Evaluation @ EMNLP 2020 (Part 2)

Presenter: Wang Chen

Outline

- Evaluating the Factual Consistency of Abstractive Text Summarization
- GRADE- Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems
- UNION-An Unreferenced Metric for Evaluating Open-ended Story Generation

Evaluating the Factual Consistency of Abstractive Text Summarization

Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher
Salesforce Research
{kryscinski,bmccann,cxiong,rsocher}@salesforce.com

Presenter: CHEN Wang

FactCC — Introduction

Source article fragments	
(CNN) The mother of a quadriplegic man who police say was left in the woods for days cannot be extradited to face charges in Philadelphia until she completes an unspecified "treatment," Maryland police said Monday. The Montgomery County (Maryland) Department of Police took Nyia Parler, 41, into custody Sunday (...)	(CNN) The classic video game "Space Invaders" was developed in Japan back in the late 1970's – and now their real-life counterparts are the topic of an earnest political discussion in Japan's corridors of power. Luckily, Japanese can sleep soundly in their beds tonight as the government's top military official earnestly revealed that (...)
Model generated claims/sentences	
Quadriplegic man Nyia Parler, 41, left in woods for days can not be extradited.	Video game "Space Invaders" was developed in Japan back in 1970.

Table 1: Examples of factually incorrect claims output by summarization models. Green text highlights the support in the source documents for the generated claims, red text highlights the errors made by summarization models.

FactCC — Related Work

- Natural language inference (NLI)
  - focuses on classifying logical entailment between **short, single sentence pairs**
  - but **verifying factual consistency** can require incorporating the entire context of the source document
- Fact checking
  - focuses on verifying facts against **the whole of available knowledge**
  - whereas **factual consistency checking** focuses on adherence of facts to information provided by a source document without guarantee that the information is true

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)
- **Building the development and test datasets**
- **Training** a BERT-based binary classifier

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)

Consistent Pairs

Inconsistent Pairs

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)

Consistent Pairs

Inconsistent Pairs

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

Back Translation
En->Ch->En
En->Fr->En
En->Ge->En
En->Sp->En
En->Ru->En

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

Consistent Pairs

Inconsistent Pairs

Randomly choose a verb:
Negation->Non-negation
Non-negation->Negation

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)

	Transformation	Original sentence	Transformed sentence
Consistent Pairs	Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Inconsistent Pairs	Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
	Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
	Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
	Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
	Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

A **randomly** chosen pronoun was **swapped** with a different one from the **same pronoun group** to ensure syntactic correctness, e.g., {him,her,them}

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)

Consistent Pairs

Inconsistent Pairs

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

An entity is randomly replaced with an entity within the same group,
Name entity->Name entity
Dates->Number values

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

Consistent Pairs

Inconsistent Pairs

the token was randomly duplicated or removed from the sequence

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)

Require:
S - set of source documents
$\mathcal{T}^+$ - set of semantically invariant transformations
$\mathcal{T}^-$ - set of semantically variant transformations

function GENERATE_DATA($\mathcal{S}, \mathcal{T}^+, \mathcal{T}^-$)
  $\mathcal{D} \leftarrow \emptyset$ ▷ set of generated data points
  for doc **in** $\mathcal{S}$ **do**
    $doc\_sents \leftarrow$ `sentence_tokenizer`(doc)
    $sent \leftarrow$ `choose_random`(doc_sents)
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, sent, +)\}$
    for `fn` **in** $\mathcal{T}^+$ **do**
      $new\_sent \leftarrow$ `fn`$(doc, sent)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, new\_sent, +)\}$
    end for
  end for

  for $example$ **in** $\mathcal{D}$ **do**
    $doc, sent, \_ \leftarrow example$
    for `fn` **in** $\mathcal{T}^-$ **do**
      $new\_sent \leftarrow$ `fn`$(doc, sent)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(doc, new\_sent, -)\}$
    end for
  end for
  return $\mathcal{D}$
end function

FactCC — Proposed Method

- **Building the training dataset** which contains factually consistent or inconsistent document-sentence pairs (key contribution)
  - using news articles from the **CNN/DailyMail** dataset as source documents. **1,003,355** training examples were created, out of which **50.2%** were labeled as negative (**INCONSISTENT**) and the remaining **49.8%** were labeled as positive (**CONSISTENT**).
- **Building the development and test datasets**
  - **manually annotated** by the authors on the summaries output by state-of-the-art summarization models. **Development: 931, Test: 503.**
- **Training** a BERT-based binary classifier
  - **FactCC**: BERT + Binary Classifier
  - **FactCCX**: BERT + Binary Classifier + Extract Support Spans

FactCC — Experiments

- Classification Accuracy

Model	Accuracy *(weighted)*	F1-score
BERT+MNLI	51.51	0.0882
BERT+FEVER	52.07	0.0857
FactCC (ours)	**74.15**	**0.5106**
FactCCX (ours)	72.88	0.5005

Table 3: Performance of models evaluated by means of weighted (class-balanced) accuracy and F1 score on the manually annotated test set.

FactCC — Experiments

- Oder Error Rate

Model	Incorrect	Δ
Random	50.0%	
DA (Falke et al., 2019)	42.6%	-7.4
InferSent (Falke et al., 2019)	41.3%	-8.7
SSE (Falke et al., 2019)	37.3%	-12.7
BERT (Falke et al., 2019)	35.9%	-14.1
ESIM (Falke et al., 2019)	32.4%	-17.6
FactCC (ours)	**30.0%**	**-20.0**

Table 5: Percentage of incorrectly ordered sentence pairs using different consistency prediction mnodels and crowdsourced human performance on the dataset.

P((document, positive sentence))

V ?

