COSY: COunterfactual SYntax for Cross-Lingual Understanding

Sicheng Yu[1], Hao Zhang[2,3], Yulei Niu[2], Qianru Sun[1], Jing Jiang[1]

[1]Singapore Management University, Singapore
[2]Nanyang Technological University, Singapore
[3] Agency for Science, Technology and Research, Singapore

ACL 2021

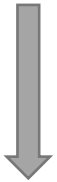
Cross-lingual Understanding

Taking natural language inference under zero-shot setting as an example:

Training

Premise: *You don't have to stay there.*
Hypothesis: *You can leave.*

English



Testing

Premise: *让我告诉你，美国人最终如何看待你作为独立顾问的表现。*
Hypothesis: *美国人完全不知道您是独立律师。*

Chinese

Premise: *Y se estremeció con el recuerdo.*
Hypothesis: *El pensamiento sobre el acontecimiento hizo su estremecimiento.*

Spanish

The examples from XNLI.

Cross-lingual Understanding

Large performance gap between English and target languages!

	En		Non-En		
XNLI	80.8	−	64.3	=	16.5
MLQA	80.2	−	58.3	=	21.9
XQUAD	83.5	−	62.6	=	20.9

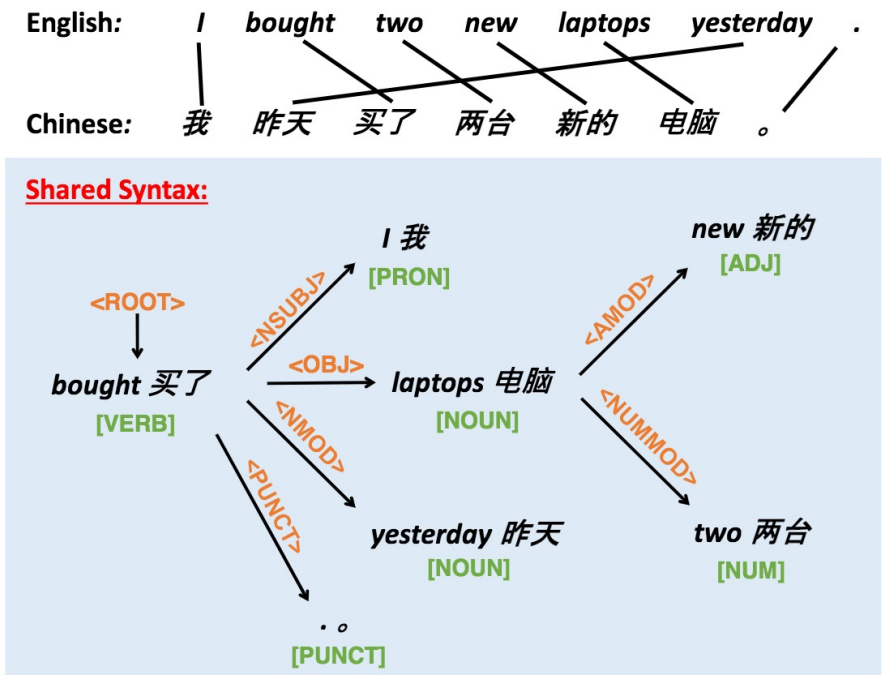
A straightforward solution:

language-agnostic feature ——— most transferable feature across languages

Our Solution: COSY

Key idea:

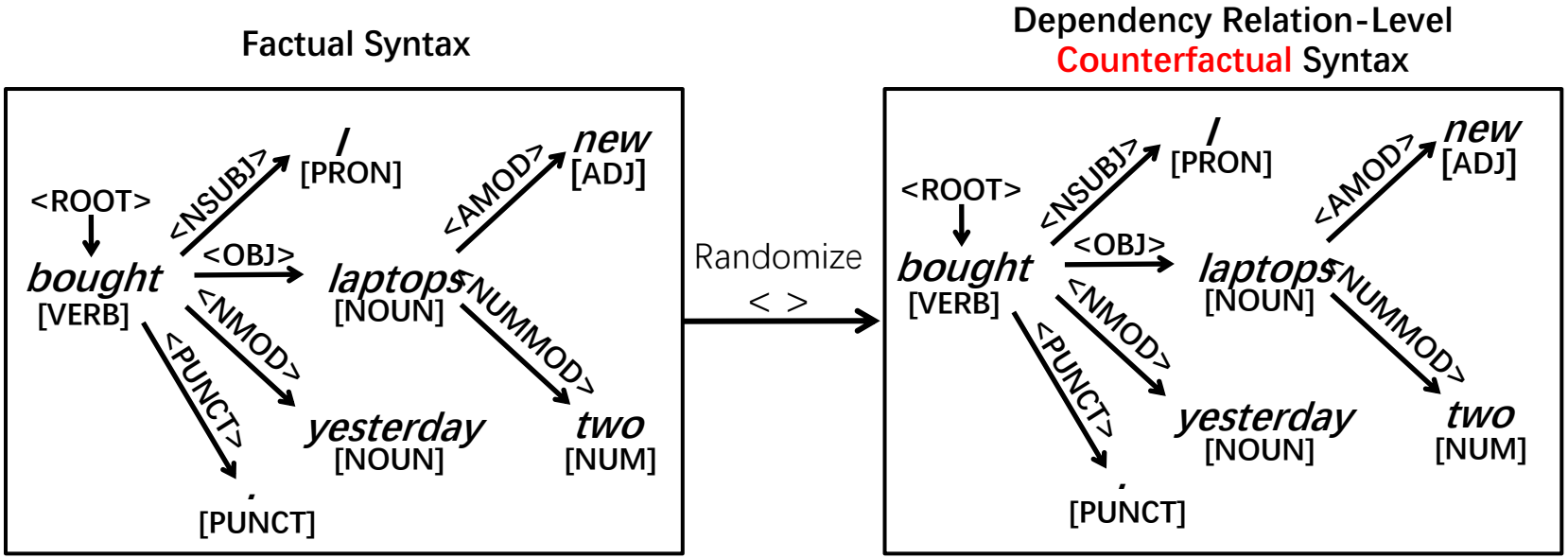
1. Adopt ***universal syntax*** as **language-agnostic feature**



