **IFFHS** World's Best Defender **in 1992** and 1993.

A model exploiting the reasoning shortcut will now find two plausible answers!

[Jiang and Bansal, ACL 2019]

Related Works (Multi-Hop QA)

- Chen & Durrett, NAACL 2019: **Understanding Dataset Design Choices for Multi-hop Reasoning**
- Min et al., ACL 2019: **Compositional Questions Do Not Necessitate Multi-hop Reasoning**

- These two useful concurrent works identified reasoning shortcuts by **building single-hop-only models** that achieve good performance in HotpotQA.

- We create adversaries to **eliminate reasoning shortcuts**, and show that models achieving strong performance in the original HotpotQA cannot solve our adversarial examples (and we then present adversarial-training and initial model development ideas).

[Jiang and Bansal, ACL 2019]

BERT (Document Retrieval Results)

* Exact-Match scores between 2 golden documents and 2 retrieved documents

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	89.44	44.67
Train = Adv	89.03	80.14

- The performance of the BERT retrieval model trained on the regular training set **dropped** a lot when evaluated on the adversarial data.
- BERT is actually exploiting the reasoning shortcut instead of performing multi-hop reasoning.

[Jiang and Bansal, ACL 2019]

BERT (Document Retrieval Results)

* Exact-Match scores between 2 golden documents and 2 retrieved documents

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	89.44	44.67
Train = Adv	89.03	80.14

- After being trained on the adversarial data, BERT achieves significantly higher EM score in adversarial evaluation.
- Adversarial training is able to teach the model to be aware of distractors and force it not to take the reasoning shortcut, but there is still a remaining drop in performance.

[Jiang and Bansal, ACL 2019]

Bi-attention + Self-attention Baseline

* Exact-Match scores

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	43.12	34.00
Train = Adv	45.12	44.65

- The performance of the baseline trained on the regular training set **dropped** a lot when evaluated on the adversarial data.
- The model that performs well in the original data is actually exploiting the reasoning shortcut instead of performing multi-hop reasoning.

[Jiang and Bansal, ACL 2019]

Bi-attention + Self-attention Baseline

* Exact-Match scores

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	43.12	34.00
Train = Adv	45.12	44.65

- After being trained on the adversarial data, the baseline achieves significantly higher EM score in adversarial evaluation.
- Adversarial training is able to teach the model a bit to be aware of distractors and force it not to take the reasoning shortcut, but still big room for improvement.

[Jiang and Bansal, ACL 2019]

Bi-attention + Self-attention Baseline

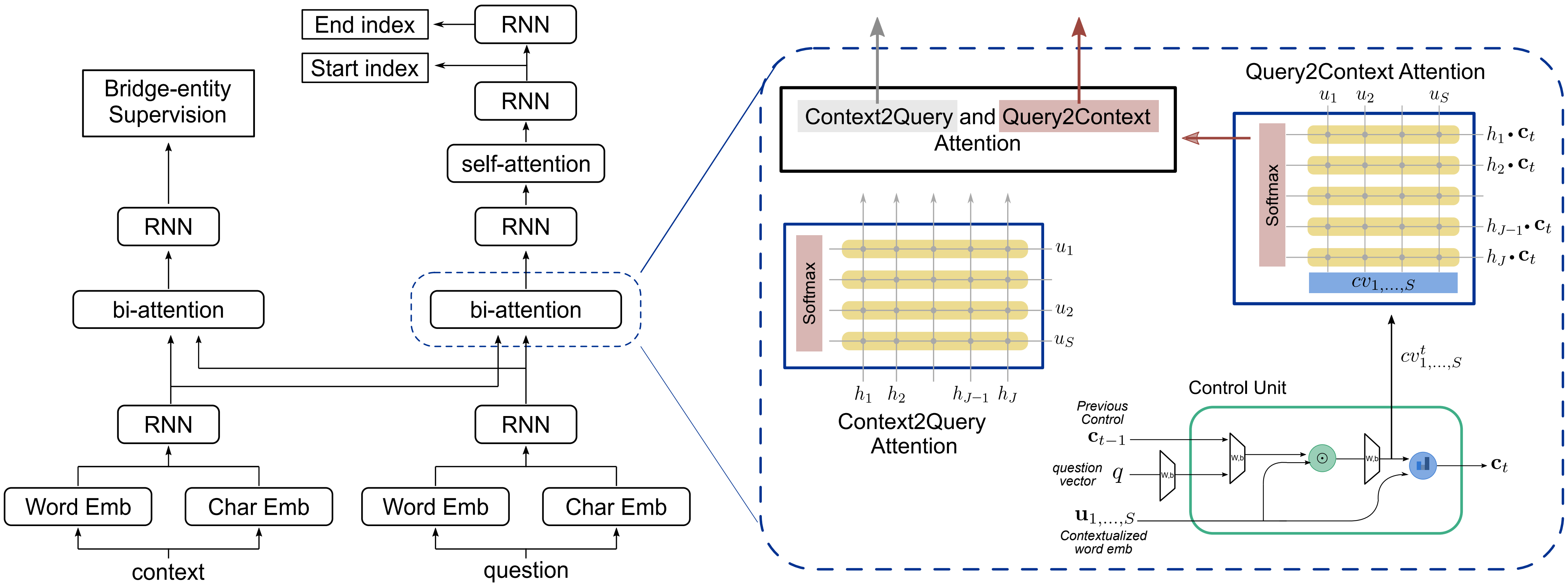
* Exact-Match scores

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	43.12	34.00
Train = Adv	45.12	44.65

- After being trained on the adversarial data, the baseline also obtains better performance in the regular evaluation.
- The multi-hop reasoning skills learnt from the adversarial data is also beneficial to the regular evaluation (and might hint that adv-trained model is not learning bad new shortcuts).

[Jiang and Bansal, ACL 2019]

An Initial 2-Hop Architecture



[Jiang and Bansal, ACL 2019]

2-Hop Model

Train \ Eval	Eval = Regular	Eval = Adv
Train = Regular	46.41	32.30
Train = Adv	47.08	46.87

[Jiang and Bansal, ACL 2019]

Analysis

- Manual Verification of Adversaries
  - 0 out of 50 examples has contradictory answers

- Model Error (Adversary Success) Analysis
  - In 96.3% of the failures, the model's prediction spans at least one of the adversarial documents

- Adversary Failure Analysis
  - Sometimes the reasoning shortcut still exists after adversarial documents are added

- **Next Steps/Questions:**
  - We might have made the model robust to one kind of attack but there might be others?
  - How do we ensure robustness to other adversaries we haven't thought of?

[Jiang and Bansal, ACL 2019]

Adversarial Over-Sensitivity and Over-Stability Strategies for Dialogue Models

Tong Niu

Mohit Bansal

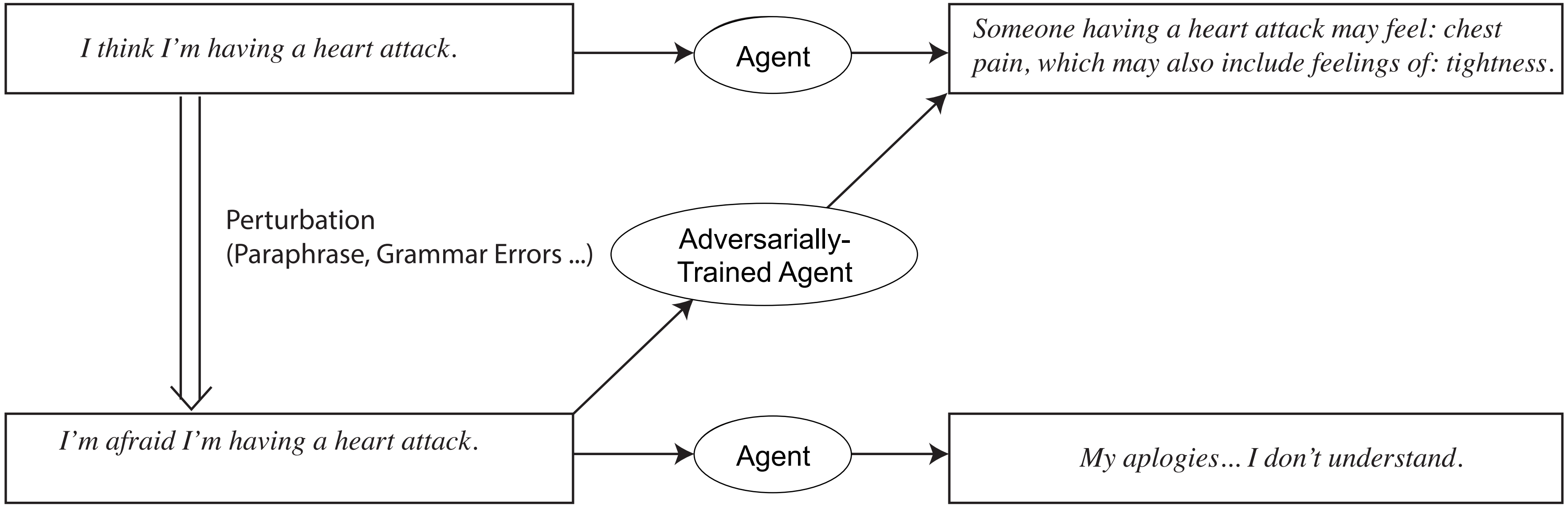
CoNLL 2018

Data/code available at
https://github.com/WolfNiu/AdversarialDialogue

Adversarial Dialogue: User-Error Robustness

We present two categories of model-agnostic adversarial strategies that reveal the weaknesses of generative, task-oriented dialogue models:

- **Should-Not-Change strategies:** evaluate over-sensitivity to small and semantics-preserving edits.
-
- **Should-Change strategies:** test if a model is over-stable against subtle yet semantics-changing modifications.



[Niu and Bansal, CoNLL 2018]

Adversarial Dialogue: User-Error Robustness

Should-Not-Change (Over-Sensitivity) Strategies on Ubuntu:

- *Random Swap*: Swap positions of neighboring words. [*I don't want you to go.* → *I don't want to you go.*]
- *Stopword Dropout*: Drop stopwords from the inputs. [*Ben ate the carrot.* → *Ben ate carrot.*]
- *Data-level Paraphrasing*: We repurpose *PPDB 2.0* (Pavlick et al., 2015) and replace words with their paraphrases. [*She bought a bike.* → *She purchased a bicycle*.]
- *Generative-level Paraphrasing*: We train *Pointer-Generator Networks* (See et al., 2017) on *ParaNMT-5M* (Wieting and Gimpel, 2017) to generate paraphrases. [How old are you? → *What's your age?*]
- *Grammar Errors*: We repurpose the *AESW* dataset (Daudaravicius, 2015), and build a look-up table to replace correct words/phrases with ungrammatical ones. [*He doesn't like cakes.* → *He don't like cake.*]

Should-Change (Over-Stability) Strategies on Ubuntu:

- *Add Negation*: Add negation to the source sequence. [*I want some coffee.* → *I don't want some coffee.*]
- *Antonym*: Change words in utterances to their antonyms. [*Please install Ubuntu.* → *Please uninstall Ubuntu.*]

[Niu and Bansal, CoNLL 2018]

Adversarial Dialogue: User-Error Robustness

Adversarial Training for Should-Not-Change Strategies: We feed "adversarial source sequence + ground-truth response pairs" as regular positive data, and teach the model that these pairs are also valid examples despite the added perturbations.