P((document, negative sentence))

FactCC — Experiments

- Quality of Extracted Spans by FactCCX

	Model Highlight Helpfulness			Model-Annotator Highlight Overlap	
Annotation subset	Helpful	Somewhat Helpful	Not Helpful	Accuracy	F1 score
Article Highlights					
Raw Data	79.21%	12.54%	8.25%	65.33%	0.6207
Golden Aligned	77.73%	12.66%	9.61%	74.87%	0.7161
Majority Aligned	81.11%	11.48%	7.41%	69.88%	0.6679
Claim Highlights					
Raw Data	64.44%	16.89%	18.67%	65.66%	0.6650
Golden Aligned	67.28%	16.05%	16.67%	80.54%	0.8190
Majority Aligned	67.17%	16.67%	16.16%	69.48%	0.6992

Table 6: Quality of spans highlighted in the *article* and *claim* by the FactCCX model evaluated by human annotators. The left side shows whether the highlights were considered helpful for the task of factual consistency annotations. The right side shows the overlap between model generated and human annotated highlights. Different rows show how the scores change depending on how the collected annotations are filtered. *Raw Data* shows results without filtering, *Golden Aligned* only considers annotations where the human-assigned label agreed with the author-assigned label, *Majority Aligned* only considers annotations where the human-assigned label agreed with the majority-vote label from all annotators.

FactCC — Experiments

- Quality of Extracted Spans by FactCCX

	Task without model highlights	Task with model highlights
Average work time (sec)	224.89	178.34
Inter-annotator agreement (κ)	0.1571	0.2526

Table 7: Annotation speed and inter-annotator agreement measured for factual consistency checking with and without assisting, model generated highlights.

FactCC — Conclusions

- A novel, weakly-supervised BERT-based model for verifying factual consistency in abstractive summary sentences
- Specialized modules that explain which portions of both the source document and generated summary are pertinent to the decision of the model
- Address one specific aspect of summarization evaluation

GRADE: Automatic Graph-Enhanced Coherence Metric for Evaluating Open-Domain Dialogue Systems

Lishan Huang[1*], Zheng Ye[1*], Jinghui Qin[1], Liang Lin[1,2], Xiaodan Liang[1,2†]
[1] Sun Yat-Sen University, [2] Dark Matter AI Inc.
{huanglsh6,yezh7,qinjingh}@mail2.sysu.edu.cn,
linliang@ieee.org, xdliang328@gmail.com

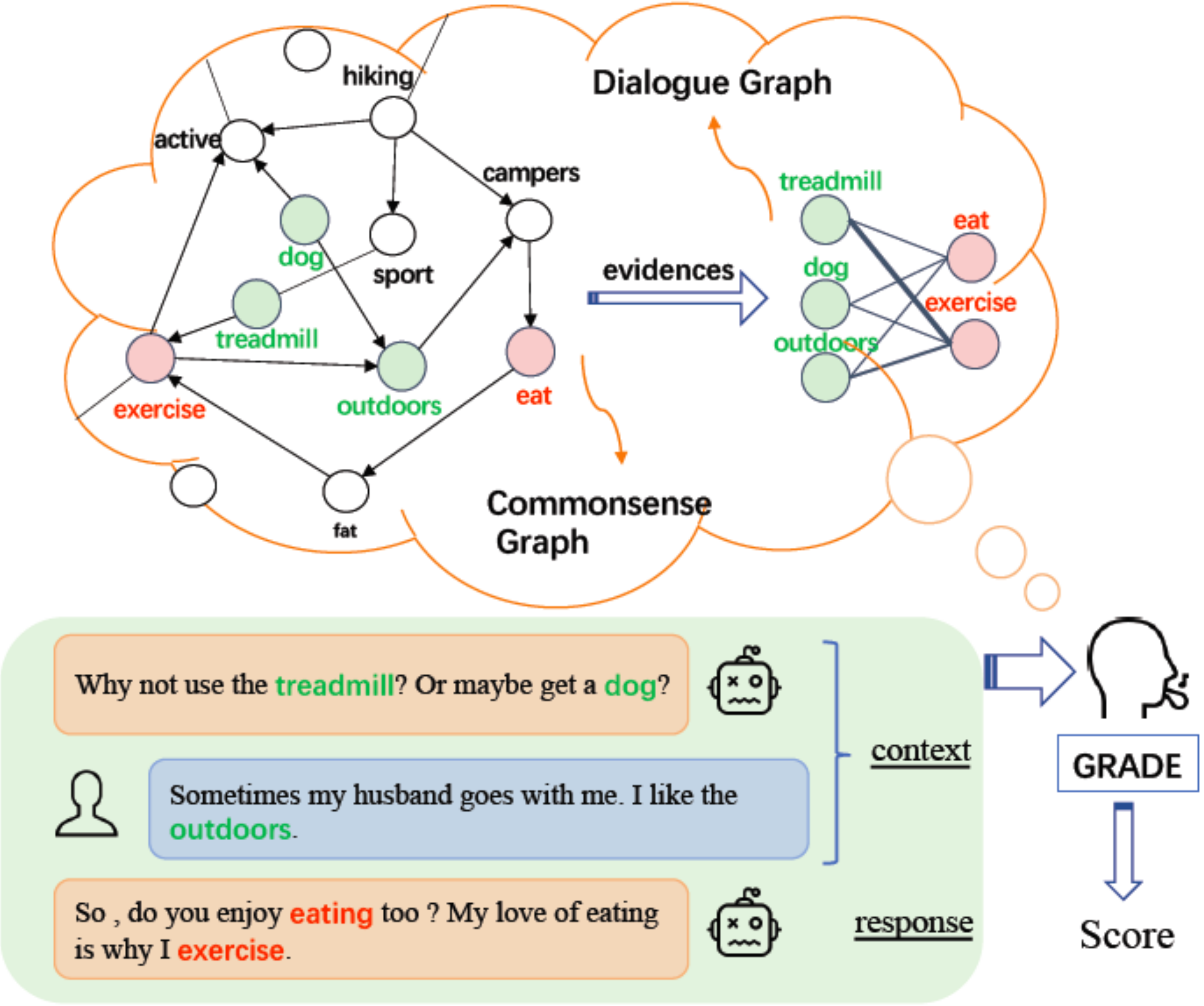
GRADE—Main Contributions

- Propose a **G**raph-enhanced **R**epresentation for **A**utomatic **D**ialogue **E**valuation (GRADE), which is the **first attempt** to introduce graph reasoning into dialogue evaluation.
- Construct and release **a new large-scale human evaluation benchmark** with 1200 context-response pairs