Our Solution: COSY

Key idea:

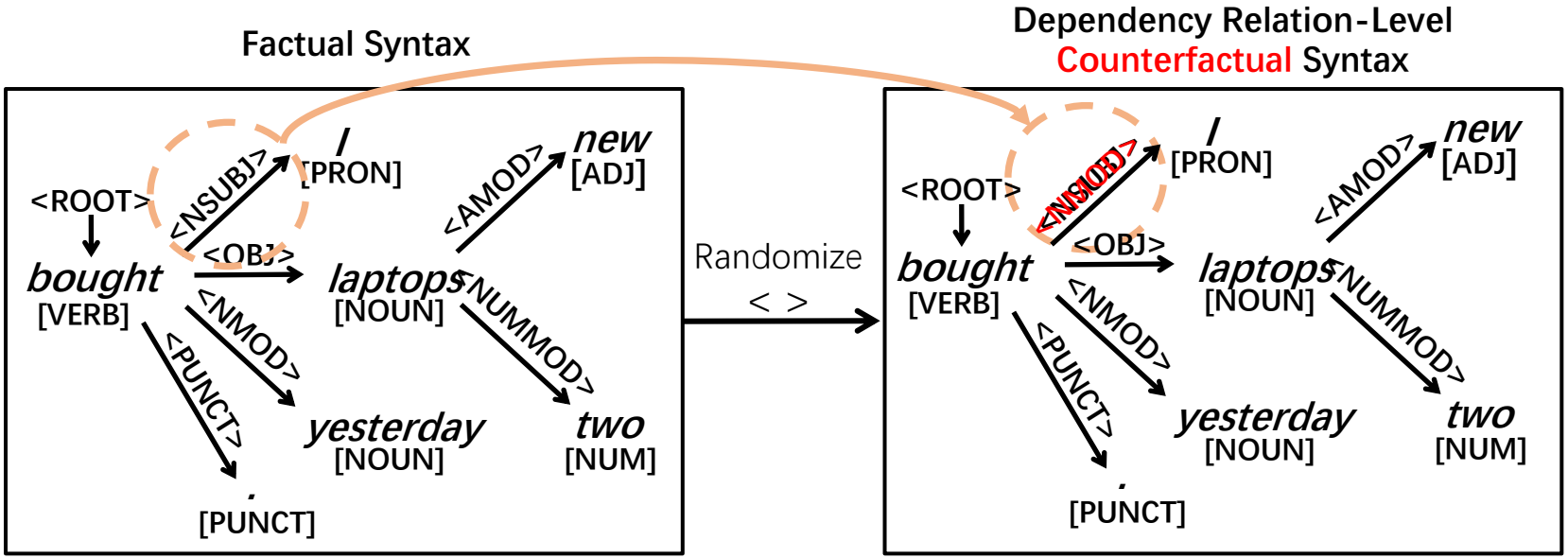
1. Adopt ***universal syntax*** as **language-agnostic feature**
2. Create ***counterfactual syntax*** to guide model to ***focus on*** the syntactic feature



Our Solution: COSY

Key idea:

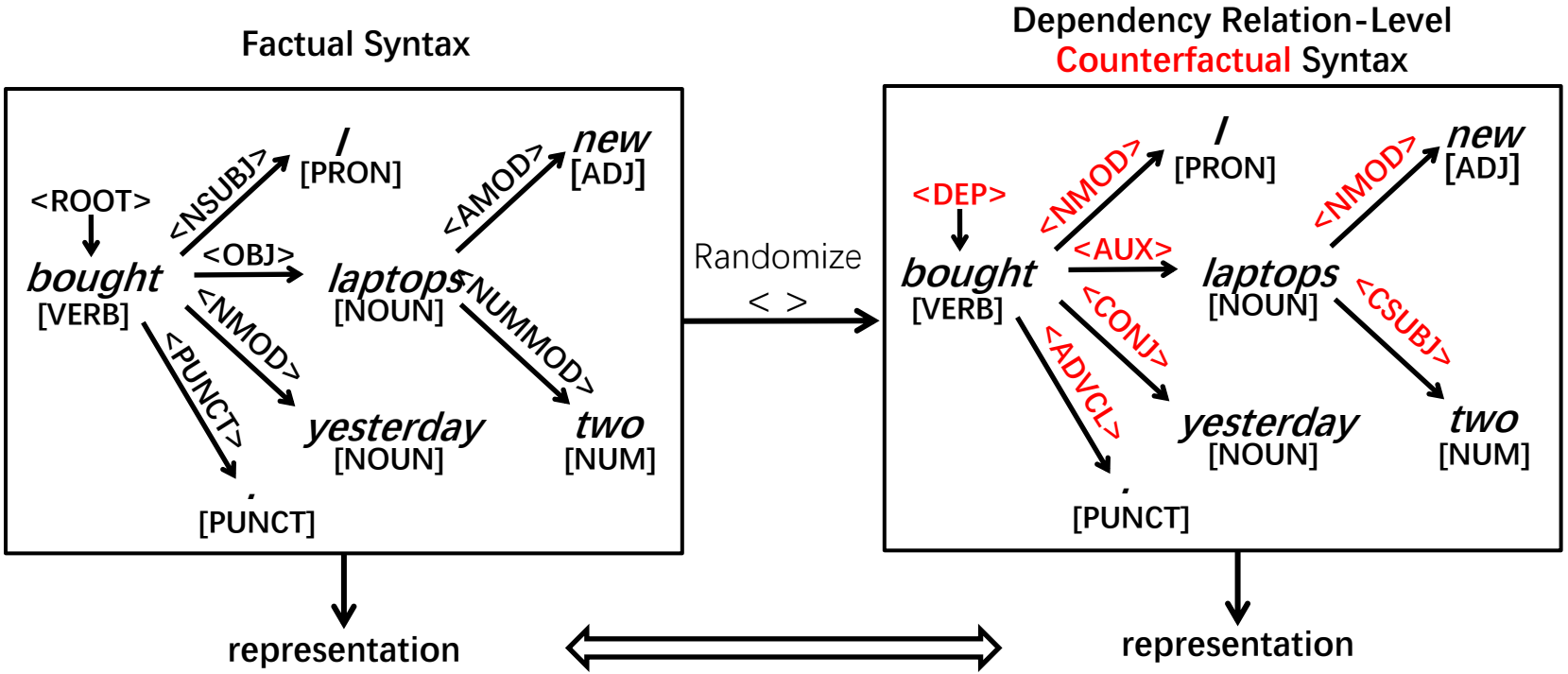
1. Adopt ***universal syntax*** as **language-agnostic feature**
2. Create ***counterfactual syntax*** to guide model to ***focus on*** the syntactic feature



