Semi-supervised Learning of MNLI by Conditional Cycle Unified Language Model with Reasoning

Haoming Jiang

Outline

- **Background**
- Challenges and Current Advances
- Proposed Method

Multi-Genre Natural Language Inference (MNLI)

- Natural Language Inference
  - (Premise, Hypothesis, Relationship)
  - Examples:
    - **Contradiction** :
    - Met my first girlfriend today. ⇔ I didn't meet my first girlfriend.
    - **Entailment**:
    - At 8:34, the Boston Center controller received a third transmission ⇔ The Boston Center controller got a third transmission
    - **Neutral**:
    - I am a lacto-vegetarian. ⇔ I enjoy eating cheese
- 5 domains for training, 10 domains for testing

Outline

- Background
- **Challenges and Current Advances**
- Proposed Method

Challenges

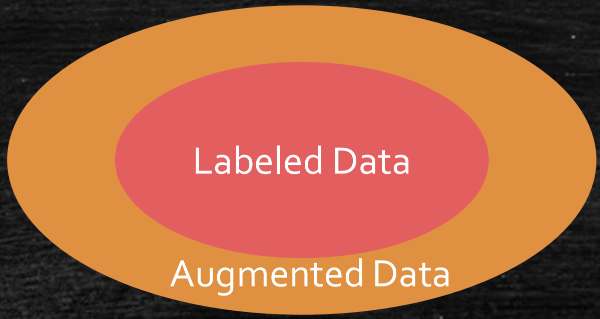
- Limited Paired Labeled Data

Challenges

- Limited Paired Labeled Data
- Solution 1: Data Augmentation
- Solution 2: Learn From Unlabeled Data: Semi-supervised Learning
- Solution 3: Learn From Unlabeled Data: Language Model Pretraining

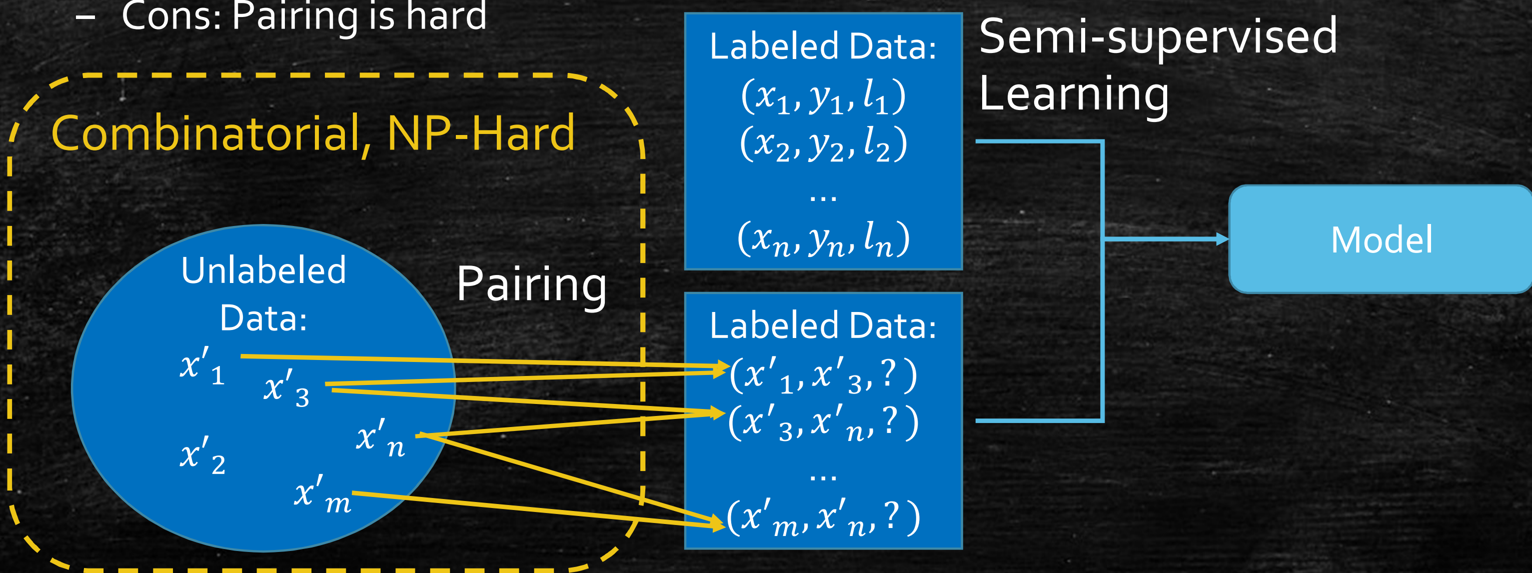
Solution

- Solution 1: Data Augmentation *(Xie et al. 2019)*
  - Back translation; TF-IDF based word replacing
  - $(x, y, l) \rightarrow (x', y', l)$
  - Pros: label is known
  - Cons: knowledge is limited