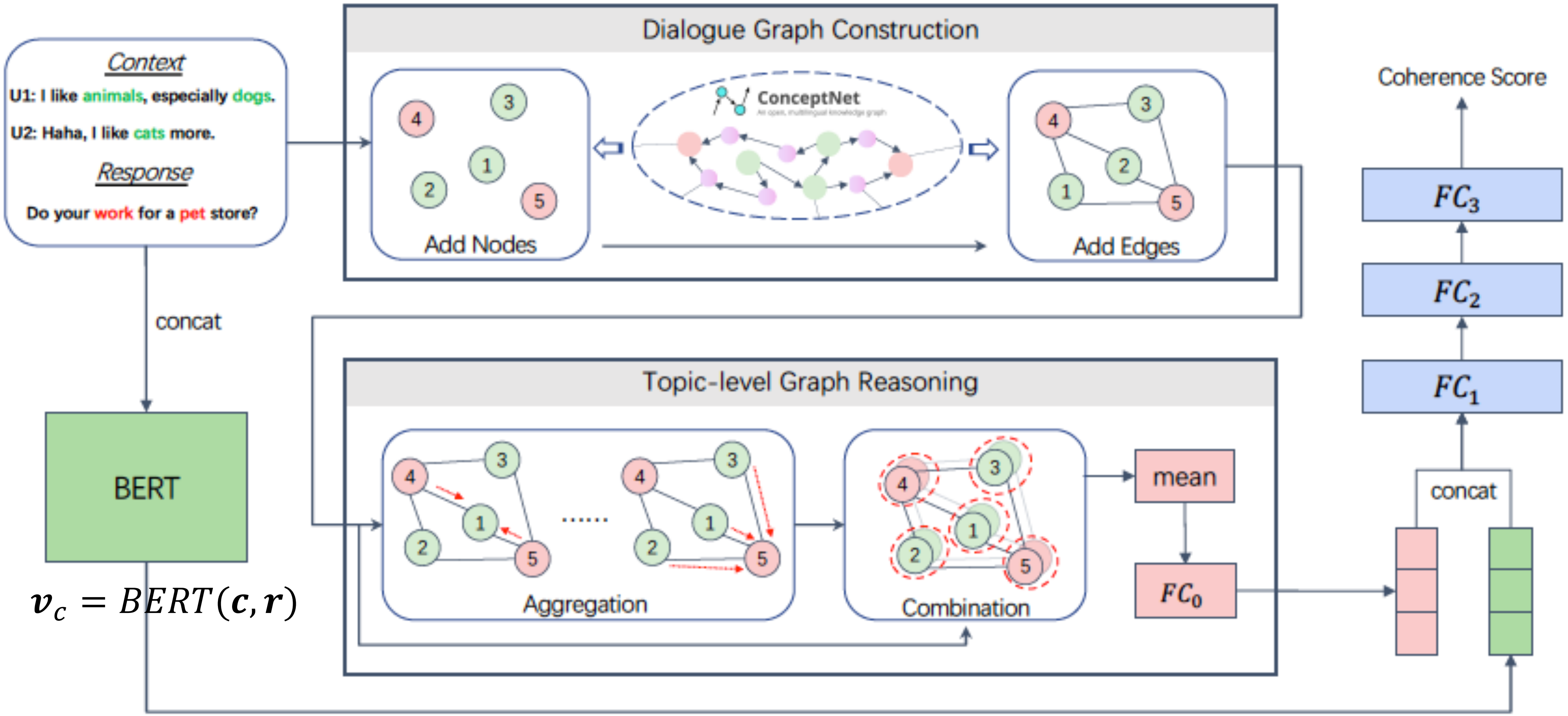
GRADE—Motivation

Green/Red words are the topic keywords of the context/response, which can be aligned to the corresponding nodes in the commonsense graph.

- Existing SOTA metrics **only** model dialogue coherence at **utterance level** without explicitly considering the fine-grained topic transition dynamics of dialogue flows