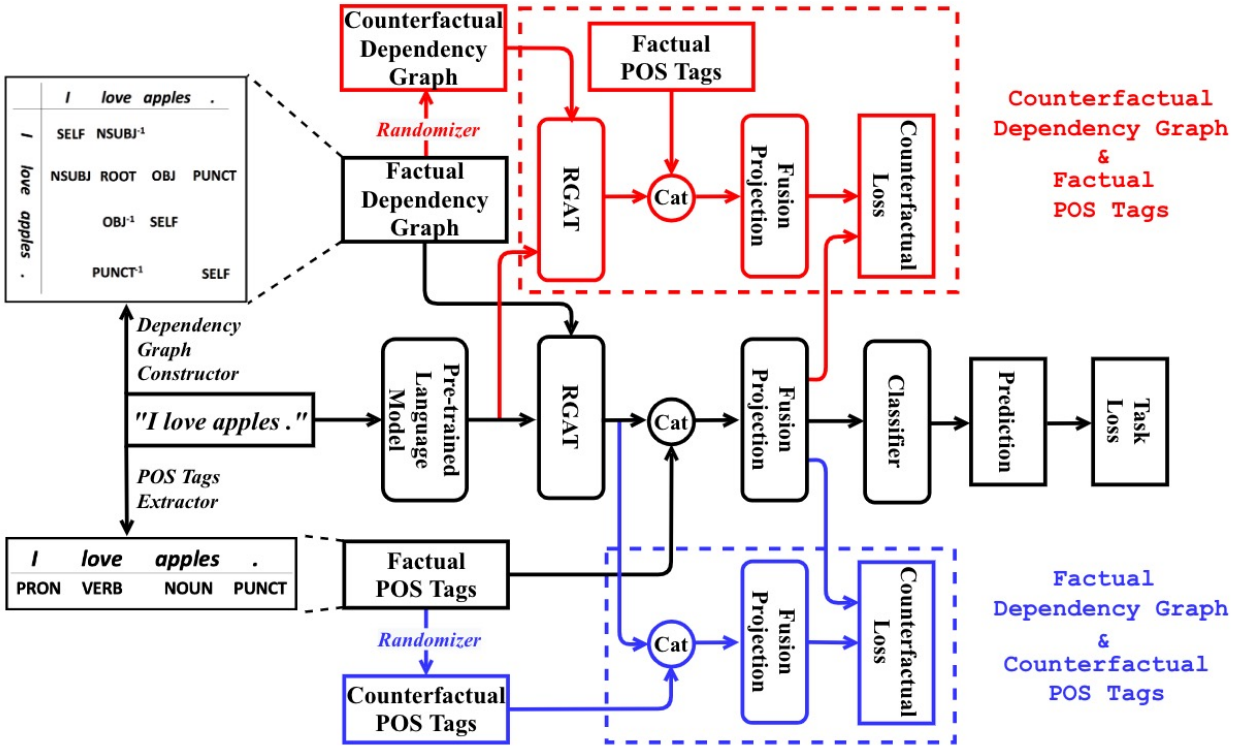
Our Solution: COSY

Key idea:

1. Adopt ***universal syntax*** as **language-agnostic feature**
2. Create ***counterfactual syntax*** to guide model to ***focus on*** the syntactic feature



COunterfactual SYntax (COSY)