Solution

- Solution 2: Learn From Unlabeled Data: Semi-supervised Learning *(Ruder et al. 2019)*
  - Self-training, Tri-training
  - Pros: Unlimited unlabeled data
  - Cons: Pairing is hard

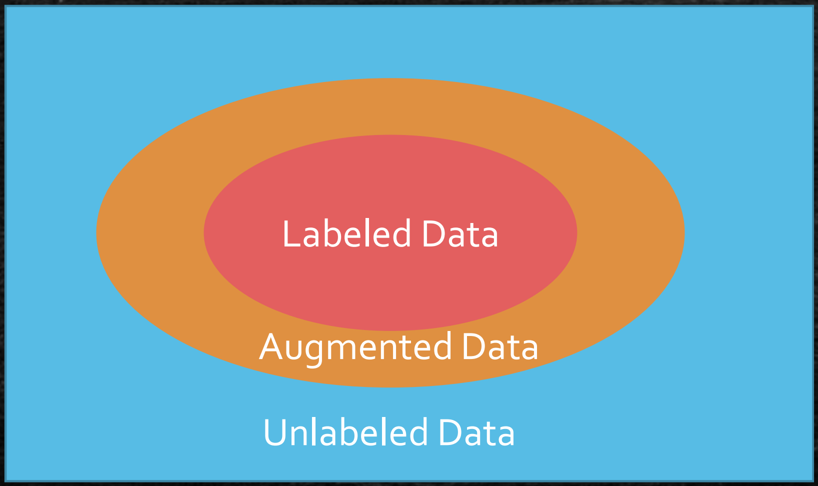


Solution

- Method 3: Learn From Unlabeled Data: Language Model Pretraining
  - Pretrain Model → Fine tuning (GPT, GPT-2, BERT,...)

Solution

- Method 3: Learn From Unlabeled Data: Language Model Pretraining
  - Pretrain Model → Fine tuning (GPT, GPT-2, BERT,…)
  - Pros: Unlimited unlabeled data
  - Cons: No label; No pair; Not task specific



Can we do better?

Method	**Amount**	**Pair**	**Label**	**Task Specific**
Data Augmentation	Limited	Yes	Yes (Ground Truth)	Yes
Semi-supervised Learning	Unlimited	Hard	Pseudo Label (Self Training)	Yes
Pretraining	Unlimited	No	No	No
Ideal	Unlimited	Reliable and Easy	Reliable and Easy	Yes

Problem 1: How to make use of unlimited unlabeled data?
Problem 2: How to find pairs for the unlabeled data?
Problem 3: How to get reliable pseudo label?

Outline

- Background
- Challenges and Current Advances
- **Proposed Method**

Proposed Method: Conditional Cycle-ULM

- Classification model $F(x)$
- Unsupervised Learning by Language Model: $Q(x)$
- Pairing by Seq2Seq: $x \to y = G(x)$

Unified Language Model (ULM)

- Label → Conditional Generation: $x, l \to y = G(x, l)$
- Reliability → Cycle Consistency:

$$y = G(x, \text{contradict}) \Leftrightarrow x = G(y, \text{contradict})$$
$$y = G(x, \text{neutral}) \Leftrightarrow x = G(y, \text{neutral})$$

Entailment is not symmetric! Two possible solutions

Chain rule: $y = G(x, \text{entail}), z = G(y, \text{entail}) \Rightarrow z = G(x, \text{entail})$

or New label: $y = G(x, \text{entail}) \Leftrightarrow x = G(y, \text{entailed by})$

Modeling:
Conditional Cycle-ULM

- Unified Language Model (ULM) *(Dong et al. 2019)*
- Powerful Probabilistic Modeling Tool
  - Language Modeling $p(x)$
  - Seq2Seq $p(y|x)$
  - Classification $p(l|x,y)$

	ELMo	GPT	BERT	UniLM
Left-to-Right LM	✓	✓		✓
Right-to-Left LM	✓			✓
Bidirectional LM			✓	✓
Seq-to-Seq LM				✓