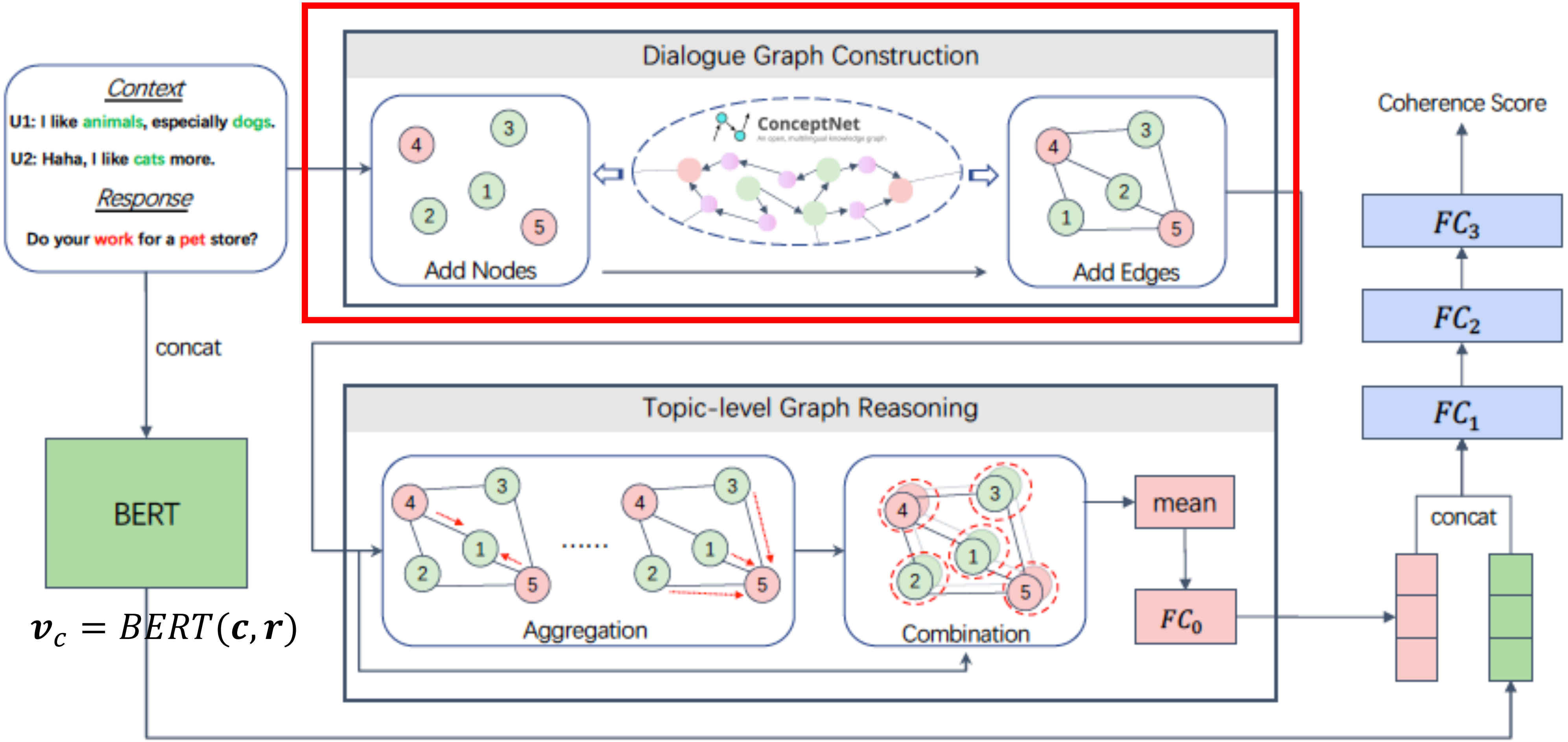


GRADE—Model



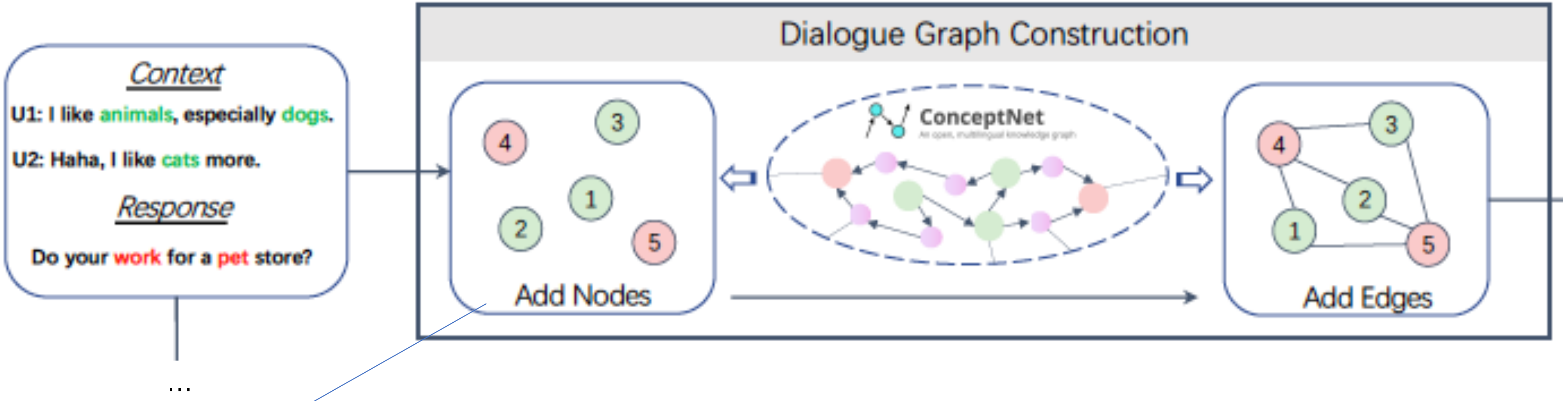
Context
U1: I like animals, especially dogs.
U2: Haha, I like cats more.
Response
Do your work for a pet store?
concat
BERT
$v_c = BERT(c, r)$
Dialogue Graph Construction
ConceptNet
Add Nodes
Add Edges
Topic-level Graph Reasoning
Aggregation
Combination
mean
FC_0
concat
FC_1
FC_2
FC_3
Coherence Score

GRADE—Model



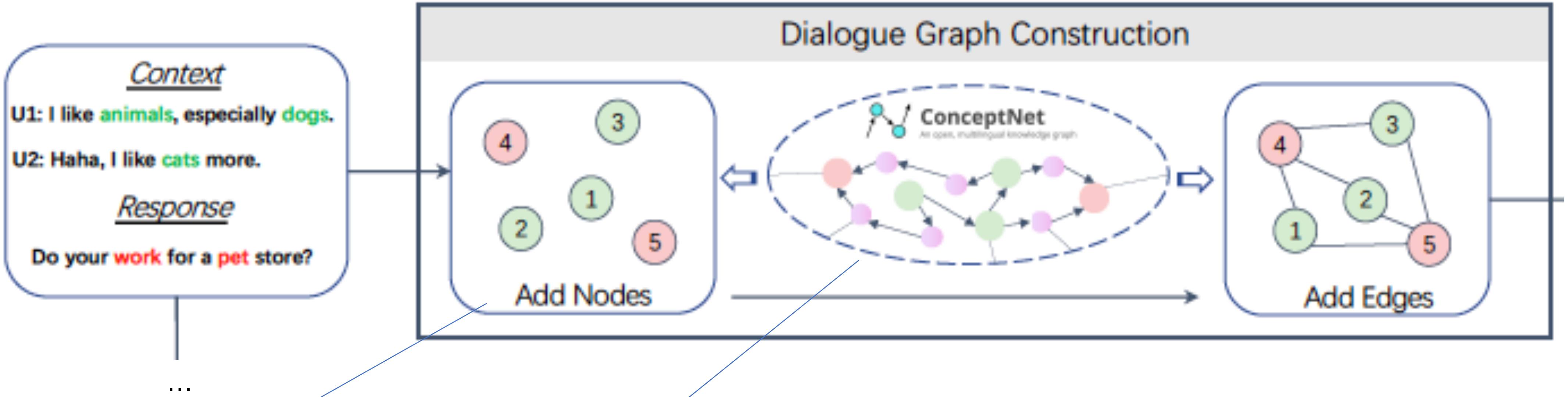
$v_c = BERT(\boldsymbol{c}, \boldsymbol{r})$