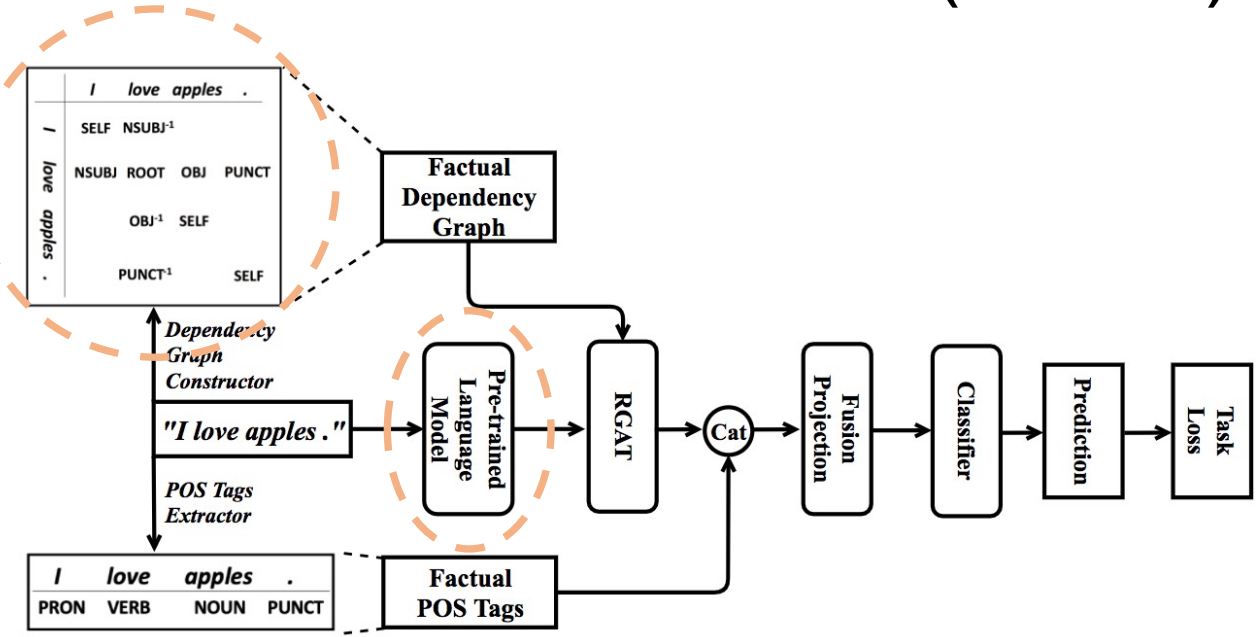


SAN-Red for counter*factual dependency*

SAN-Black for *factual syntax*

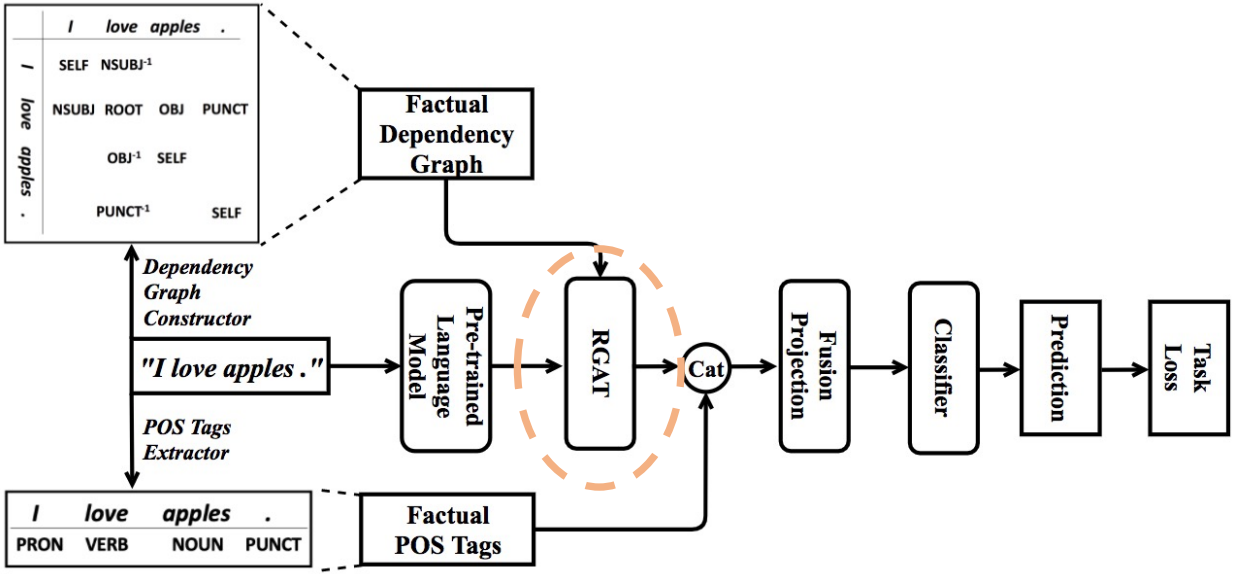
SAN-Blue for counter*factual POS Tags*

COunterfactual SYntax (COSY) :SAN-Black



- Pre-trained language model: extract contextual representation, $\boldsymbol{H} = \{\boldsymbol{h}_i\}_{i=1}^{S} \in \mathbb{R}^{S \times d}$, *e.g.*, mBERT, XLM-R.
- Construct ***factual dependency graph***: G.
  - Forward relation
  - Backward relation
  - Self-loop

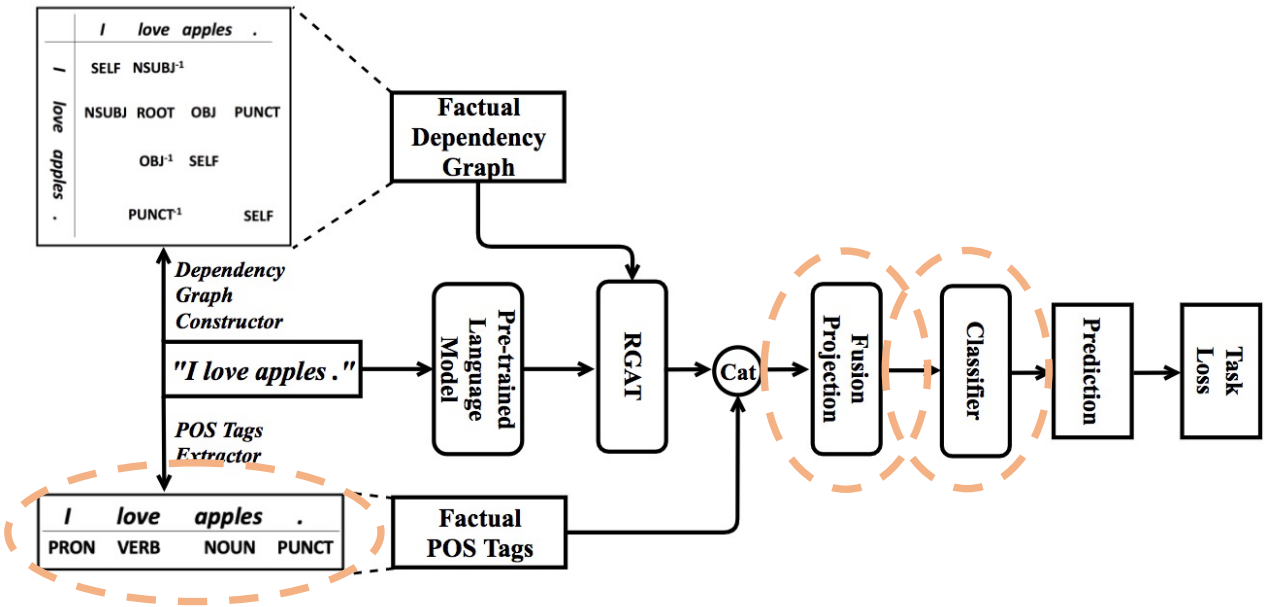
COunterfactual SYntax (COSY) :SAN-Black



- Relational graph attention networks (RGAT): incorporating dependency graph to obtain relation aware representation, $\boldsymbol{H}' = \{\boldsymbol{h}_i\}_{i=1}^{S} \in \mathbb{R}^{S \times d}$:

$$\boldsymbol{H}' = \mathrm{RGAT}(\boldsymbol{H}, G)$$

COunterfactual SYntax (COSY) :SAN-Black

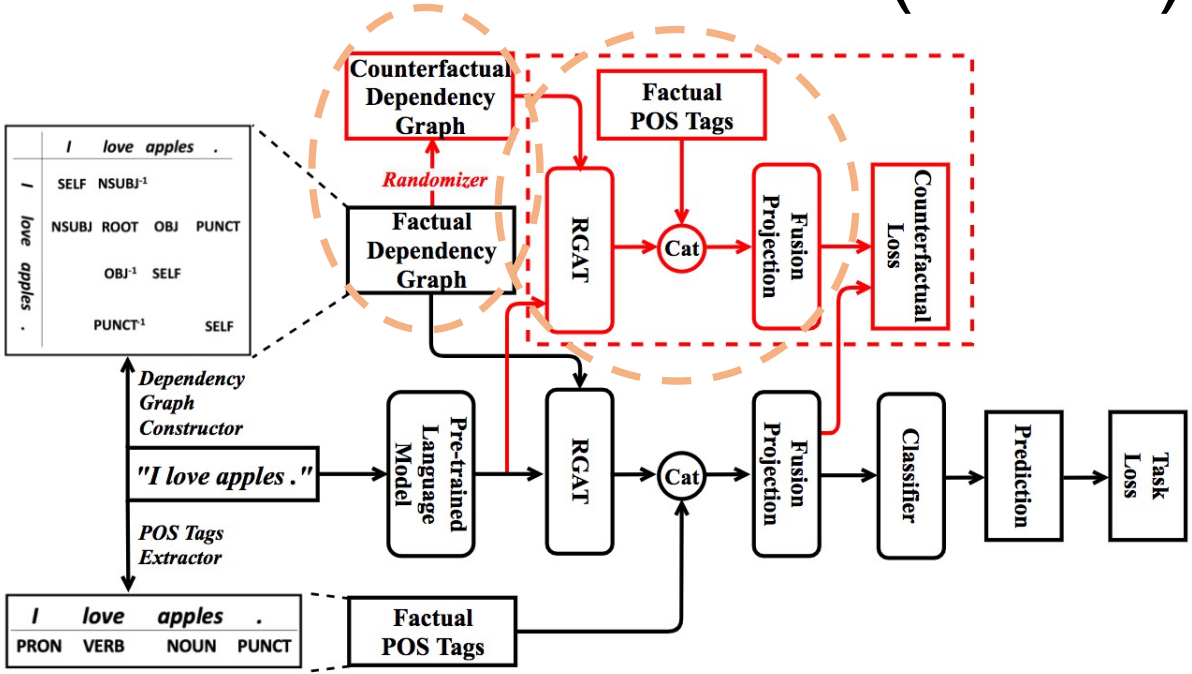


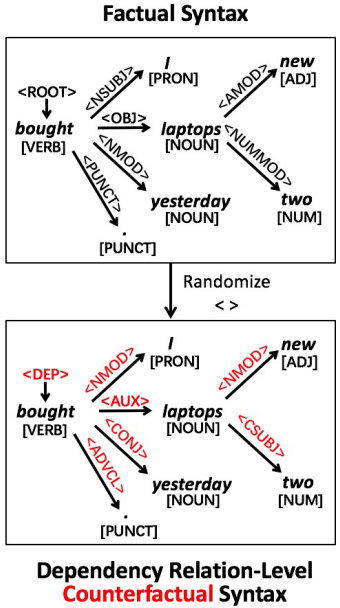
- Extract ***factual POS Tags***: P.
- Fuse relation aware representation with POS tags by fusion projection, the factual fused feature can be obtained by:

$$\boldsymbol{F} = \mathrm{Concat}(\boldsymbol{H}', P) \cdot \boldsymbol{W}_F$$

- Task-specific classifier and task-specific loss, *i.e.,* cross-entropy loss.

COunterfactual SYntax (COSY) :SAN-Red

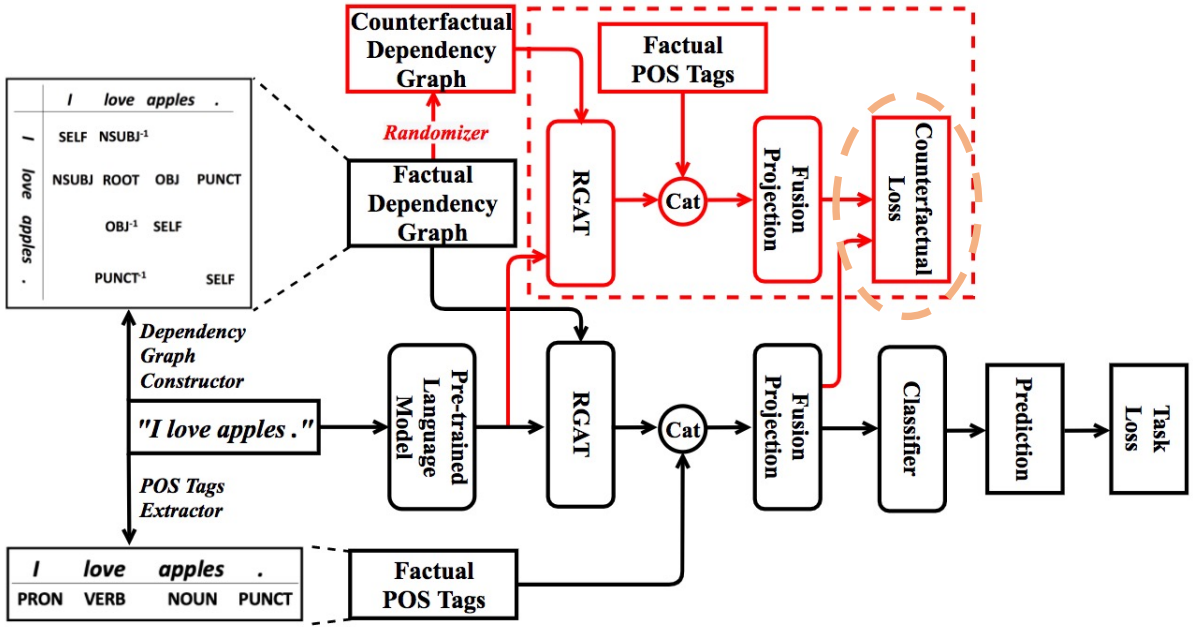




- Randomize ***factual dependency graph*** to obtain ***counterfactual dependency graph*** G^{-}.