Adversarial Training for Should-Change Strategies: We use a linear combination of maximum likelihood and max-margin loss to train on negative examples.

$$L = L_{\mathrm{ML}} + \alpha L_{\mathrm{MM}}$$

$$L_{\mathrm{ML}} = \sum_i \log P(t_i | s_i)$$

$$L_{\mathrm{MM}} = \sum_i \max\left(0, M + \log P(t_i | a_i) - \log P(t_i | s_i)\right)$$

L_{ML} is the maximum likelihood loss, L_{MM} is the max-margin loss, α is the weight of the max-margin loss, M is the margin and t_i, s_i and a_i are the target sequence, normal input, and adversarial input.

- Tasks/Datasets: Ubuntu (Activity/Entity F1, Human Eval), CoCoA (Completion Rate)
- Models: VHRED, Reranking-RL, DynoNet

[Niu and Bansal, CoNLL 2018]

Adversarial Dialogue: User-Error Robustness

Strategy Name	N-train + A-test	A-train + A-test	A-train + N-test	N-train + N-test
Normal Input	-	-	-	5.94, 3.52
Random Swap	6.10, 3.42	6.47, 3.64	6.42, 3.74	-
Stopword Dropout	5.49, 3.44	6.23, 3.82	6.29, 3.71	-
Data-Level Para.	5.38, 3.18	6.39, 3.83	6.32, 3.87	-
Generative-Level Para.	4.25, 2.48	5.89, 3.60	6.11, 3.66	-
Grammar Errors	5.60, 3.09	5.93, 3.67	6.05, 3.69	-
All Should-Not-Change	-	-	6.74, 3.97	-
Add Negation	6.06, 3.42	5.01, 3.12	6.07, 3.46	-
Antonym	5.85, 3.56	5.43, 3.43	5.98, 3.56	-

Table 2: Activity and Entity F1 results of adversarial strategies on the **VHRED** model.

	VHRED	Tie	Combined-VHRED
Winning %	28	22	49

Table 5: Human evaluation results on comparison between VHRED and VHRED train on all Should-Not-Change strategies combined.

[Niu and Bansal, CoNLL 2018]

Adversarial Dialogue: User-Error Robustness

Context	Response
N: ... you could save your ubuntu files and reinstall Windows , then install ubuntu as a dual boot option __eou__ __eot__ aight buddy , so how do i get that **unknown** space back__eou__ **Random Swap:** ... you could your save ubuntu and files Windows reinstall , then install ubuntu as dual a option boot __eou__ __eot__ aight buddy , so do how i that get space **unknown** back __eou__	**NN:** you can use the Live CD , you can install Ubuntu on the same partition as the Windows partition __eou__ **NA:** I am using ubuntu . __eou__ **AA:** you can use Windows XP on the Windows partition , and then install Ubuntu on the same drive __eou__
N: ... ok then the problem is you don't have the nvidia driver __eou__ dpkg -l — grep nvidia __eou__ __eot__ i can see some nvidia thing :P __eou__ **Stopword Dropout:** ... ok then problem is you don't have nvidia driver __eou__ dpkg -l — grep nvidia __eou__ __eot__ i see nvidia thing :P	**NN:** sudo apt-get install nvidia-glx __eou__ **NA:** ok __eou__ **AA:** you can use the nvidia driver __eou__

Code: https://github.com/WolfNiu/AdversarialDialogue

[Niu and Bansal, CoNLL 2018]

Adversarial Dialog: User-Error Robustness

Context	Response
N: ... if using hdmi , the display should be set properly . except perhaps for any overscan settings __eou__ __eot__ which package do i have to install for that ? __eou__ **Data-Level Para.:** ... if using hdmi , the displaying ought to be set correctly . except maybe for any overscan setting __eou__ __eot__ which packets do i have to installs for that ? __eou__	**NN:** you can use the nvidia-settings tool to install the drivers . __eou__ **NA:** the monitor is the monitor . __eou__ **AA:** you can use the nvidia-settings tool to see if the video card is installed . __eou__
N: at this point driver activated but not in use __eou__ __eot__ did you uninstall the old driver first ? __eou__ **Generative-Level Para.:** driver activated but not in use __eou__ __eot__ did you uninstall the old driver ? __eou__	**NN:** i installed the driver from nvidia website __eou__ **NA:** ok __eou__ **AA:** i installed the driver from nvidia website __eou__
N: i would go with a thinkpad . they are pretty support and have decent hardware __eou__ **Grammar Errors:** ii would go with a thinkpad . they are quite support and have descent hardwares __eou__	**NN:** I know , but I don't want to buy a new laptop . __eou__ **NA:** Ubuntu . __eou__ **AA:** I know , but I don't want to buy a new laptop . __eou__

[Niu and Bansal, CoNLL 2018]

Adversarial Dialog: User-Error Robustness

Context	Response
N: suggest you reinstall grub . assuming no other changes , your root system and data should be present and safe . perhaps the bootloader scrambled somehow ? __eou__ **Add Negation:** suggest you don't reinstall grub . assuming no other changes , your root system and data should be present and safe . perhaps the bootloader scrambled somehow ? __eou__	**NN:** I have a Windows XP partition , and I want to install Ubuntu on it . __eou__ **NA:** I have a Windows XP partition , and I want to install Ubuntu on it . __eou__ **AA:** I don't know how to do that , but I don't know how to do that . __eou__
N: 100% of your RAM . You have it to be used . __eou__ __eot__ what what ? __eou__ **Antonym:** 100% of your RAM . You refuse it to be used . __eou__ __eot__ what what ? __eou__	**NN:** you have a lot of RAM . __eou__ **NA:** you have a lot of RAM . __eou__ **AA:** I don't know . I don't use it much . __eou__

[Niu and Bansal, CoNLL 2018]

Analyzing Compositionality-Sensitivity of NLI Models

Yixin Nie Yicheng Wang Mohit Bansal

AAAI 2019

Data/code available at
https://github.com/easonnie/analyze-compositionality-sensitivity-NLI

Overall Analysis Process

- **Adversarial Evaluation**
  - Reveal NLI models' limited compositionality-awareness and their over-reliance on lexical features.
- **Compositionality-Removal Analysis**
  - Reveal the limitations of current evaluation.
- **Compositional-Sensitivity Testing**
  - Provide a tool to explicitly analyze models' compositionality-sensitivity and better evaluation subsets.

[Gururangan et al. (2018); Poliak et al. (2018b); Tsuchiya (2018); Zhao, Dua, Singh, 2018; Nie, Wang, Bansal, AAAI 2019]

Importance of NLI

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

(Premise, Hypothesis) → Label { Entailment, Contradiction, Neutral }

[Dagan et al., 2006; Harabagiu and Hickl, 2006; Bowman et al., 2015; Williams et al., 2017]

Importance and Difficulty of NLI

The concepts of entailment and contradiction are central to all aspects of natural language understanding.

Building computation systems that can recognize these relationships is essential to many NLP tasks such as question answering and summarization.

Intuitively, success in natural language inference needs a high degree of sentence-level understanding.

Sentence-level understanding requires a model to capture both lexical and compositional semantics.

[Dagan et al., 2006; Harabagiu and Hickl, 2006; Bowman et al., 2015; Williams et al., 2017]

Importance and Difficulty of NLI

SNLI leaderboard

Other neural network models				
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Jianpeng Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code)	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Chuanqi Tan et al. '18	150D Multiway Attention Network	14m	94.5	88.3
Xiaodong Liu et al. '18	Stochastic Answer Network	3.5m	93.3	88.5
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Boyuan Pan et al. '18	300D DMAN	9.2m	95.4	88.8
Zhiguo Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	88.8
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code) **Ensemble**	17m	92.3	88.9
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network	6.7m	93.1	88.9
Zhuosheng Zhang et al. '18	SLRC	6.1m	89.1	89.1
Qian Chen et al. '17	KIM **Ensemble**	43m	93.6	89.1
Ghaeini et al. '18	450D DR-BiLSTM **Ensemble**	45m	94.8	89.3

Despite their **high** performance, it is unclear if models employ compositional understanding or are simply performing **shallow** pattern matching.

Model designs indicate an **over-focus** on **lexical** information, which is **different** from human reasoning.

This motivates our analytic study of models' **compositionality-sensitivity**.

[Nie, Wang, Bansal, AAAI 2019]

Importance and Difficulty of NLI

SNLI leaderboard

Other neural network models				
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Jianpeng Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code)	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Chuanqi Tan et al. '18	150D Multiway Attention Network	14m	94.5	88.3
Xiaodong Liu et al. '18	Stochastic Answer Network	3.5m	93.3	88.5
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Boyuan Pan et al. '18	300D DMAN	9.2m	95.4	88.8
Zhiguo Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	88.8
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, code) **Ensemble**	17m	92.3	88.9
Seonhoon Kim et al. '18	Densely-Connected Recurrent and Co-Attentive Network	6.7m	93.1	88.9
Zhuosheng Zhang et al. '18	SLRC	6.1m	89.1	89.1
Qian Chen et al. '17	KIM **Ensemble**	43m	93.6	89.1
Ghaeini et al. '18	450D DR-BiLSTM **Ensemble**	45m	94.8	89.3

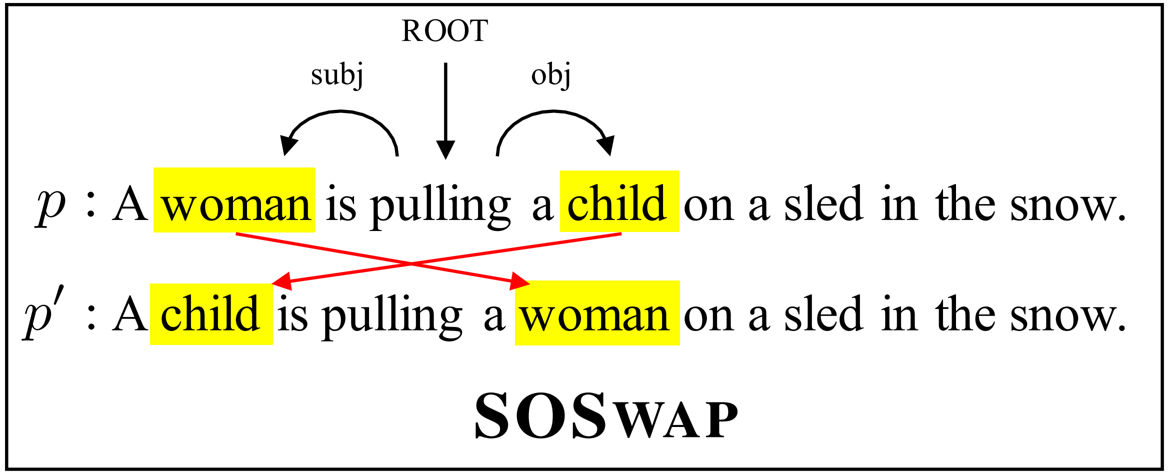
Model	SNLI	Type	Representation
RSE	86.47	Enc	Sequential
G-TLSTM	85.04	Enc	Recursive (latent)
DAM	85.88	CoAtt	Bag-of-Words
ESIM	88.17	CoAtt	Sequential
S-TLSTM	88.10	CoAtt	Recursive (syntax)
DIIN	88.10	CoAtt	Sequential
DR-BiLSTM	88.28	CoAtt	Sequential

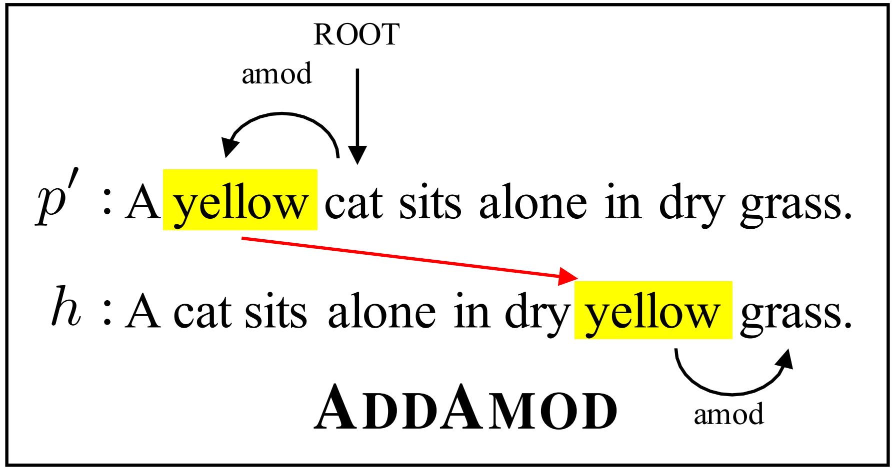
[Nie, Wang, Bansal, AAAI 2019]

Semantics-based Adversaries

Goal: To show that models are *over-reliant* on word-level information and have limited ability to process compositional structures.