GRADE—Dialogue Graph Construction



Rule-based extractor,
tf-idf+POS

GRADE—Dialogue Graph Construction



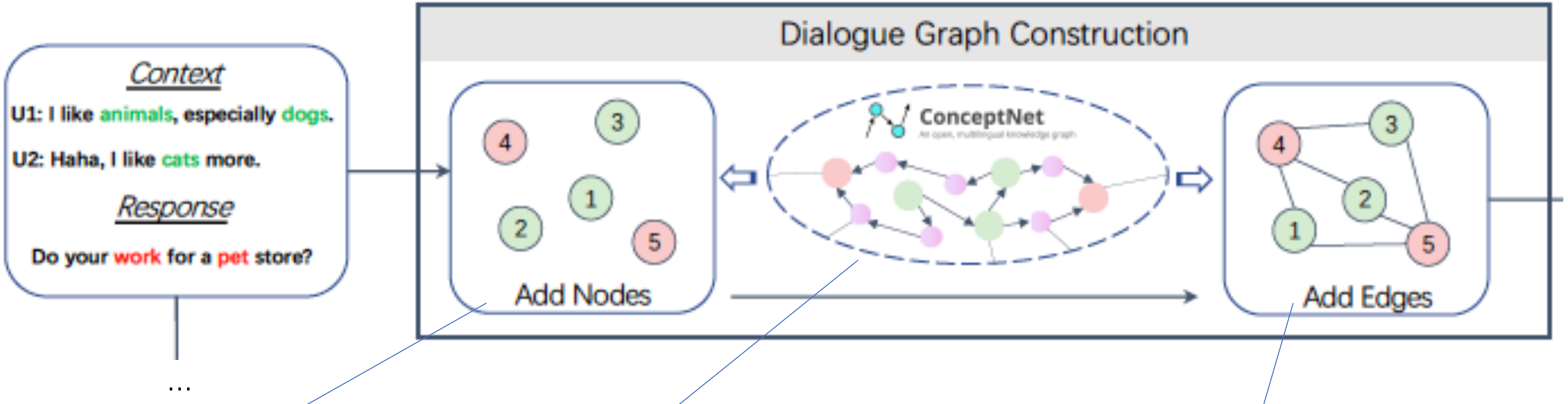
Rule-based extractor, tf-idf+POS

ConceptNet, initial: $\boldsymbol{h}_i = CN(t_i)$, where CN means the ConceptNet Numberbatch embeddings

$$\boldsymbol{h}_{\bar{\mathcal{N}}_i^k} = \frac{1}{|\bar{\mathcal{N}}_i^k|} \sum_{t_j \in \bar{\mathcal{N}}_i^k} CN(t_j), \tag{2}$$

$$\bar{\boldsymbol{h}}_i = \boldsymbol{h}_i + \sum_{k=1}^{K} (\boldsymbol{W}_k \boldsymbol{h}_{\bar{\mathcal{N}}_i^k} + \boldsymbol{b}), \tag{3}$$

GRADE—Dialogue Graph Construction



Rule-based extractor, tf-idf+POS

ConceptNet, initial: $\boldsymbol{h}_i = CN(t_i)$, where CN means the ConceptNet Numberbatch embeddings

$$h_{\bar{\mathcal{N}}_i^k} = \frac{1}{|\bar{\mathcal{N}}_i^k|} \sum_{t_j \in \bar{\mathcal{N}}_i^k} CN(t_j), \qquad (2)$$

$$\bar{h}_i = h_i + \sum_{k=1}^{K} (W_k h_{\bar{\mathcal{N}}_i^k} + b), \qquad (3)$$

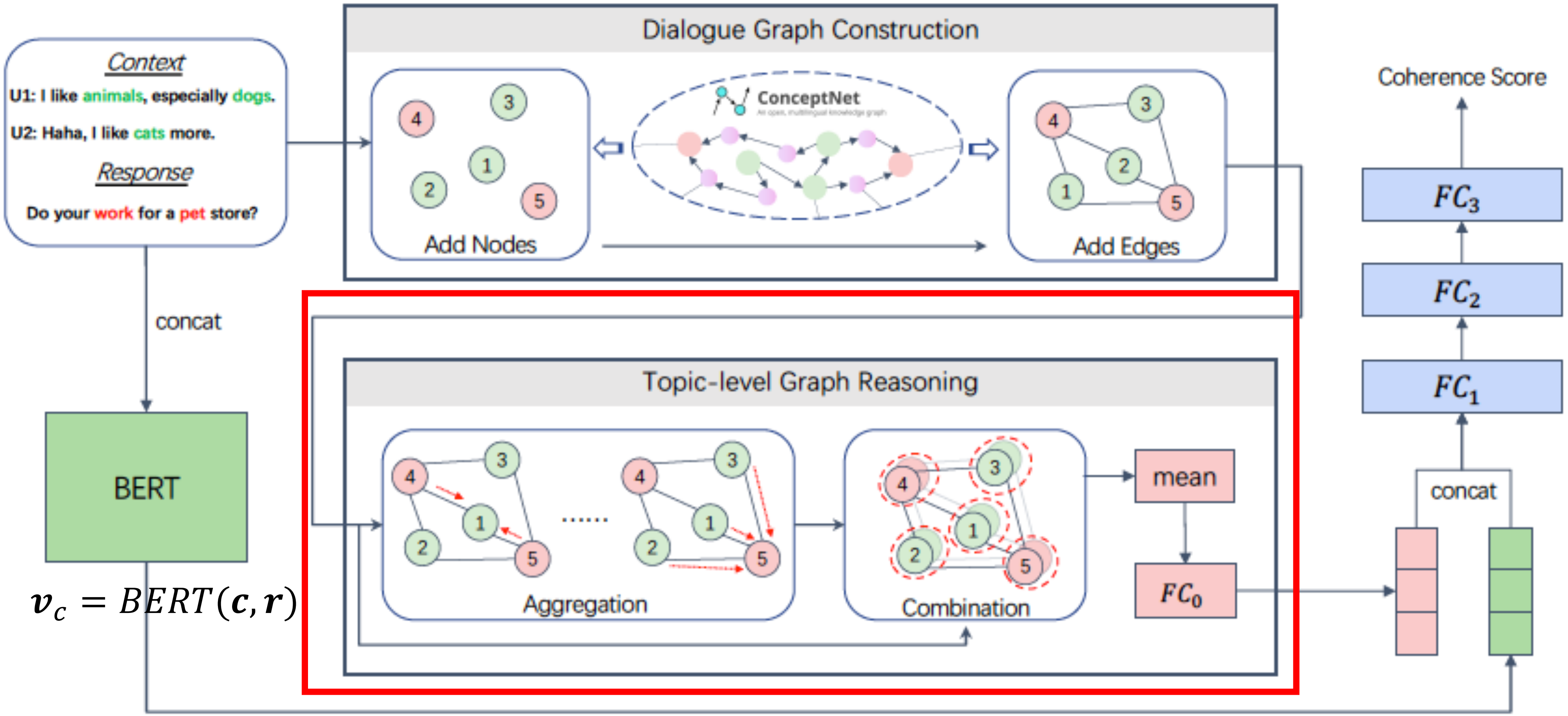
Only consider the edges between the context nodes V_c and the response nodes V_r