- Feed RGAT and fusion projection with ***counterfactual dependency graph*** and ***factual POS tags*** to obtain first counterfactual fused feature $\boldsymbol{F^{cf1}}$.

COunterfactual SYntax (COSY) :SAN-Red

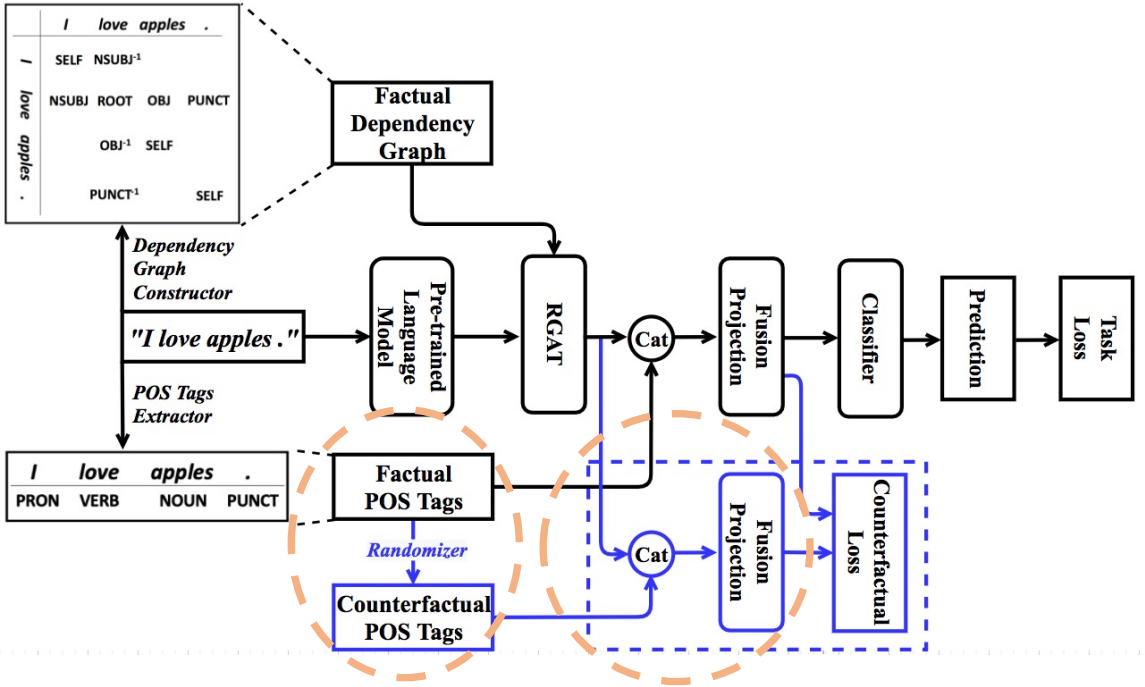


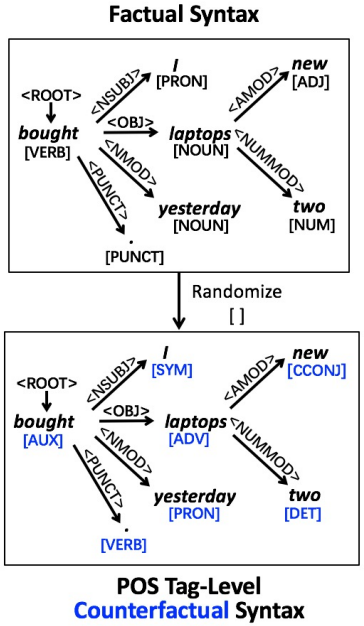
Target: guide model to focus on dependency graph.

Feature from ***factual dependency graph*** and ***counterfactual dependency graph*** should be different.

$$Loss_{cf1} = \sum f \cdot f^{cf1}$$

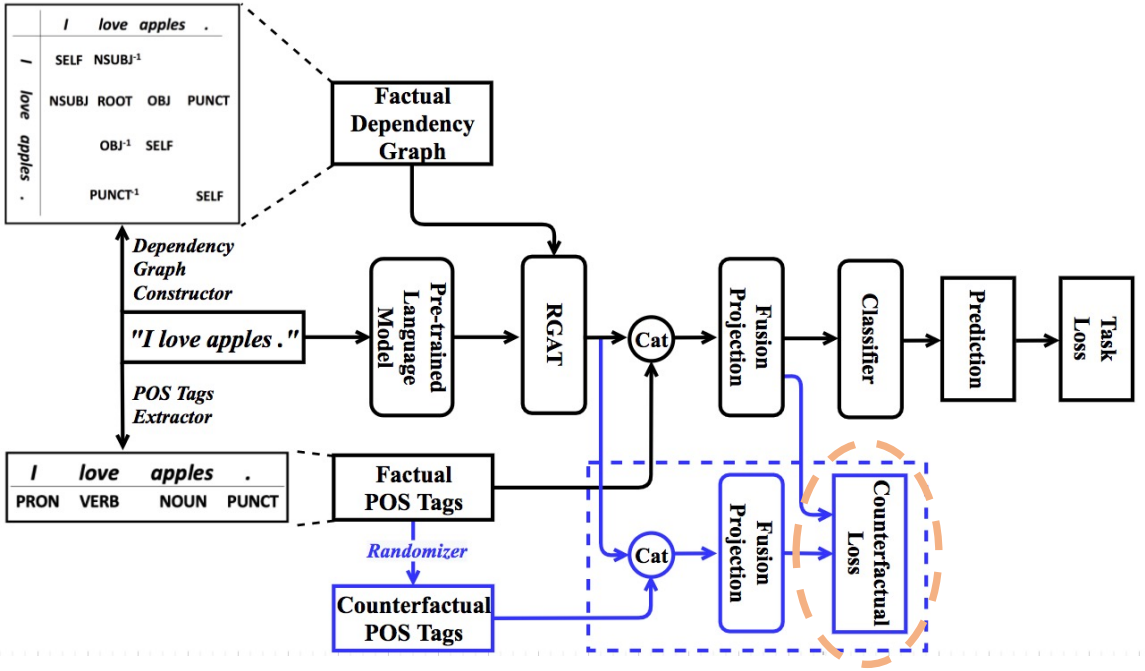
COunterfactual SYntax (COSY) :SAN-Blue





- Randomize ***factual POS tags*** to obtain ***counterfactual POS tags***.
- Feed fusion projection with ***counterfactual POS tags*** to obtain second counterfactual fused feature F^{cf2}.