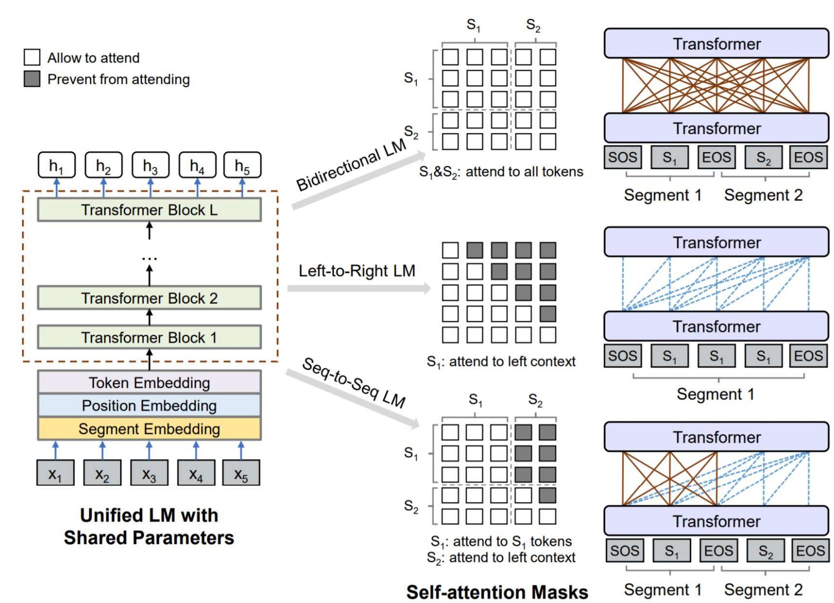


Figure 1: Overview of unified LM pre-training. The model parameters are shared across the LM objectives (i.e., bidirectional LM, unidirectional LM, and sequence-to-sequence LM). We use different self-attention masks to control the access to context for each word token. The right-to-left LM is similar to the left-to-right one, which is omitted in the figure for brevity.

Proposed Method: Conditional Cycle-ULM

- Classification model $F(x)$
- Unsupervised Learning by Language Model: $Q(x)$
- Pairing by Seq2Seq: $x \rightarrow y = G(x)$

Unified Language Model (ULM)

- Label → Conditional Generation: $x, l \rightarrow y = G(x, l)$

Conditional ULM

- Reliability → Cycle Consistency:

$$y = G(x, \text{contradict}) \Leftrightarrow x = G(y, \text{contradict})$$
$$y = G(x, \text{neutral}) \Leftrightarrow x = G(y, \text{neutral})$$

Entailment is not symmetric! Two possible solutions

Chain rule: $y = G(x, \text{entail}), z = G(y, \text{entail}) \Rightarrow z = G(x, \text{entail})$

or New label: $y = G(x, \text{entail}) \Leftrightarrow x = G(y, \text{entailed by})$

Modeling:
Conditional Cycle-ULM

- Seq2Seq models the probability in a recursive way:
$$p(y|x) = \prod p(y_t|y_{<t}, x)$$
- Conditional Seq2Seq models the conditional probability:
$$p(y|x, l) = \prod p(y_t|y_{<t}, x, l)$$
- Implementation:
  - A shared model with different heads:
$$p(y_t|y_{<t}, x, l_1) = H_1(F(x, y_{<t}))$$
$$p(y_t|y_{<t}, x, l_2) = H_2(F(x, y_{<t}))$$
$$p(y_t|y_{<t}, x, l_3) = H_3(F(x, y_{<t}))$$

Modeling:
Conditional Cycle-ULM

- Conditional ULM:

Classification $p(l|x, y)$

LM $p(x)$

Conditional Seq2Seq
$p(y|x, \text{contradict})$
$p(y|x, \text{neutral})$
$p(y|x, \text{entail})$