$$\boldsymbol{A}[i][j] = \frac{1}{\#hops(V_c^i, V_r^j)}, \qquad (4)$$

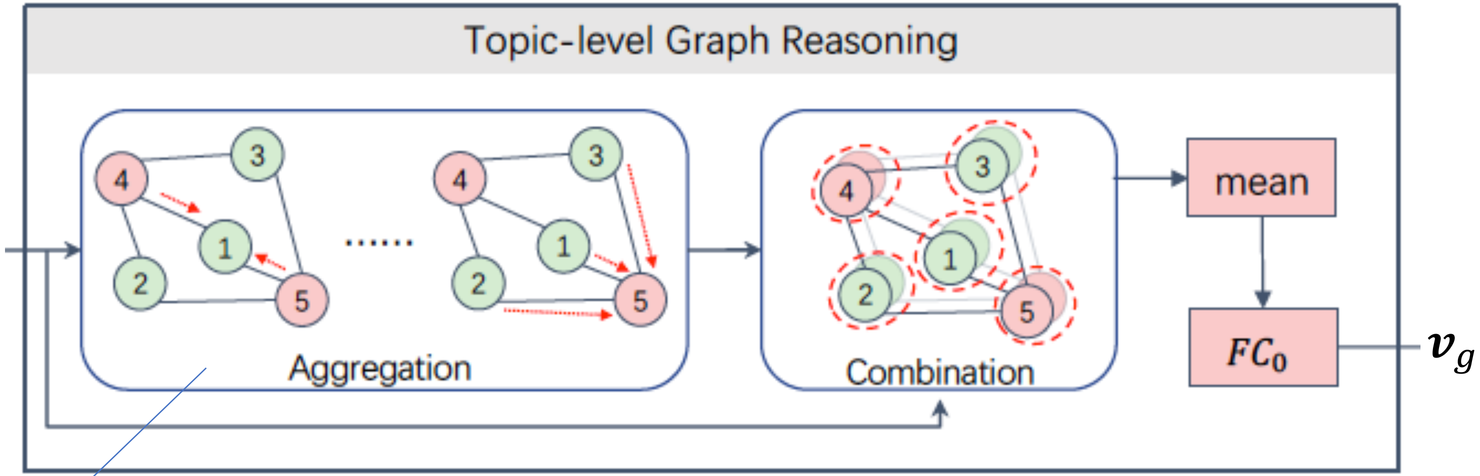
Hop-attention weights

$$\bar{\boldsymbol{A}} = (\boldsymbol{D}+\boldsymbol{I})^{-1/2}(\boldsymbol{A}+\boldsymbol{I})(\boldsymbol{D}+\boldsymbol{I})^{-1/2}, \qquad (5)$$

GRADE—Model


Dialogue Graph Construction
Context
U1: I like animals, especially dogs.
U2: Haha, I like cats more.
Response
Do your work for a pet store?
ConceptNet
Add Nodes
Add Edges
Coherence Score
FC_3
FC_2
FC_1
concat
Topic-level Graph Reasoning
BERT
Aggregation
Combination
mean
FC_0
concat
$v_c = BERT(c, r)$

GRADE—Topic-level Graph Reasoning



$$\boldsymbol{z}_i^{(l)} = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}_l \boldsymbol{h}_j^{(l)}, \tag{6}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{n \in \mathcal{N}_i} \exp(e_{in})}, \tag{7}$$

$$e_{ij} = \bar{\boldsymbol{A}}[i][j] * \rho\left(\boldsymbol{a}_l^T \left[\mathbf{W}_l \boldsymbol{h}_i^{(l)} \| \mathbf{W}_l \boldsymbol{h}_j^{(l)}\right]\right), \tag{8}$$