Method: Created adversaries whose logical relations *cannot* be extracted from lexical information *alone*.





SubObjSwap: Take a premise with a subject-verb-object structure; create the hypothesis by swapping the subject and object.

AddAmod: Take a premise that has at least two different noun entities; pick an adjective modifier; create the premise by adding the modifier to one of the nouns, and the hypothesis by adding it to the other

[Nie, Wang, Bansal, AAAI 2019]

Semantics-based Adversaries

Most types of models fail to recognize the effects of our compositional manipulations!

Model	SNLI dev	SOSWAP E	SOSWAP **C**	SOSWAP N	ADDAMOD E	ADDAMOD C	ADDAMOD **N**
RSE	86.5	**92.5**	2.1	5.5	**95.2**	0.2	4.6
G-TLSTM	85.9	**97.2**	1.2	1.5	**95.9**	1.2	2.9
DAM	85.0	**99.7**	0.3	0.0	**99.9**	0.0	0.1
ESIM	88.2	**96.4**	2.1	1.5	**85.6**	9.6	4.8
S-TLSTM	88.1	**92.1**	4.4	3.5	**90.4**	1.1	8.5
DIIN	88.1	**84.9**	4.5	10.6	**55.0**	0.4	44.6
DR-BiLSTM	88.3	**89.7**	5.5	4.8	**82.1**	8.9	9.0
Human	-	2	**84**	14	10	2	**88**

[Nie, Wang, Bansal, AAAI 2019]

Failed Adversarial-Training Generalization

- We adv-trained the ESIM model with data augmentation from 2 adversaries, and re-evaluated. While adversarial data-augmentation leads to improvement on the same type of adversary, it does not generalize to other types of adversaries (in fact, leads to over-fitting on that particular adversary)

- This indicates that models' success on a fixed set of adversarial evaluation is still far from validating its general compositionality ability. Thus, we propose an alternative evaluation strategy that leverages existing data to evaluate a model's general compositional understanding capabilities.

	SOSWAP E/C/N	ADDAMOD E/C/N
None	96.4/2.1/1.5	85.6/9.6/4.8
SOSWAP	0.9/99.1/0.0	66.7/26.9/6.5
ADDAMOD	73.1/1.0/25.9	0.3/0.1/99.6

The percentages of predicting E/C/N by ESIM with different types of adversarial training, where an underlined number indicates the accuracy on the correct label.

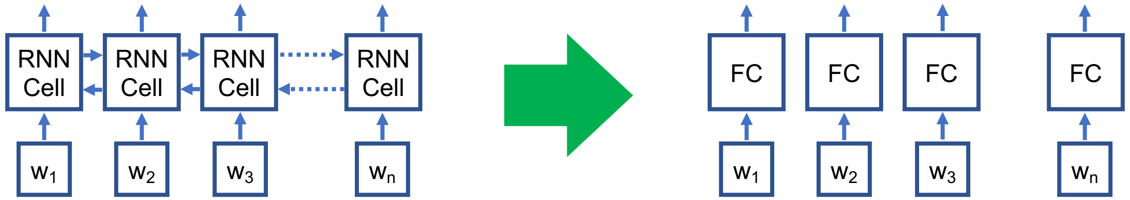
[Nie, Wang, Bansal, AAAI 2019]

Limitations of Regular Evaluation

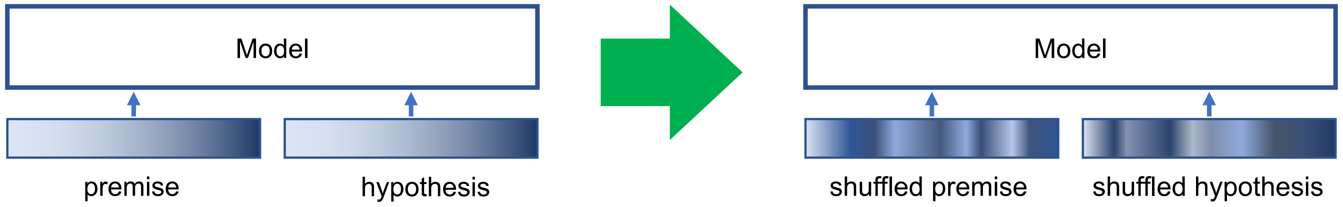
Goal: To show that regular evaluation *fails* to assess model's deeper compositional understanding.

Method: Train models with *compositional structures explicitly removed* and compare their results with those before, on regular evaluation.

<u>RNN Replacement</u>: Create strong bag-of-words-like models by replacing RNN layers with fully-connected layers, and train them on the standard training set.



<u>Word-Shuffled Training</u>: We train the NLI models with the words of the two input sentences shuffled, such that the compositional information is diluted and hard to learn.



[Nie, Wang, Bansal, AAAI 2019]

Limitations of Regular Evaluation

Removing compositional structures doesn't induce as much performance drop as expected.

Model	**SNLI**			**MNLI Matched**			**MNLI MisMatched**		
	Original	BoW	WS	Original	BoW	WS	Original	BoW	WS
RSE	86.47	85.02	–	72.80	70.02	–	74.00	71.10	–
ESIM	88.17	82.37	86.79	76.16	68.98	73.70	76.22	69.77	74.20
DR-BiLSTM	88.28	82.81	86.90	76.90	70.11	73.27	77.49	70.70	73.25

The 'Original' columns show results for vanilla models on the resp. validation sets. The 'BoW' column show results for BoW-like variants created replacing their RNNs with fully-connected layers. The 'WS' columns show results for models trained with shuffled input sentences.

[Nie, Wang, Bansal, AAAI 2019]

Lexically-Misleading Score (LMS)

Formally, we define the Lexically-Misleading Score (LMS) of an NLI datapoint (x, c^*) as:

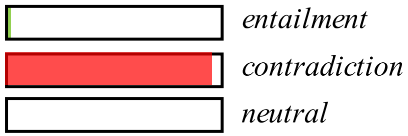
$$f_{LMS}(x, c^*) = \max_{c \in L \setminus \{c^*\}} p(c \mid x)$$

where c^* is the ground truth label, $p(c \mid x)$ is the probability generated by our regression model, and L = {entailment, contradiction, neutral} is the label set.

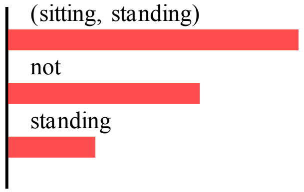
Premise: Two people are sitting in a station.
Hypothesis: A couple of people are inside and not standing.

True Label: *entailment*

Lexical Linear Model Prediction:



Top 3 misleading features

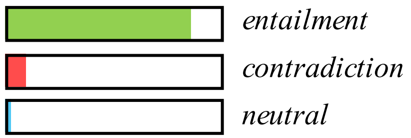


LMS: 0.9632 (to *contradiction*)

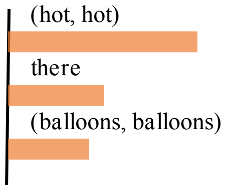
Premise: A group of people prepare hot air balloons for takeoff.
Hypothesis: There are hot air balloons on the ground and air.

True Label: *neutral*

Lexical Linear Model Prediction:



Top 3 misleading features



LMS: 0.8643 (to *entailment*)

(correct prediction for this example requires recognizing that 'not standing' and 'sitting' are the same state, rather than focusing on superficial lexical clues such as 'not' and the cross unigram ('sitting', 'standing') that both mislead to 'contradiction')

(for this example, word-overlap misleads the classifier to predict 'entailment)

[Nie, Wang, Bansal, AAAI 2019]

Compositionality-Sensitivity Evaluation Sets

Given a standard evaluation set and associated 'ground-truth' labels, $D = \{(x_i, c_i)\}_{i=1}^{N}$, we create CS_λ, the compositionality-sensitivity evaluation set of confidence λ:

$$\mathrm{CS}_\lambda = \{(x_i, c_i) \in D \mid f_{LMS}(x_i, c_i) \geq \lambda\}$$

[Nie, Wang, Bansal, AAAI 2019]

Compositionality-Sensitivity Results

Results of models, human, and majority-vote baseline on different levels of compositionality-sensitivity testing. Results of models with limited compositional information are in the bottom:

	Model	SNLI				MNLI (Matched)				MNLI (MisMatched)			
		Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**	Whole Dev	$CS_{0.5}$	$CS_{0.6}$	**$CS_{0.7}$**
1	RSE	86.47	59.01	55.59	52.73	72.80	48.48	43.57	39.62	74.00	49.30	45.84	40.85
2	G-TLSTM	85.88	57.27	53.68	50.28	70.70	45.32	41.20	38.14	70.81	46.33	42.03	38.87
3	**ESIM**	88.17	62.76	58.58	55.28	76.16	52.76	49.96	48.31	76.22	54.06	51.26	48.32
4	**S-TLSTM**	88.10	64.60	60.57	**57.51**	76.06	53.92	51.54	**48.90**	76.04	55.60	52.40	**50.61**
5	**DIIN**	88.08	64.28	60.57	**57.17**	78.70	59.49	56.12	**54.05**	78.38	59.79	57.44	**53.66**
6	DR-BiLSTM	88.28	62.92	58.50	55.28	76.90	55.26	52.72	50.07	77.49	57.39	55.37	53.04
7	Human	88.32	81.87	80.40	80.76	88.45	86.00	86.03	86.45	89.30	85.53	85.35	84.45
8	Majority Vote	33.82	42.13	42.96	43.27	35.45	36.23	35.04	35.20	35.22	34.22	35.39	34.00
	Models in which compositional information removed or diluted												
9	RSE (BoW)	85.02	52.82	47.93	43.60	70.02	40.69	34.57	31.66	71.10	43.66	38.60	34.30
10	ESIM (BoW)	82.37	48.64	44.18	40.49	68.98	38.59	33.44	30.34	69.77	41.00	35.93	32.32
11	DR-BiLSTM (BoW)	82.81	48.97	44.33	41.38	70.11	37.97	33.07	28.42	70.70	40.73	35.09	30.79
12	ESIM (WS)	86.79	58.41	50.61	45.49	73.70	44.20	41.20	41.09	74.20	49.39	45.39	41.77
13	DR-BiLSTM (WS)	86.90	58.46	50.39	44.77	73.27	45.77	41.20	37.85	73.25	46.33	42.03	38.26

[Nie, Wang, Bansal, AAAI 2019]

Next Steps / Food for Thought

- How far can adversarial training go in bringing back robustness?
- What types of direct model enhancements and better evaluations are needed for robust representation learning?
- We might have made the model robust to one kind of attack but there might be others?
- How do we ensure robustness to other adversaries we haven’t thought of?
- Should we focus on automated adversary generation or on linguistically-motivated probes?
- Important: Generalizing to other domains and languages

Part 2:
Knowledge-Rich Model Representations

Motivation and Topics

- How can we make neural models’ representations more knowledge-rich, e.g., via weak relational supervision, or via multi-task and reinforcement learning methods?