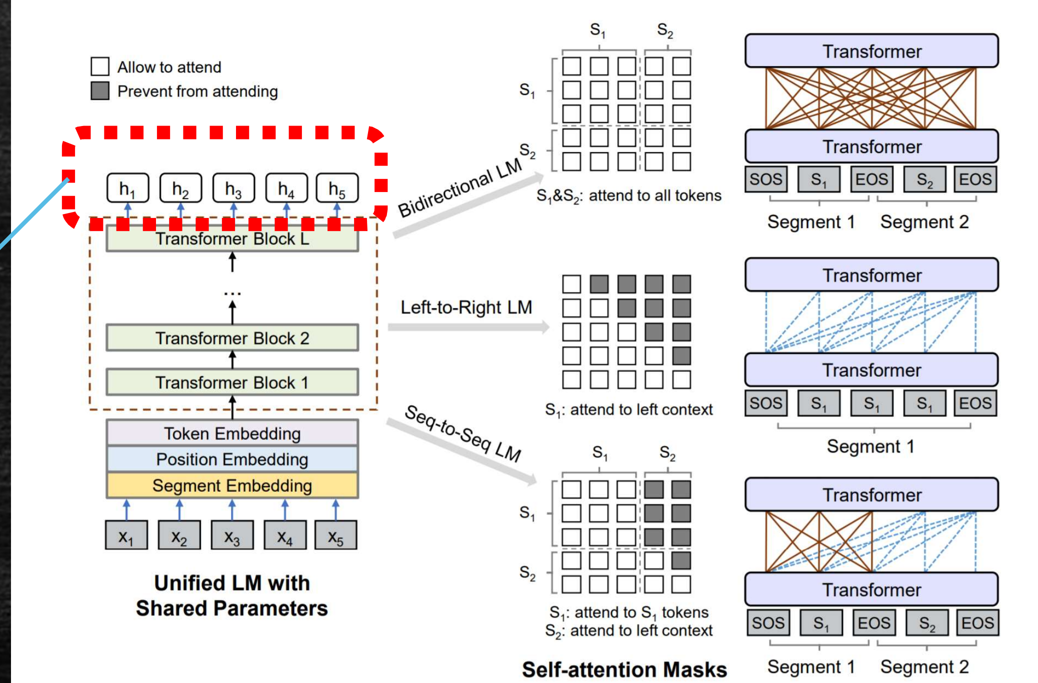


Figure 1: Overview of unified LM pre-training. The model parameters are shared across the LM objectives (i.e., bidirectional LM, unidirectional LM, and sequence-to-sequence LM). We use different self-attention masks to control the access to context for each word token. The right-to-left LM is similar to the left-to-right one, which is omitted in the figure for brevity.

Modeling:
Conditional Cycle-ULM

- Relationship between classification and conditional Seq2Seq
- Integrate generative model with classification model

$$p(l|x,y) \propto p(y|x,l)p(l) \quad \& \quad p(l|x,y) \propto p(x|y,l')p(l')$$

take $p(l|x,y) = \left[\frac{p(y|x,l)p(l)}{\sum_l p(y|x,l)p(l)} + \frac{p(x|y,l')p(l')}{\sum_{l'} p(x|y,l')p(l')}\right]/2$

Proposed Method:
Conditional Cycle-ULM

- Classification model $F(x)$
- Unsupervised Learning by Language Model: $Q(x)$
- Pairing by Seq2Seq: $x \to y = G(x)$

Unified Language Model (ULM)

- Label → Conditional Generation: $x, l \to y = G(x, l)$

Conditional ULM

- Reliability → Cycle Consistency:

$$y = G(x, \text{contradict}) \Leftrightarrow x = G(y, \text{contradict})$$
$$y = G(x, \text{neutral}) \Leftrightarrow x = G(y, \text{neutral})$$

Cycle Consistent Loss

Entailment is not symmetric! Two possible solutions

Chain rule: $y = G(x, \text{entail}), z = G(y, \text{entail}) \Rightarrow z = G(x, \text{entail})$

or New label: $y = G(x, \text{entail}) \Leftrightarrow x = G(y, \text{entailed by})$

Training Objective of Conditional Cycle-ULM

- Part 1: Conditional Seq2Seq LM on Labeled Data:

$$\min_G L_{csl}(D) = -\sum_{(x,y,l)\in D} \log p(y|x,l;G) + \log p(x|y,l';G)$$

- Part 2: Supervised Learning on Labeled Data:

$$\min_G L_{sll}(D) = -\sum_{(x,y,l)\in D} \log p(l|y,x;G)$$

- Remark:
  - The right part of Part1 is from cycle consistency, with the dual label l'

Training Objective of Conditional Cycle-ULM

- Part 3: Cycle Consistency on Unlabeled Data:

$$\min_G L_{ccu}(D') = -\sum_{x \in D'} \log p(x \mid \hat{y}, l'; G)$$

- Part 4: Supervised Learning on Unlabeled Data:

$$\min_G L_{slu}(D') = -\sum_{x \in D'} \log p(l \mid x, \hat{y}; G)$$

- Here

$$\hat{y} = \arg\max p(y \mid x, l)$$

$$l \sim \text{Uniform}(\{\text{contradiction, neutral, entailment, entailed by}\})$$