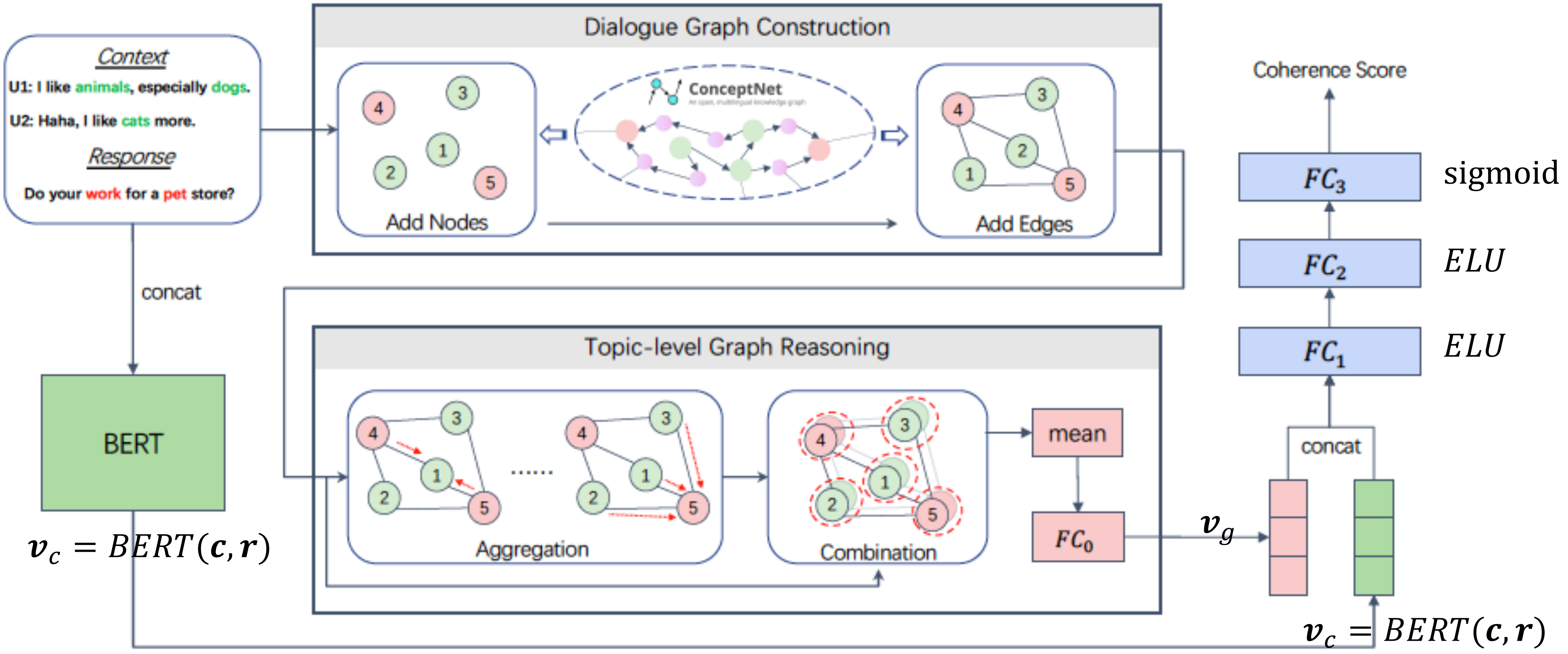
where $\boldsymbol{h}_i^{(0)} = \overline{\boldsymbol{h}}_i$, N_i is the neighboring nodes of t_i in the dialogue graph G, α_{ij} is the attention coefficient, ρ is LeakyReLU.

$$\boldsymbol{h}_i^{(l+1)} = ELU\left(\mathbf{V}_l \boldsymbol{h}_i^{(l)} + \boldsymbol{z}_i^{(l)}\right), \tag{9}$$

$$\boldsymbol{v}_g = FC_0(mean(\{\boldsymbol{h}_i^{(L)} | i \in [1, p+q]\})), \tag{10}$$

where ELU represents an exponential linear unit and FC_0 is a fully-connected layer with an ELU activation.

GRADE—Model



Context
U1: I like animals, especially dogs.
U2: Haha, I like cats more.
Response
Do your work for a pet store?
Dialogue Graph Construction
ConceptNet
Add Nodes
Add Edges
concat
BERT
$v_c = BERT(c, r)$
Topic-level Graph Reasoning
Aggregation
Combination
mean
FC_0
v_g
concat
$v_c = BERT(c, r)$
Coherence Score
FC_3
sigmoid
FC_2
ELU
FC_1
ELU

GRADE—Training

- **Training objective**: margin ranking loss

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} max(0, \bar{s}_i - s_i + m), \qquad (12)$$

Sampled false response

GRADE—Training

- **Training objective**: margin ranking loss

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} max(0, \underbrace{\bar{s}_i}_{\text{Sampled false response}} - s_i + m), \qquad (12)$$

- **Negative sampling:** select negative $\bar{r}$ which is **similar** to ground-truth r
  - **Lexical sampling**: use **Lucene** to retrieve utterances that is related to r from the training dataset and select the middle one as $\bar{r}$
  - **Embedding-based sampling**: randomly sample 1000 utterances -> compute the cosine similarity with r -> randomly select one from the top-5 utterances as $\bar{r}$.

GRADE—Experiments

- **Datasets:**
  - **Training**: DailyDialog
  - **Testing**: totally 1200 context-response pairs with human-annotated coherence score.
    - DailyDialog
      - 150 for Transformer-Ranker
      - 150 for Transformer-Generator
    - ConvAI2
      - 150 for Transformer-Ranker
      - 150 for Transformer-Generator
      - 150 for Bert-Ranker
      - 150 for DialoGPT
    - EmpatheticDialogues
      - 150 for Transformer-Ranker
      - 150 for Transformer-Generator

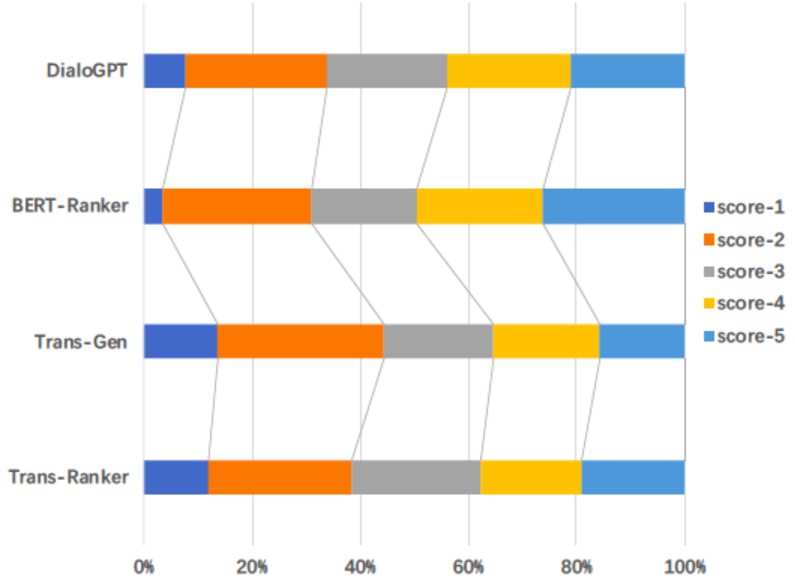


Figure 3: Score distributions of human judgements on the ConvAI2 dataset. Trans-Gen and Trans-Ranker refer to the Transformer-Generator and Transformer-Ranker dialogue models respectively.