- What kinds of knowledge sources and auxiliary skills are useful?

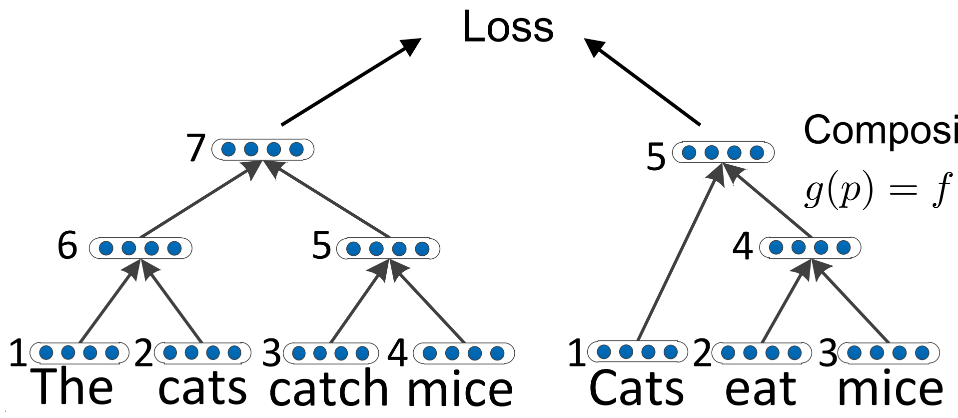
- How can we automate inductive bias and hand-designed decisions in multi-task learning?

Our Past Embeddings Work: Motivation

- Vector space representations learned on *unlabeled* linear context: distributional semantics (Harris, 1954; Firth, 1957)
- Various drawbacks:
  - capture a very generic similarity (usually topical)
  - may help one task but harm another
  - mix synonyms and antonyms, senses, similarity/relatedness (e.g., hypernymy)
- Use weak relational supervision/labels, e.g., lexicon/KB, multilingual, or task-specific (e.g., syntactic dependencies):
  - Paraphrase relation (monolingual alignments)
  - Translation relation (multilingual alignments)
  - Syntactic relation (dependency context)

[Wieting et al., TACL 2015; Bansal et al., ACL 2014; Lu et al., NAACL 2015]

Paraphrastic Embeddings



Composition =
$g(p) = f(W[g(c_1); g(c_2)] + b)$

Positive training pairs

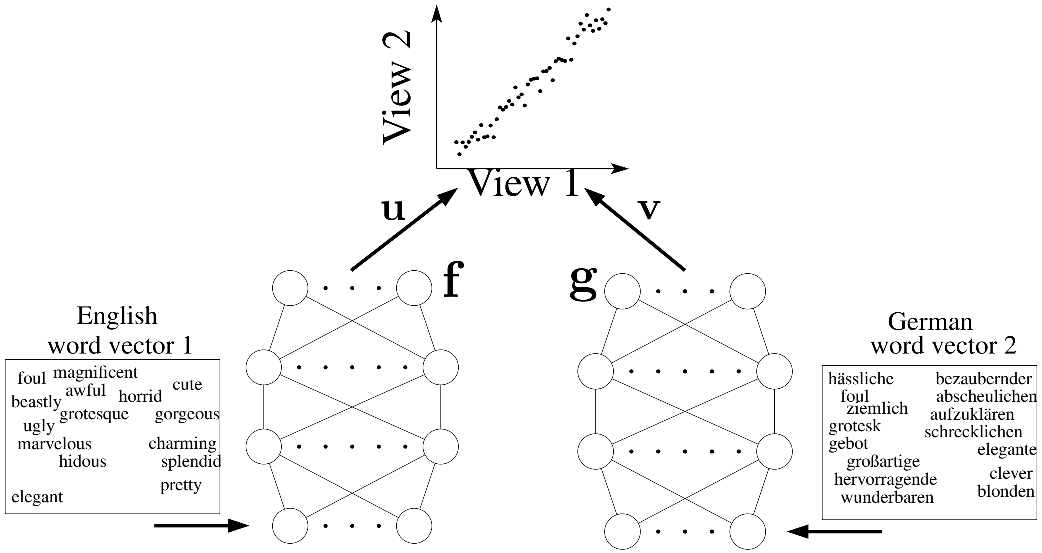
Negative training pairs

$$\min_{W,b,W_w} \frac{1}{|X|} \left(\boxed{\sum_{\langle x_1, x_2 \rangle \in X}} \max(0, \delta - \boxed{g(x_1) \cdot g(x_2)} + \boxed{g(x_1) \cdot g(t_1)}) + \max(0, \delta - \boxed{g(x_1) \cdot g(x_2)} + \boxed{g(x_2) \cdot g(t_2)}) \right)$$

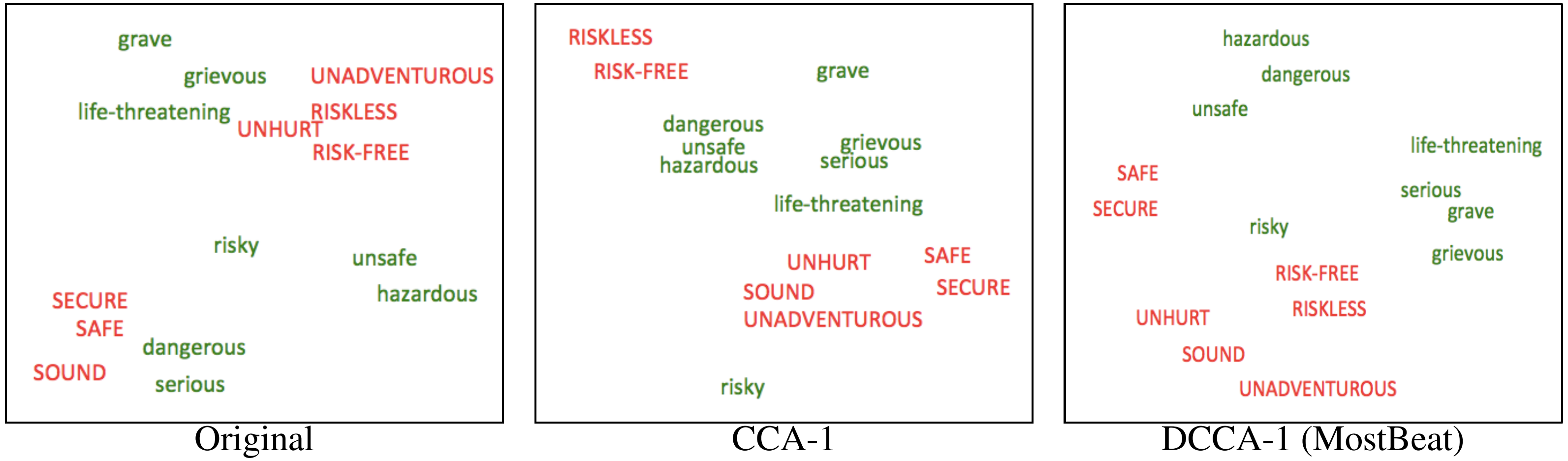
$$+ \underbrace{\lambda_W (\|W\|^2 + \|b\|^2)}_{\text{Regularization terms}} + \underbrace{\lambda_{W_w} \|W_{w_{initial}} - W_w\|^2}$$

[Wieting et al., TACL 2015; ICLR 2016]

Multilingual Deep-CCA Embeddings

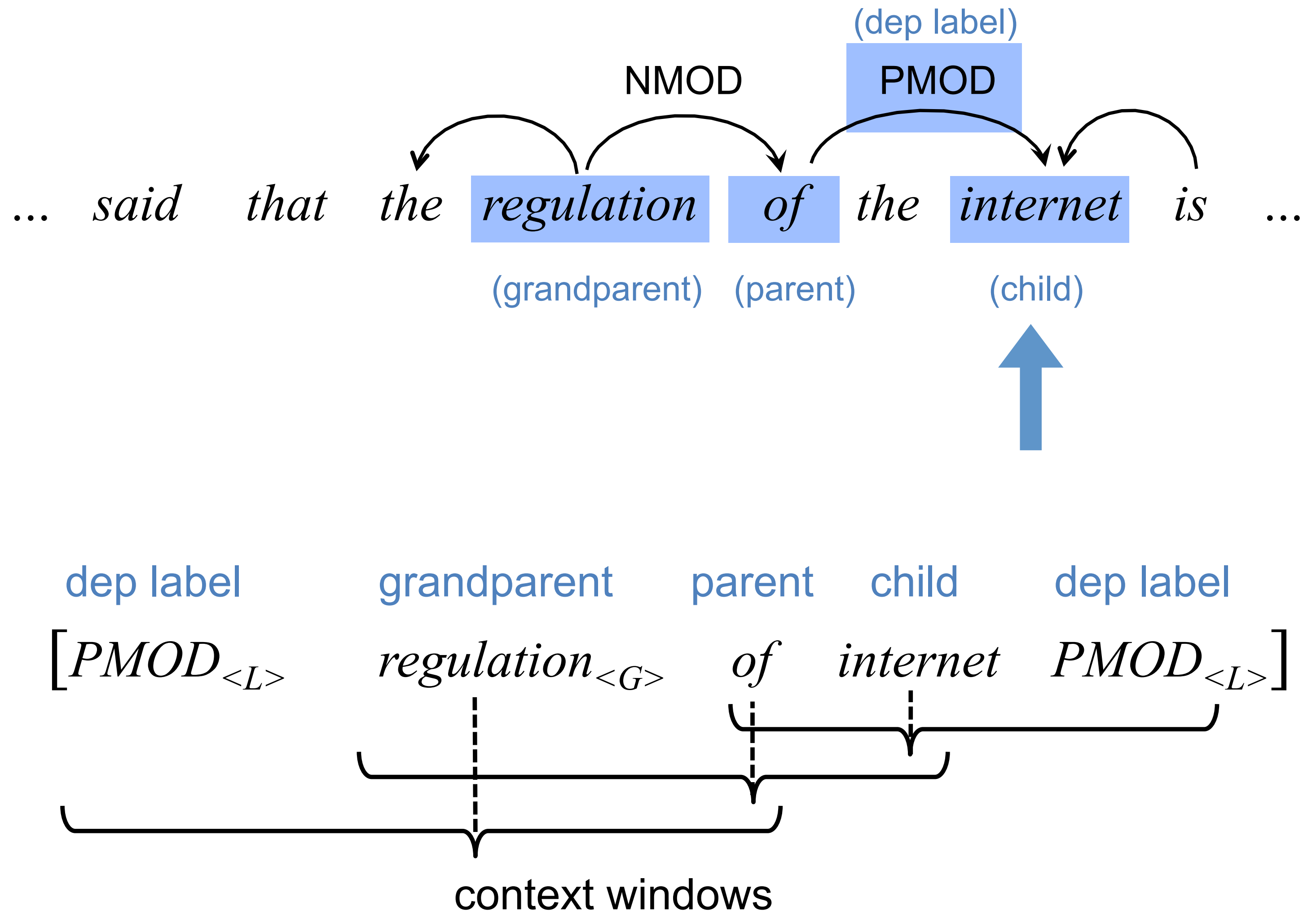


$$\max_{\mathbf{W_f},\mathbf{W_g},\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^\top \Sigma_{fg} \mathbf{v}}{\sqrt{\mathbf{u}^\top \Sigma_{ff} \mathbf{u}} \sqrt{\mathbf{v}^\top \Sigma_{gg} \mathbf{v}}}$$



[Faruqui & Dyer, EACL 2014; Lu et al., NAACL 2015]

Syntactic Dependency Embeddings



[Bansal et al., ACL 2014; Levy & Goldberg, ACL 2014]

Auxiliary Knowledge via Multi-Task Learning

- MTL: Paradigm to improve generalization performance of a task using related tasks.