Training Objective of Conditional Cycle-ULM

- Final Objective:

$$L = \lambda_{csl} L_{csl}(D) + \lambda_{sll} L_{sll}(D)$$
$$+\lambda_{ccu} L_{ccu}(D') + \lambda_{slu} L_{slu}(D')$$
$$+\lambda_{ULM} L_{ULM}(\{D, D'\})$$

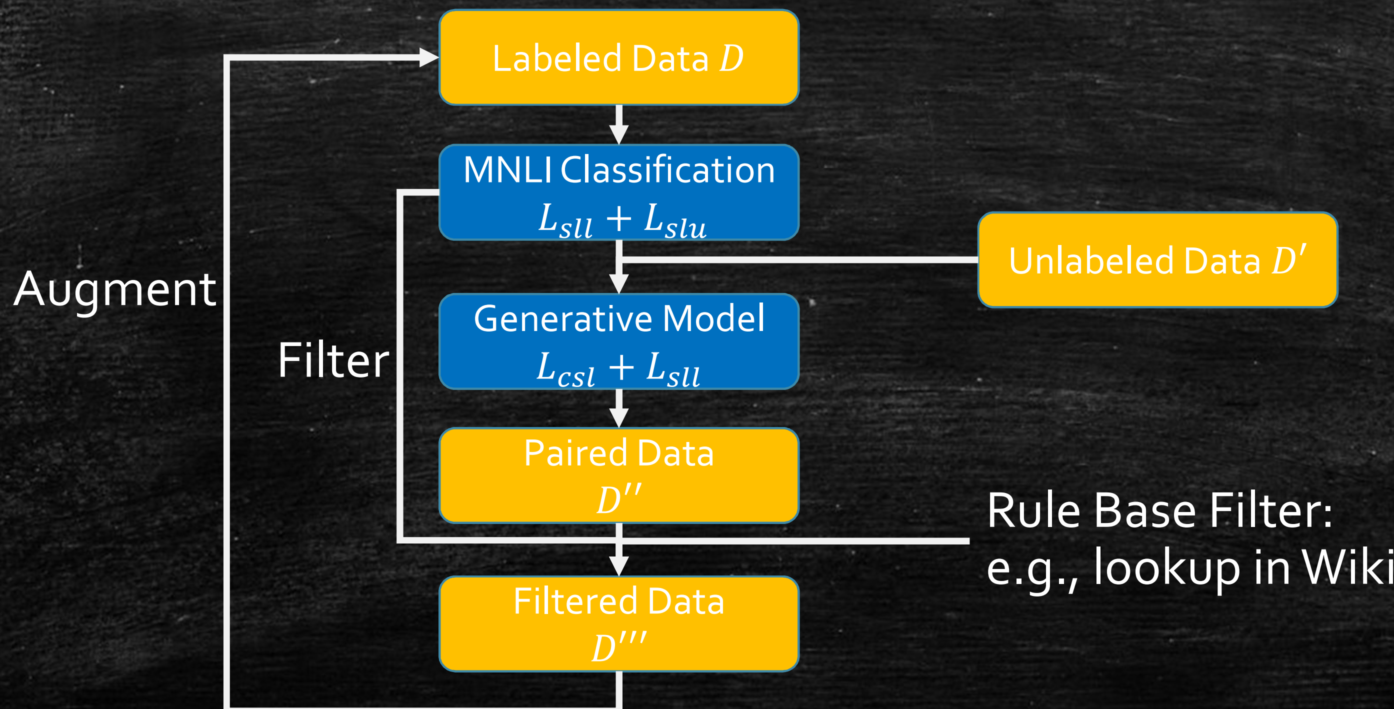
- $\lambda_{csl}, \lambda_{sll}, \lambda_{ccu}, \lambda_{slu}, \lambda_{ULM}$ are trade off parameters
- L_{ULM} is from the unsupervised pretraining *(Dong et al. 2019)*
  - Unidirectional LM
  - Bidirectional LM
  - Seq2Seq LM

Training Scheme

- Scheme 1: joint training of all objectives in a multitask way
  - Adjust trade off parameters $\lambda_{csl}, \lambda_{sll}, \lambda_{ccu}, \lambda_{slu}, \lambda_{ULM}$ during the training. Gradually increasing the noisy signal from unlabeled data to prevent overfitting the noise.

Training Scheme

- Scheme 2: Multi-Stage Training



Other Challenges

- Diversity
  - Sample suboptimal: $\hat{y} \simeq \arg\max p(y|x, l)$
  - Inject random noise: $\hat{y} = \arg\max p(y|x + \delta, l)$
- Domain Mismatch
  - Hopefully, the large scale out-domain unlabeled data can help with that