COunterfactual SYntax (COSY) :SAN-Blue

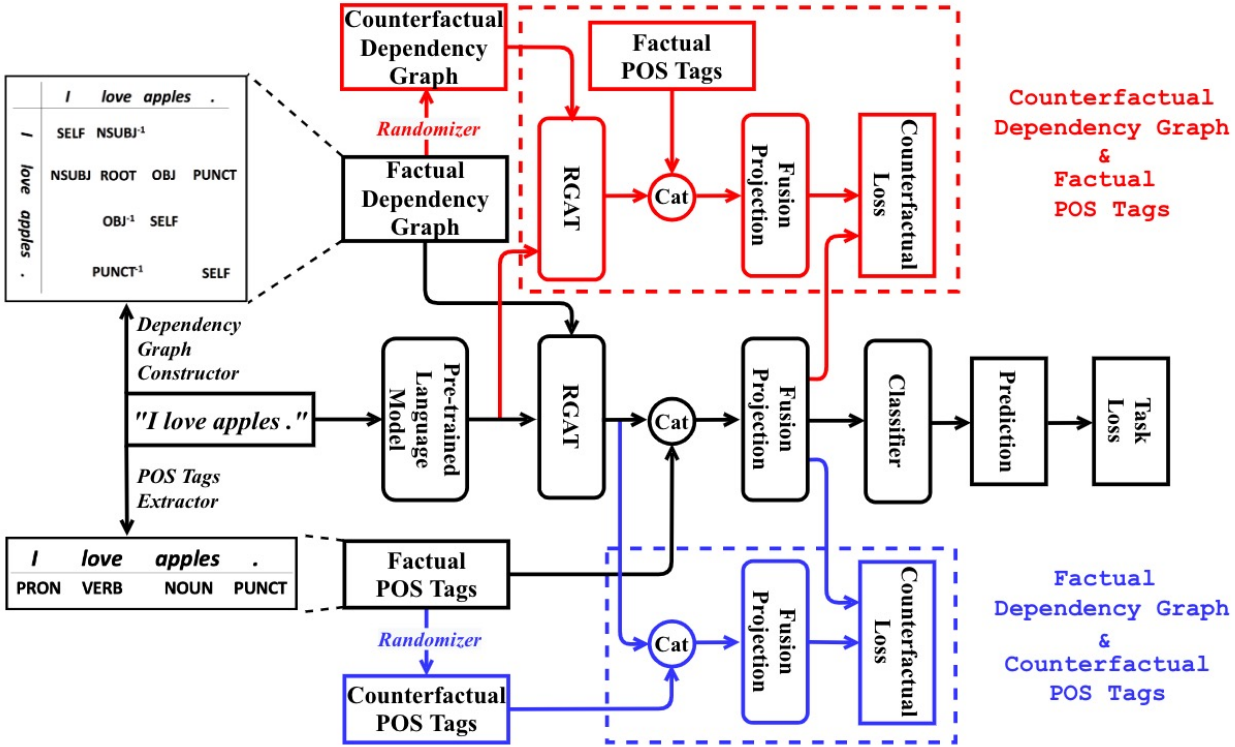


Target: guide model to focus on POS tags.

Feature from ***factual POS tags*** and ***counterfactual POS tags*** should be different.

$$Loss_{cf2} = \sum f \cdot f^{cf2}$$

COunterfactual SYntax (COSY): Overall



$$Loss = Loss_{task} + \lambda(Loss_{cf1} + Loss_{cf2})$$

Experiments

Datasets:
- XNLI (Natural Language Inference)
- MLQA (Question Answering)
- XQUAD (Question Answering)

Metrics
- Accuracy for XNLI
- EM / F1 for MLQA and XQUAD

Settings:
- Zero-shot
- Few-shot

Backbones
- mBERT
- XLM-R base and XLM-R large

Experiments: Main results

	Method	#T	#M	A.D.	XNLI		MLQA		XQUAD	
					en.	avg.	en.	avg.	en.	avg.
mBERT	Naive F.T.	1	1	✘	82.1	68.4	67.0 / 80.2	44.2 / 61.4	72.2 / 83.5	51.0 / 66.7
	XMAML-One	L	$O(L)$	✔	82.1	69.6	-	-	-	-
	LAKM	1	1	✔	-	-	66.8 / 80.0	-	-	-
	COSY (Ours)	1	1	✘	82.2	**70.1**	67.2 / 80.4	**45.2 / 62.1**	72.6 / 83.6	**53.2 / 68.1**
X-R_{base}	Naive F.T.	1	1	✘	84.6	75.1	- / 80.1	- / 65.1	71.6 / 83.1	55.9 / 71.8
	XMAML-One	L	$O(L)$	✔	-	-	- / 80.2	- / 66.1	-	-
	COSY (Ours)	1	1	✘	84.3	**75.6**	67.7 / 80.7	**48.5 / 66.5**	74.0 / 85.1	**57.3 / 73.4**
X-R_{large}	Naive F.T.	1	1	✘	88.7	80.0	70.6 / 83.5	53.2 / 71.6	75.7 / 86.5	60.6 / 76.8
	STILT	9	1	✔	89.6	81.6	70.8 / 84.1	54.4 / 72.8	77.4 / 88.3	63.3 / 78.7
	XMAML-One	L	$O(L)$	✔	-	-	- / 84.3	- / **73.2**	-	-
	COSY (Ours)	1	1	✘	89.2	**81.9**	70.9 / 84.2	**54.7 / 73.2**	77.7 / 88.0	**64.0 / 79.7**

Zero-shot setting

Method	en.	non-en. avg.	avg.
Naive F.T.*	81.9	70.3	71.2
XMAML-One*	82.4	70.7	71.6
COSY (Ours)	82.6	**71.9**	**72.7**

Few-shot setting (XNLI)

Baselines

- Naïve F.T.: Naïve Fine-tuning
- XMAML-One : Meta-learning
- LAKM: External data from web
- STILT: Augments intermediate task

COSY surpasses Naïve F.T. significantly.