- The multiple tasks are learned in parallel (alternating optimization mini-batches) while using shared model representations/parameters.

- Each task benefits from extra information in the training signals of related tasks.

- Useful survey+blog by Sebastian Ruder for details of diverse MTL papers!

[Caruana, 1998; Argyriou et al., 2007; Kumar and Daume, 2012; Luong et al., 2016; Ruder, 2017]

Auxiliary Knowledge in Language Generation

- Multi-Task & Reinforcement Learning for Entailment+Saliency Knowledge/Control in NLG (Video Captioning, Document Summarization, and Sentence Simplification)

Ground truth: A woman is slicing a red pepper.
SotA Baseline: A woman is slicing a carrot.
Our model: A woman is slicing a pepper.

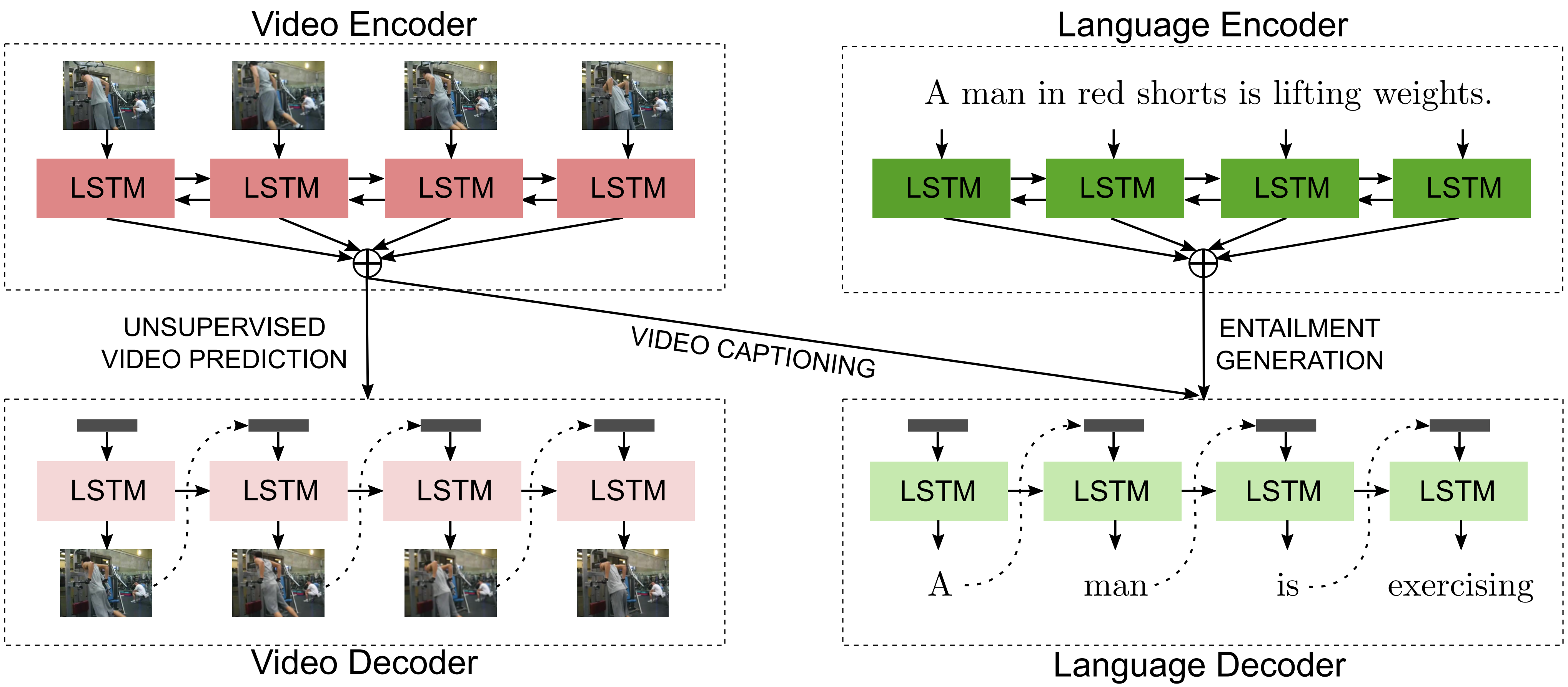
Ground truth: A group of boys are fighting.
SotA Baseline: A group of men are dancing.
Our model: Two men are fighting.

Document: *top activists arrested after last month 's anti-government rioting are in good condition , a red cross official said saturday .*
Ground-truth: *arrested activists in good condition says red cross*
SotA Baseline: *red cross says it is good condition after riots*
Our model: *red cross says detained activists in good condition*

Document: *canada 's prime minister has dined on seal meat in a gesture of support for the sealing industry .*
Ground-truth: *canadian pm has seal meat*
SotA Baseline: *canadian pm says seal meat is a matter of support*
Our model: *canada 's prime minister dines with seal meat*

Auxiliary Knowledge in Language Generation

- Many-to-Many Multi-Task Learning for Video Captioning (with Video and Entailment Generation)



[Pasunuru and Bansal, ACL 2017 (*Outstanding Paper Award*)]

Auxiliary Knowledge in Language Generation

- Reverse Multi-Task Benefits: Improved Entailment Generation

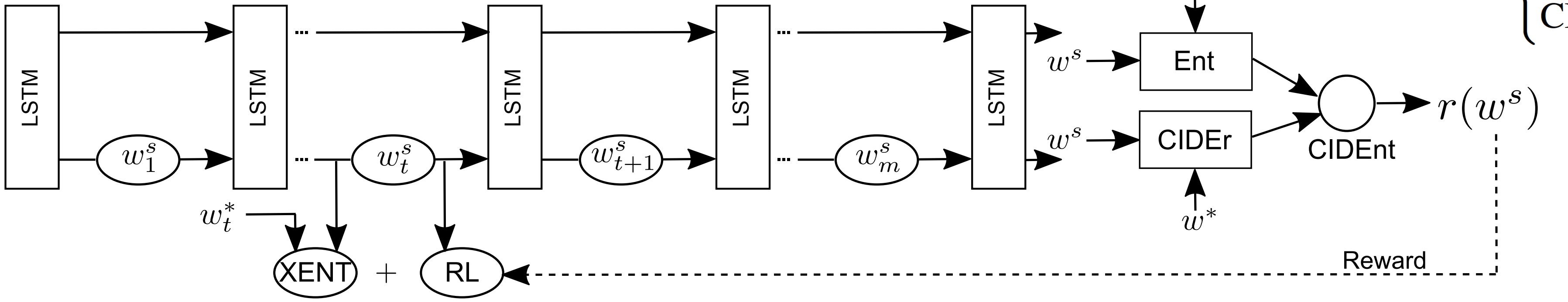
Given Premise	Generated Entailment
a man on stilts is playing a tuba for money on the boardwalk	a man is playing an instrument
a child that is dressed as spiderman is ringing the doorbell	a child is dressed as a superhero
several young people sit at a table playing poker	people are playing a game
a woman in a dress with two children	a woman is wearing a dress
a blue and silver monster truck making a huge jump over crushed cars	a truck is being driven

[Pasunuru and Bansal, ACL 2017 (*Outstanding Paper Award*)]

Auxiliary Knowledge in Language Generation

- RL Reward = Entailment-corrected phrase-matching metrics such as CIDEr → CIDEnt

Ground-truth caption	Generated (sampled) caption	CIDEr	Ent
a man is spreading some butter in a pan	puppies is melting butter on the pan	140.5	0.07
a panda is eating some bamboo	a panda is eating some fried	256.8	0.14
a monkey pulls a dogs tail	a monkey pulls a woman	116.4	0.04
a man is cutting the meat	a man is cutting meat into potato	114.3	0.08
the dog is jumping in the snow	a dog is jumping in cucumbers	126.2	0.03
a man and a woman is swimming in the pool	a man and a whale are swimming in a pool	192.5	0.02

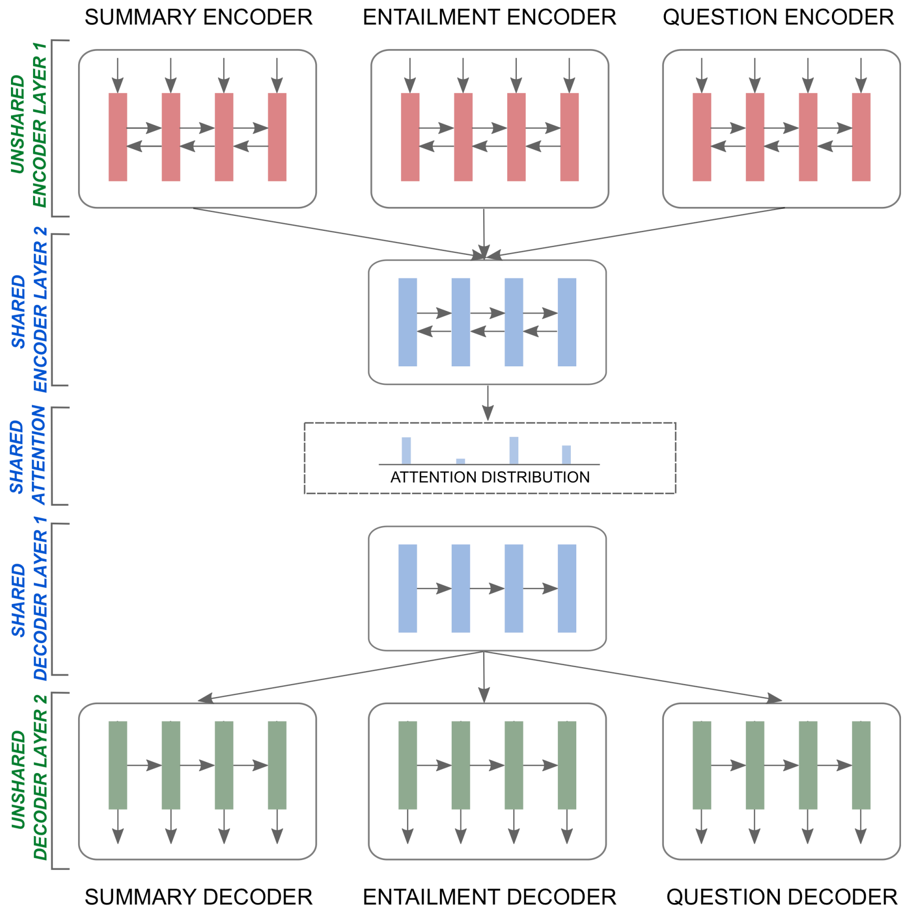


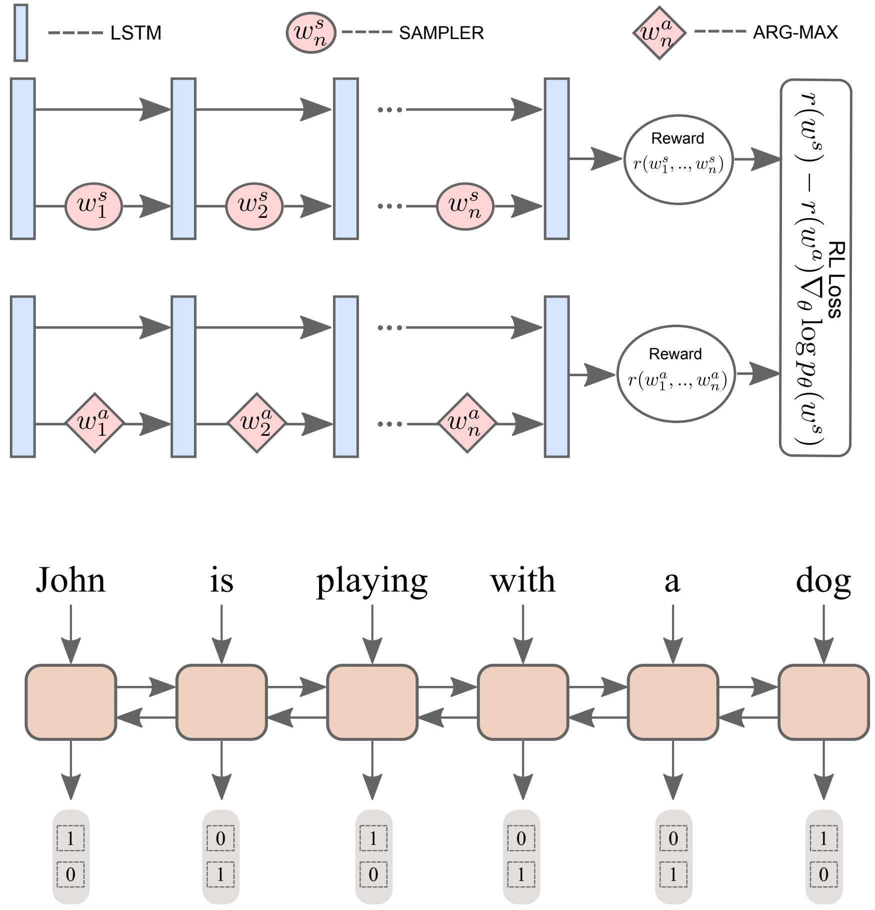
$$\text{CIDEnt} = \begin{cases} \text{CIDEr} - \lambda, & \text{if Ent} < \beta \\ \text{CIDEr}, & \text{otherwise} \end{cases}$$

[Pasunuru and Bansal, EMNLP 2017]

Auxiliary Knowledge in Language Generation

- Multi-Task & Reinforcement Learning with Entailment+Saliency Knowledge for Summarization





[Guo, Pasunuru, and Bansal, ACL 2018; Pasunuru and Bansal, NAACL 2018]

Auxiliary Knowledge in Language Generation

Input Document: celtic have written to the scottish football association in order to gain an ‘understanding’ of the refereeing decisions during their scottish cup semi-final defeat by inverness on sunday . the hoops were left outraged by referee steven mclean ’s failure to award a penalty or red card for a clear handball in the box by josh meekings to deny leigh griffith ’s goal-bound shot during the first-half . caley thistle went on to win the game 3-2 after extra-time and denied rory delia ’s men the chance to secure a domestic treble this season . celtic striker leigh griffiths has a goal-bound shot blocked by the outstretched arm of josh meekings . ……after the restart for scything down marley watkins in the area . greg tansey duly converted the resulting penalty . edward ofere then put caley thistle ahead , only for john guidetti to draw level for the bhoys . with the game seemingly heading for penalties , david raven scored the winner on 117 minutes , breaking thousands of celtic hearts . celtic captain scott brown -lrb- left -rrb- protests to referee steven mclean but the handball goes unpunished . griffiths shows off his acrobatic skills during celtic ’s eventual surprise defeat by inverness . celtic pair aleksandar tonev -lrb- left -rrb- and john guidetti look dejected as their hopes of a domestic treble end .

Ground-truth Summary: celtic were defeated 3-2 after extra-time in the scottish cup semi-final . leigh griffiths had a goal-bound shot blocked by a clear handball. however, no action was taken against offender josh meekings. the hoops have written the sfa for an ‘understanding’ of the decision .

See et al. (2017): **john hartson** was once on the end of a major **hampden injustice** while playing for celtic . but he can not see any point in his old club writing to the scottish football association over the latest controversy at the national stadium . hartson had a goal wrongly disallowed for offside while celtic were leading 1-0 at the time but went on to lose 3-2 .

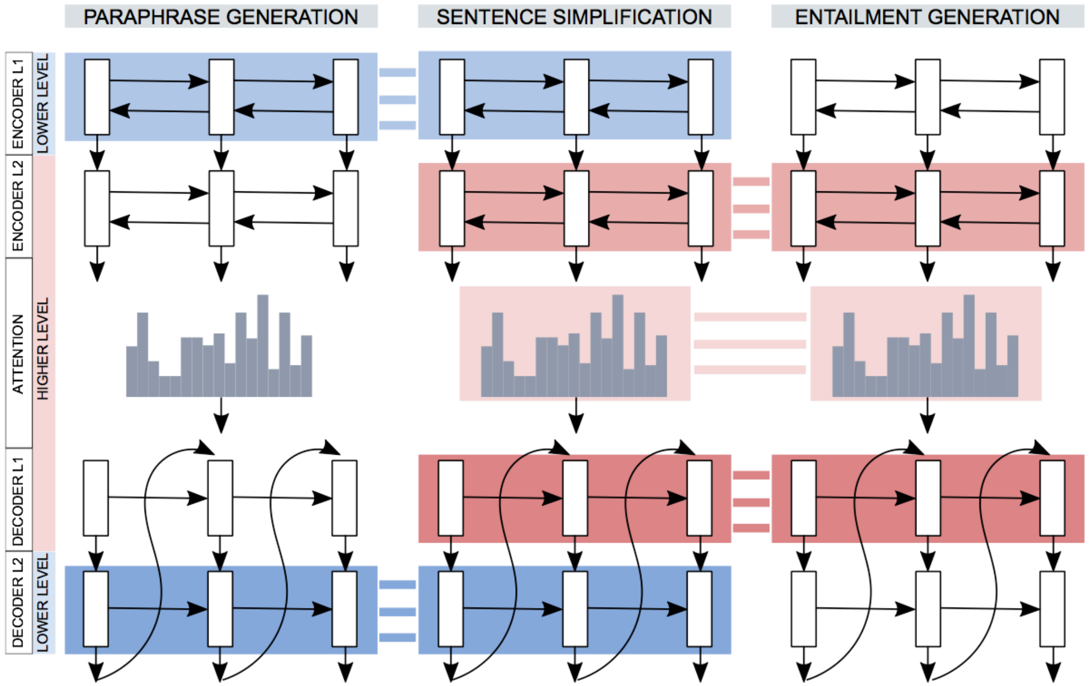
Our Baseline: **john hartson** scored the late winner in 3-2 win against celtic . celtic were leading 1-0 at the time but went on to lose 3-2 . some fans have questioned how referee steven mclean and **additional assistant alan muir** could have missed the infringement .

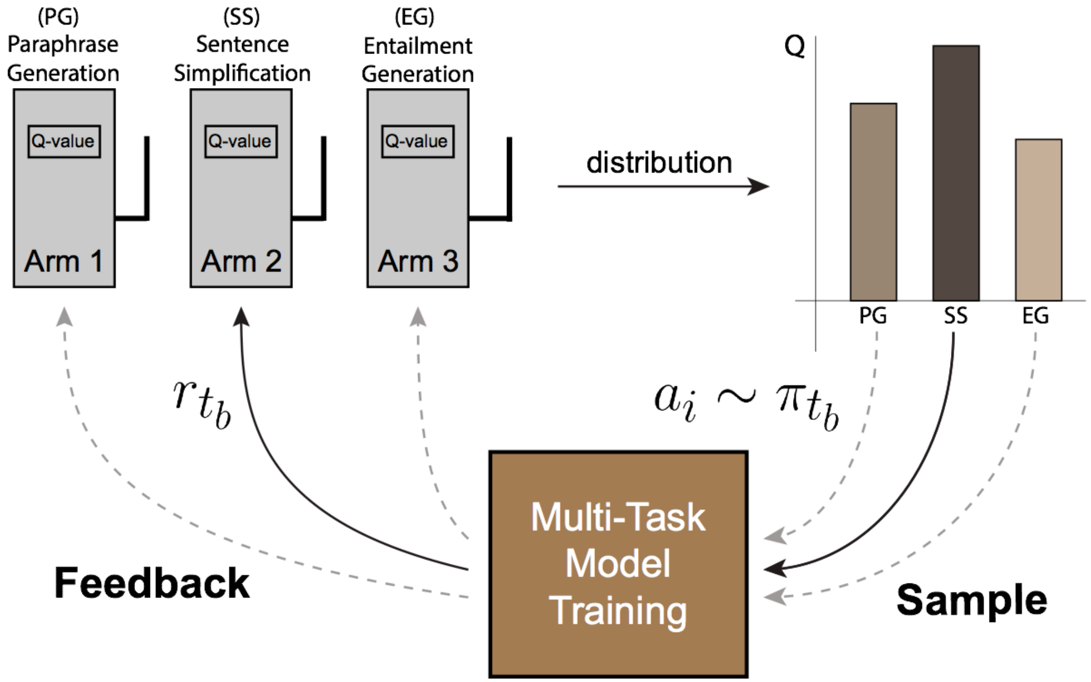
Our Multi-task Summary: celtic have written to the scottish football association in order to gain an ‘ understanding ’ of the refereeing decisions . the hoops were left outraged by referee steven mclean ’s failure to award a penalty or red card for a clear handball in the box by josh meekings . celtic striker leigh griffiths has a goal-bound shot blocked by the outstretched arm of josh meekings .