COSY is comparable to or better than other compared methods.

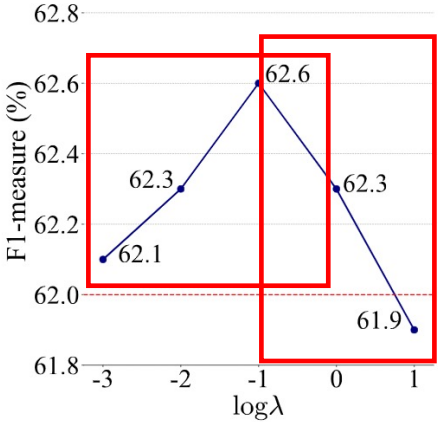
Experiments: Analysis

Ablative Setting	MLQA		XQUAD		XNLI
	EM	F1	EM	F1	Acc
Naive F.T.	44.2	61.4	51.0	66.7	68.4
(1) SAN-Black	44.3	61.4	51.6	66.9	68.7
(2) SAN-Black+Gate	44.5	61.5	51.9	67.1	68.7
(3) SAN-Black, Red	44.9	61.7	52.8	67.8	69.9
(4) SAN-Black, Blue	44.7	61.8	52.2	67.4	69.7
(5) COSY	**45.2**	**62.1**	**53.2**	**68.1**	**70.1**

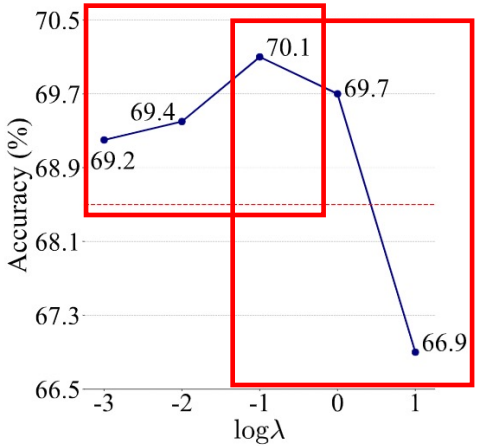
Ablative settings

- Syntax feature is helpful
- Dependency graph > POS tags
  Red Blue
- (Dependency graph + POS tags) < Dependency graph + POS tags
  (Red + Blue) Red + Blue

Experiments: Analysis



Effect of λ on MLQA



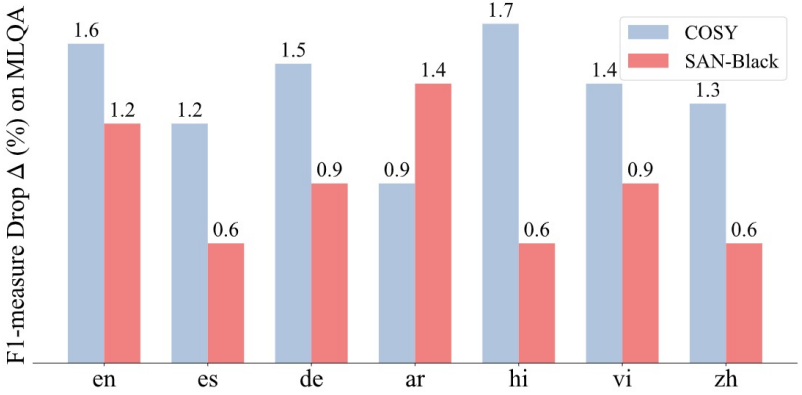
Effect of λ on XNLI

- When λ increases, model over-emphasizes the syntactic feature
- When λ decreases, model over-looks the syntactic

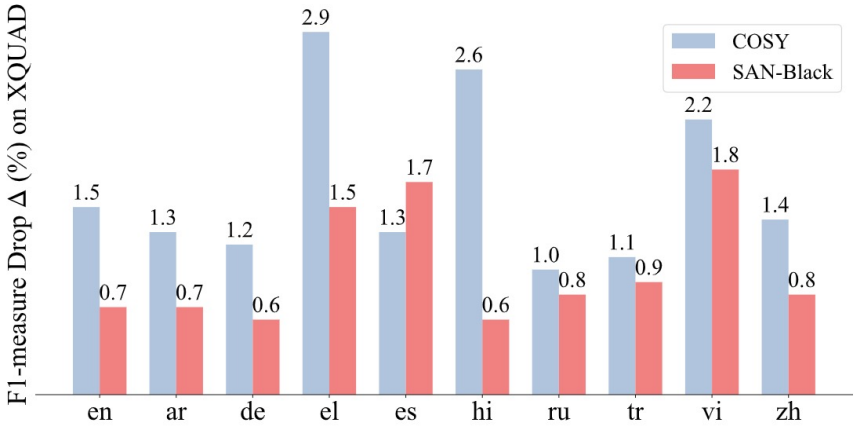
Experiments: Analysis

Does COSY really make the model concentrate on syntactic information?

- Comparing COSY and SAN-Black (syntactic information is simply encoded)
- The model more sensitive to noise on syntactic feature pays more attention to syntax



Performance drop on MLQA



Performance drop on XQUAD

Conclusion

- We design a syntax-aware network that incorporates transferable syntax in language models.

- We propose a novel counterfactual training method that addresses the technical challenge of emphasizing syntax.

- Experiments show that COSY achieves SOTA performance on several cross-lingual understanding tasks under different backbones and settings.

Thank You!