[Guo, Pasunuru, and Bansal, ACL 2018; Pasunuru and Bansal, NAACL 2018]

Auxiliary Knowledge in Language Generation

- Dynamic-Curriculum MTL with Entailment+Paraphrase Knowledge for Sentence Simplification

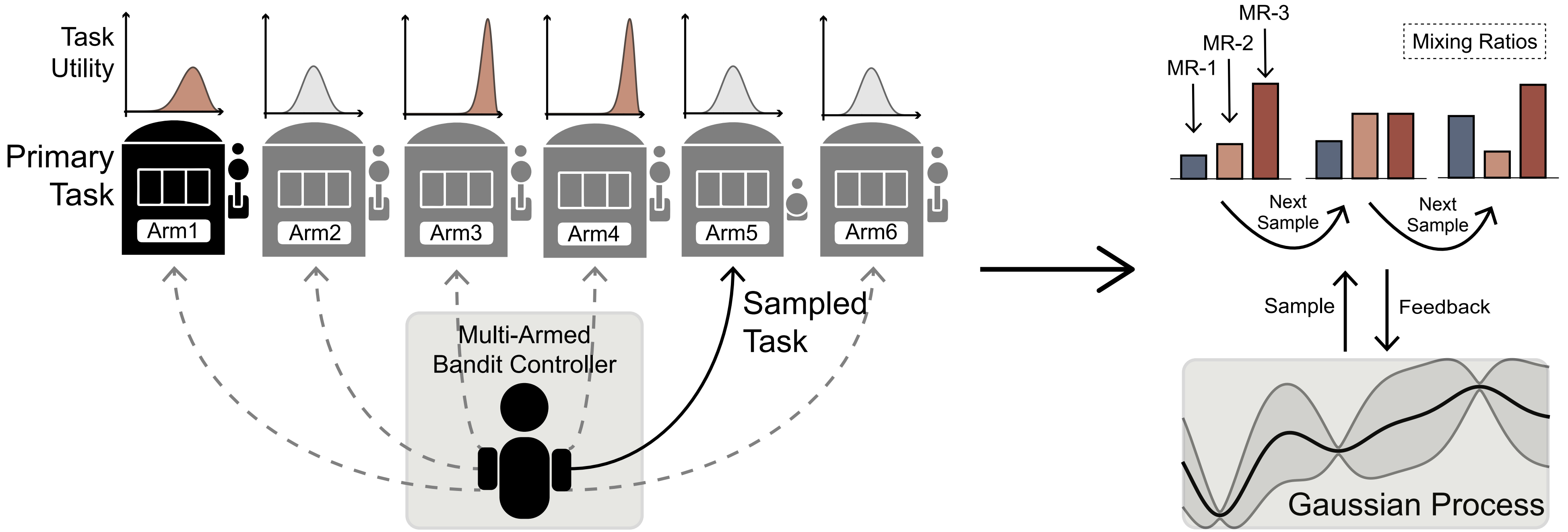




Code: https://github.com/HanGuo97/MultitaskSimplification

[Guo, Pasunuru, and Bansal, COLING 2018 (Area Chair Favorites)]

AutoSeM: Automatic Auxiliary Task Selection+Mixing



Left: the multi-armed bandit controller used for task selection, where each arm represents a candidate auxiliary task. The agent iteratively pulls an arm, observes a reward, updates its estimates of the arm parameters, and samples the next arm. Right: the Gaussian Process controller used for automatic mixing ratio (MR) learning. The GP controller sequentially makes a choice of mixing ratio, observes a reward, updates its estimates, and selects the next mixing ratio to try, based on the full history of past observations.

Code: https://github.com/HanGuo97/AutoSeM

[Guo, Pasunuru, and Bansal, NAACL 2019]

Automatic Auxiliary Task Selection

Beta Distribution

Prior Knowledge

Thompson Sampling

$$\hat{\theta}_k \sim \mathrm{Beta}\left(\alpha_k, \beta_k\right)$$

Action

Train

Observation

Bayesian Update

$$p\left(\theta_k | r\right) \propto \mathrm{Bern}_\theta(r)\, \mathrm{Beta}_{\alpha,\beta}\left(\theta_k\right)$$
$$\propto \mathrm{Beta}_{\alpha+r,\beta+1-r}\left(\theta_k\right)$$

Bernoulli Variable

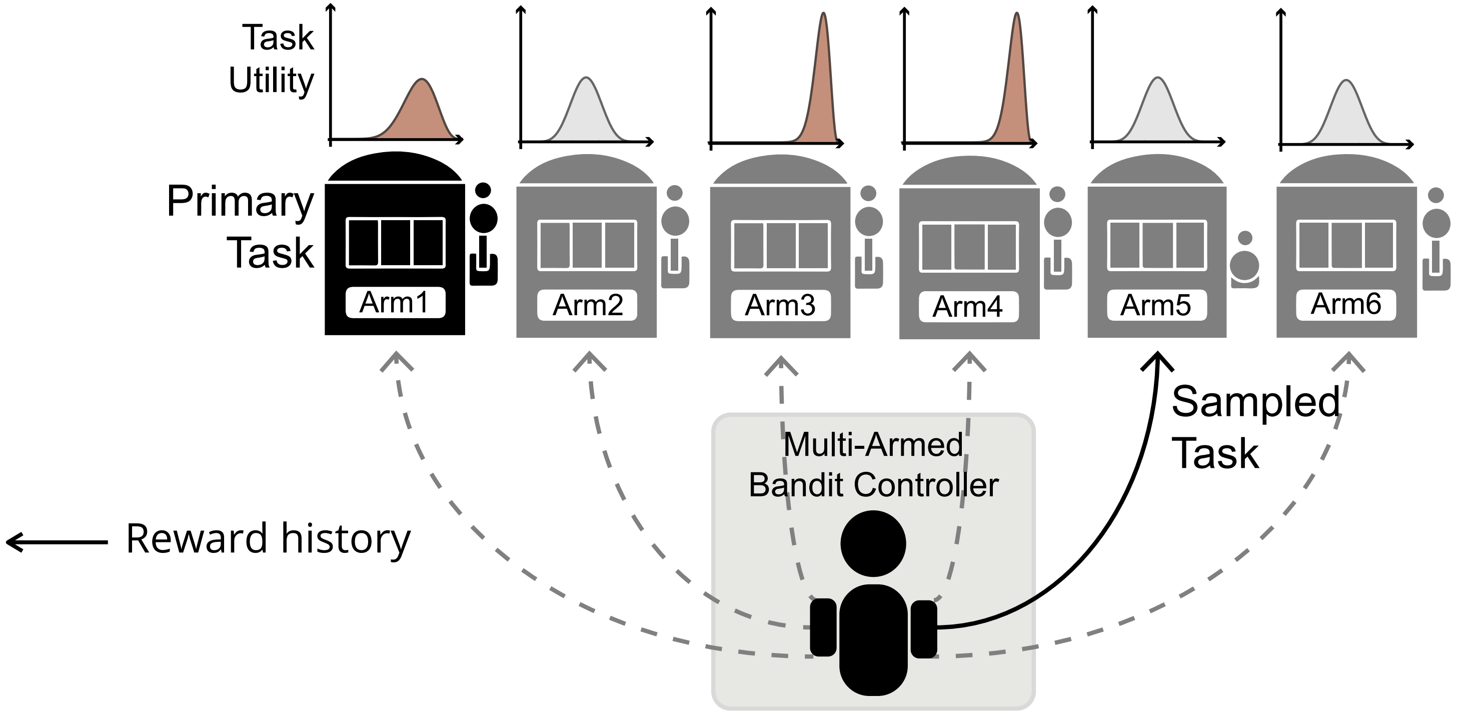
$$r_{t_b} = \begin{cases} 1, & \text{if } R_{t_b} \geq R_{t_b - 1} \\ 0, & \text{otherwise} \end{cases}$$

[Chapelle & Li, 2011; Russo et al., 2018; Guo, Pasunuru, and Bansal, NAACL 2019]

Automatic Auxiliary Task Selection

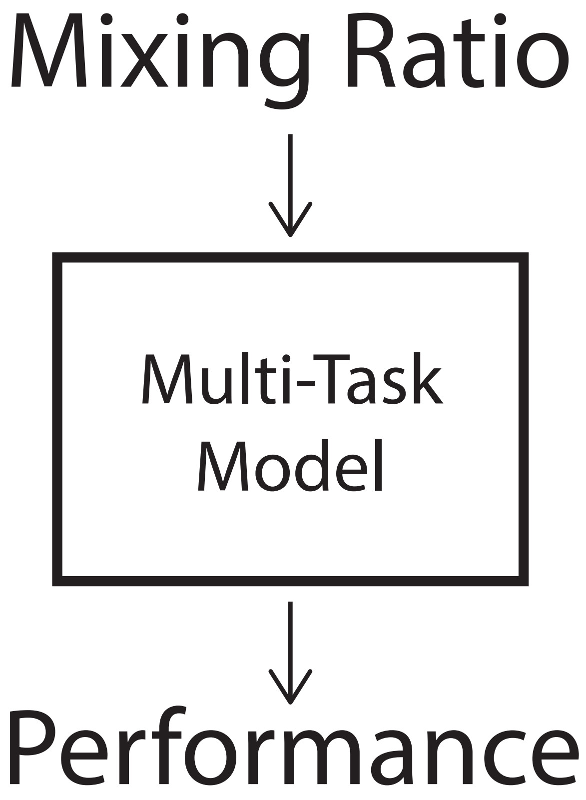
Algorithm 1 BernThompson$(N, \alpha, \beta, \gamma, \alpha_0, \beta_0)$

for $t_b = 1, 2, \ldots$ **do**
  # sample model:
  for $k = 1, \ldots, N$ **do**
    Sample $\hat{\theta}_k \sim \text{Beta}(\alpha_k, \beta_k)$
  end for
  # select and apply action:
  $x^s_{t_b} \leftarrow \arg\max_k \hat{\theta}_k$
  Apply $x^s_{t_b}$ and observe r_{t_b}
  # non-stationarity
  for $k = 1, \ldots, N$ **do**
    $\hat{\alpha}_k = (1-\gamma)\alpha_k + \gamma\alpha_0$
    $\hat{\beta}_k = (1-\gamma)\beta_k + \gamma\beta_0$
    if $k \neq x^s_{t_b}$ **then**
      $(\alpha_k, \beta_k) \leftarrow (\hat{\alpha}_k, \hat{\beta}_k)$
    else
      $(\alpha_k, \beta_k) \leftarrow (\hat{\alpha}_k, \hat{\beta}_k) + (r_{t_b}, 1 - r_{t_b})$
    end if
  end for
end for



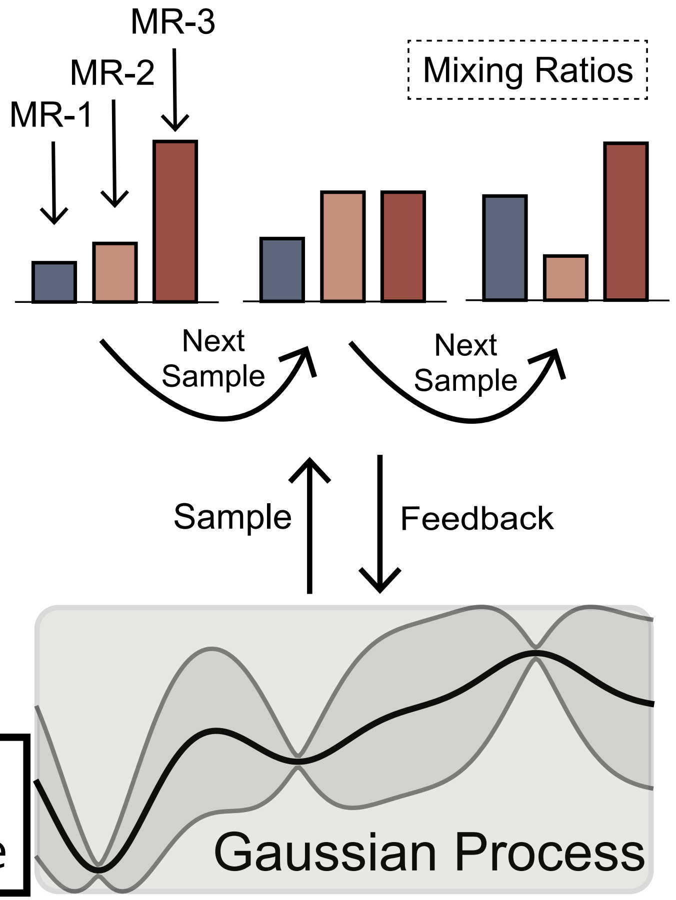
[Chapelle & Li, 2011; Russo et al., 2018; Guo, Pasunuru, and Bansal, NAACL 2019]

Automatic Mixing Ratio Curriculum Learning



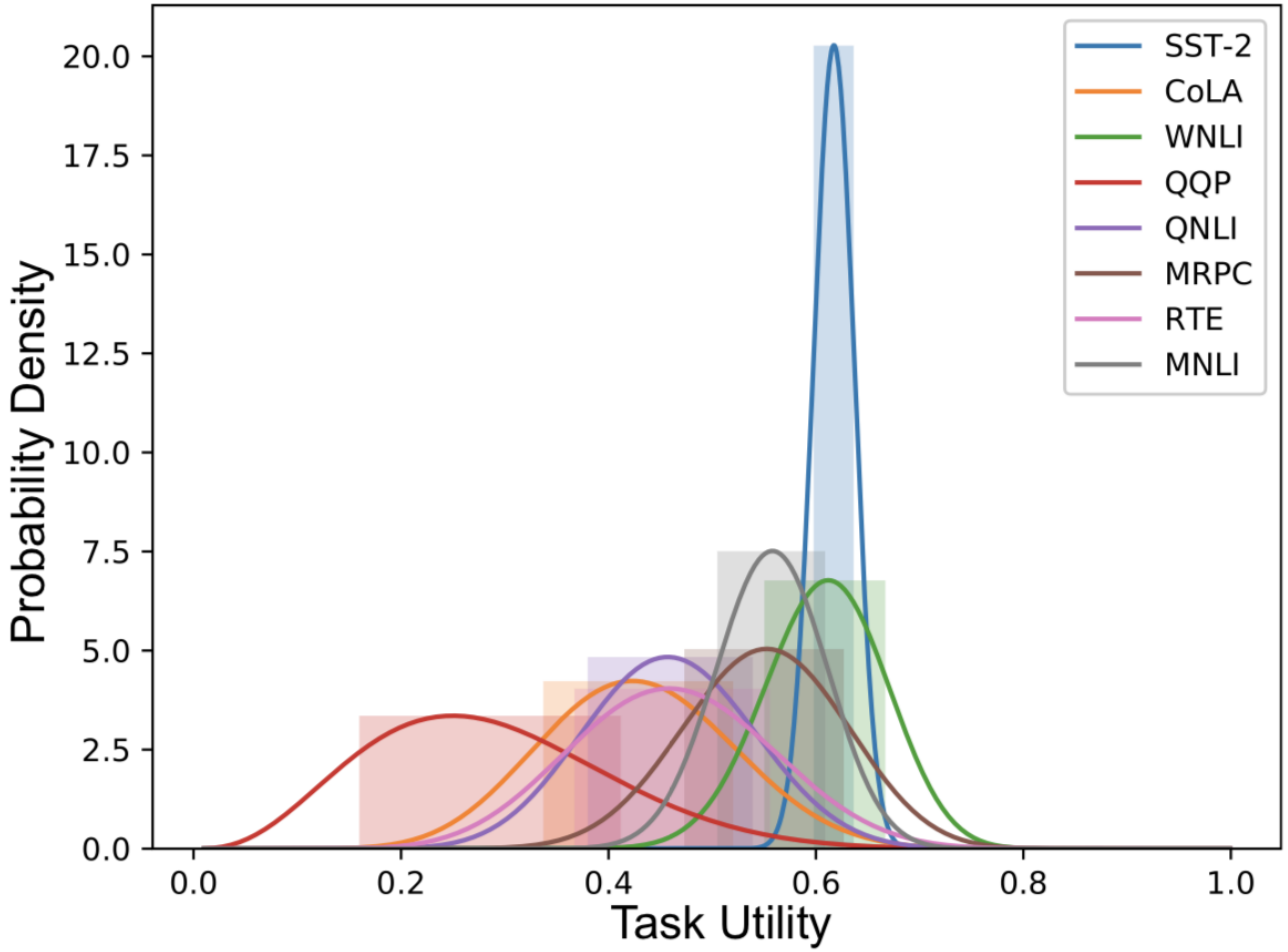
$$\boldsymbol{f}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K})$$
$$\boldsymbol{y}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 \boldsymbol{I})$$

Kernel: Matern Kernel
Acquisition Function: Hedge



[Rasmussen, 2004; Snoek et al., 2012; Shahriari et al., 2016; Guo, Pasunuru, and Bansal, NAACL 2019]

Visualization of Stage-1 Task Selection

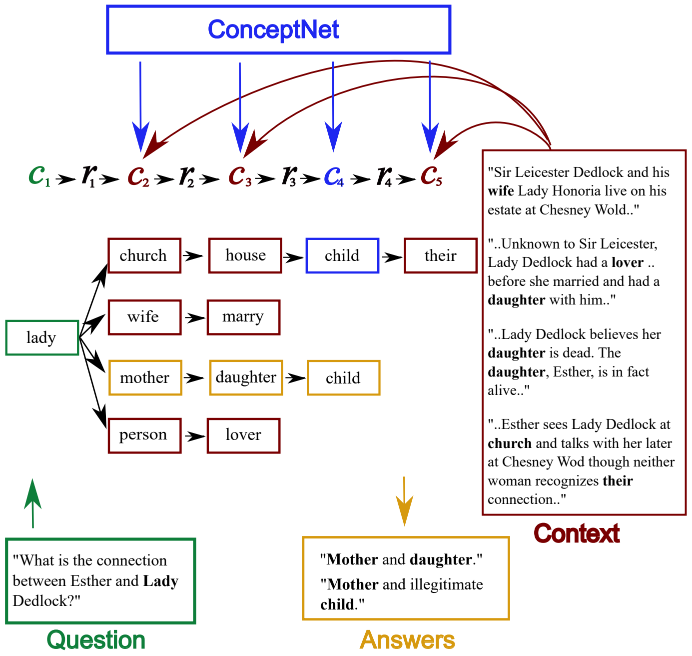


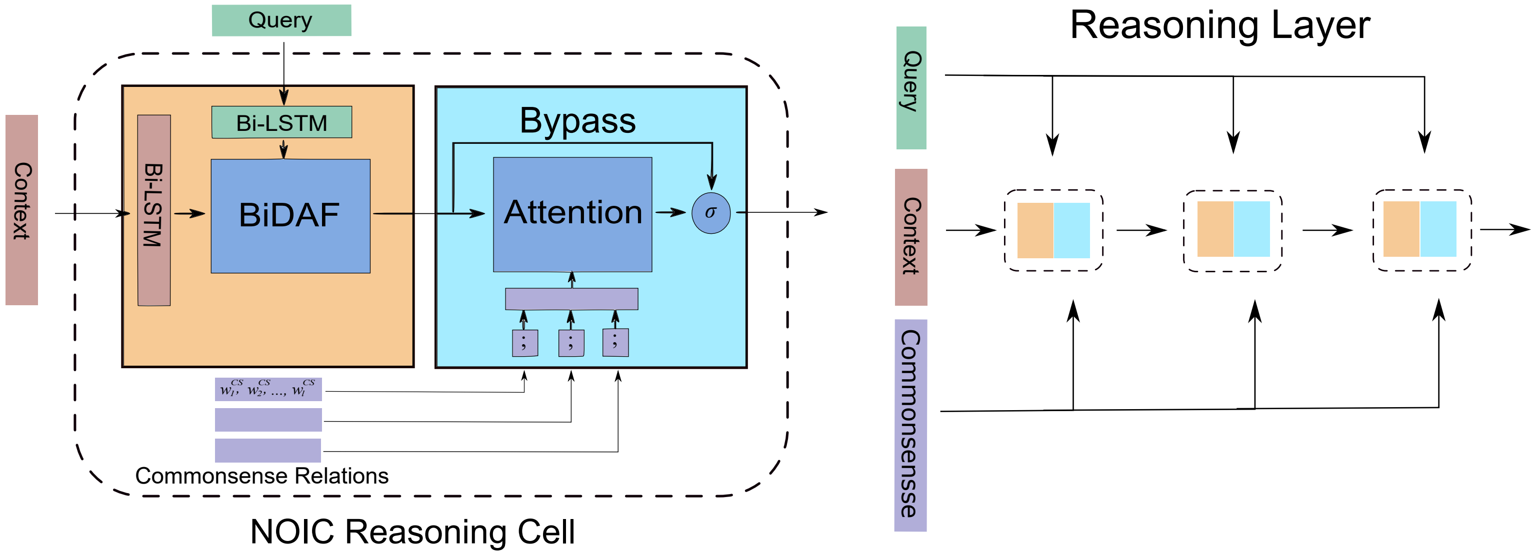
Visualization of task utility estimates from the multi-armed bandit controller on SST-2 (primary task). The x-axis represents the task utility, and the y- axis represents the corresponding probability density. Each curve corresponds to a task and the bar corresponds to their confidence interval.

[Guo, Pasunuru, and Bansal, NAACL 2019]

Commonsense in Generative Q&A Reasoning

- We use 'bypass-attention' mechanism to reason jointly on both internal context and external commonsense, and essentially learn when to fill 'gaps' of reasoning and with what information





[Bauer, Wang, and Bansal, EMNLP 2018]

Next Steps / Food for Thought

- Use of such auxiliary skill enhances MTL models for better generalization? (e.g., our MTL models transfer well to DUC test-only summarization setup in Guo et al., ACL 2018).

- Strongly promote evaluations on completely unseen and out-of-domain evaluation setups?

- Human inductive bias vs. everything learned from data?: we interpreted the learned decisions from AutoSeM (Guo et al., NAACL2019) and sometimes results do no match human intuition (e.g., the selected auxiliary tasks are not always the ones closest to the primary task), which might be due to subtle dataset noise/distribution reasons that are hard to see manually.

Students:

PhD Students


Lisa Bauer
PhD at UNC


Darryl Hannan
PhD at UNC


Peter Hase
PhD at UNC


Yichen Jiang
PhD at UNC


Hyounghun Kim
PhD at UNC
(co-advised w/ H. Fuchs)


Jie Lei
PhD at UNC
(co-advised w/ T. Berg)


Yixin Nie
PhD at UNC


Ramakanth Pasunuru
PhD at UNC


Swarnadeep Saha
PhD at UNC


Hao Tan
PhD at UNC


Shiyue Zhang
PhD at UNC


Yubo Zhang
PhD at UNC
(co-advised w/ A. Tropsha)


Xiang Zhou
PhD at UNC

Masters Students


Yen-Chun Chen
MS at UNC

Undergraduate Students


Tsion Coulter
UG at UNC


Han Guo
UG at UNC


Akshay Jain
UG at UNC


Sweta Karlekar
UG at UNC


Yicheng Wang
UG at UNC


Songhe Wang
UG at UNC

Thank you!

Webpage: http://www.cs.unc.edu/~mbansal/

Email: mbansal@cs.unc.edu

UNC-NLP Lab: http://nlp.cs.unc.edu/

Postdoc Openings!!: ~mbansal/postdoc-advt-unc-nlp.pdf