GRADE—Experiments

Metric		Transformer-Ranker Pearson	Transformer-Ranker Spearman	Transformer-Generator Pearson	Transformer-Generator Spearman	Average
		DailyDialog				
Statistic-based	BLEU	0.065 *	0.114 *	0.084 *	0.246	0.127
	ROUGE	0.163	0.169	0.138 *	0.126 *	0.149
	METEOR	0.079 *	0.036 *	0.115 *	0.016 *	0.062
Learning-based	BERTScore	0.163	0.138 *	0.214	0.156	0.168
	ADEM	0.162	0.179	0.077 *	0.092 *	0.128
	BERT-RUBER	0.185	0.225	0.142 *	0.182	0.184
	BLEURT	0.230	**0.258**	0.347	0.299	0.284
	GRADE	**0.261**	0.187	**0.358**	**0.368**	**0.294**
		ConvAI2				
Statistic-based	BLEU	0.161	0.240	0.130 *	0.013 *	0.136
	ROUGE	0.177	0.240	0.130 *	0.126 *	0.168
	METEOR	0.215	0.274	0.101 *	0.131 *	0.180
Learning-based	BERTScore	0.310	0.344	0.266	0.241	0.290
	ADEM	-0.015 *	-0.040 *	0.063 *	0.057 *	0.016
	BERT-RUBER	0.204	0.274	0.160	0.173	0.203
	BLEURT	0.259	0.229	0.195	0.200	0.221
	GRADE	**0.535**	**0.558**	**0.606**	**0.617**	**0.579**
		EmpatheticDialogues				
Statistic-based	BLEU	-0.073 *	0.081 *	-0.056 *	-0.089 *	-0.034
	ROUGE	0.170	0.143 *	-0.200	-0.202	-0.022
	METEOR	0.275	0.269	-0.126 *	-0.130 *	0.072
Learning-based	BERTScore	0.184	0.181	-0.087 *	-0.115 *	0.041
	ADEM	0.001 *	-0.004 *	0.087 *	0.086 *	0.042
	BERT-RUBER	0.021 *	-0.034 *	-0.128 *	-0.177	-0.080
	BLEURT	0.187	0.181	0.017 *	-0.031 *	0.090
	GRADE	**0.375**	**0.338**	**0.257**	**0.223**	**0.298**

Table 1: Correlations between automatic evaluation metrics and human judgements on three different datasets (DailyDialog, ConvAI2 and EmpatheticDialogues) and two dialogue models (Transformer-Ranker and Transformer-Generator). The star * indicates results with p-value > 0.05, which are not statistically significant.

GRADE—Experiments

	Bert-Ranker		DialoGPT	
	Pearson	Spearman	Pearson	Spearman
ROUGE	0.157	0.121 *	0.084 *	0.098 *
METEOR	0.070 *	0.088 *	0.020 *	0.029 *
BERTScore	0.165	0.135 *	0.208	0.177
BERT-RUBER	0.141 *	0.111 *	0.113 *	0.085 *
BLEURT	0.133 *	0.071 *	0.273	0.275
GRADE	**0.502**	**0.425**	**0.487**	**0.485**

Table 2: Correlations between auto-metrics and human judgements on the ConvAI2 dataset and two dialogue models, Bert-Ranker and DialoGPT, respectively.

GRADE—Experiments

Metric	Transformer-Ranker		Transformer-Generator		Average
	Pearson	Spearman	Pearson	Spearman	
Our GRADE ($N_1 = 10, N_2 = 10$)	**0.227** ±0.018	**0.162** ±0.015	**0.364** ±0.017	**0.372** ±0.018	**0.281** ±0.008
random sampling	0.225 ±0.022	0.153 ⋆ ±0.016	0.237 ±0.034	0.245 ±0.028	0.215 ±0.023
no graph branch	0.211 ±0.028	0.146 ⋆ ± 0.020	0.324 ±0.034	0.336 ±0.029	0.254 ±0.024
no k-hop neighboring representations	0.219 ±0.011	0.153 ⋆ ± 0.008	0.347 ±0.032	0.356 ±0.034	0.269 ±0.019
no hop-attention weights	**0.227** ±0.013	**0.162** ±0.012	0.349 ±0.019	0.352 ±0.015	0.273 ±0.007
1-hop neighboring representations ($N_1 = 10$)	0.211 ±0.022	0.150 ⋆ ±0.019	0.347 ±0.014	0.352 ±0.017	0.265 ±0.018
1-hop neighboring representations ($N_1 = 20$)	0.206 ±0.025	0.148 ⋆ ±0.015	0.356 ±0.030	0.358 ±0.032	0.267 ±0.025
2-hop neighboring representations ($N_1 = 20, N_2 = 20$)	0.216 ±0.016	0.150 ⋆ ±0.014	0.360 ±0.019	0.364 ±0.017	0.273 ±0.015

Table 3: Ablation results on the DailyDialog dataset, averaged across five random seeds, with standard deviations presented in gray color. N_1 and N_2 refer to the numbers of the 1^{st} and 2^{nd} hop neighboring nodes in ConceptNet, respectively. The symbol $\star$ indicates that three or more than three correlation results over the five random seeds are not statistically significant, namely, p-value > 0.05.

GRADE—Experiments

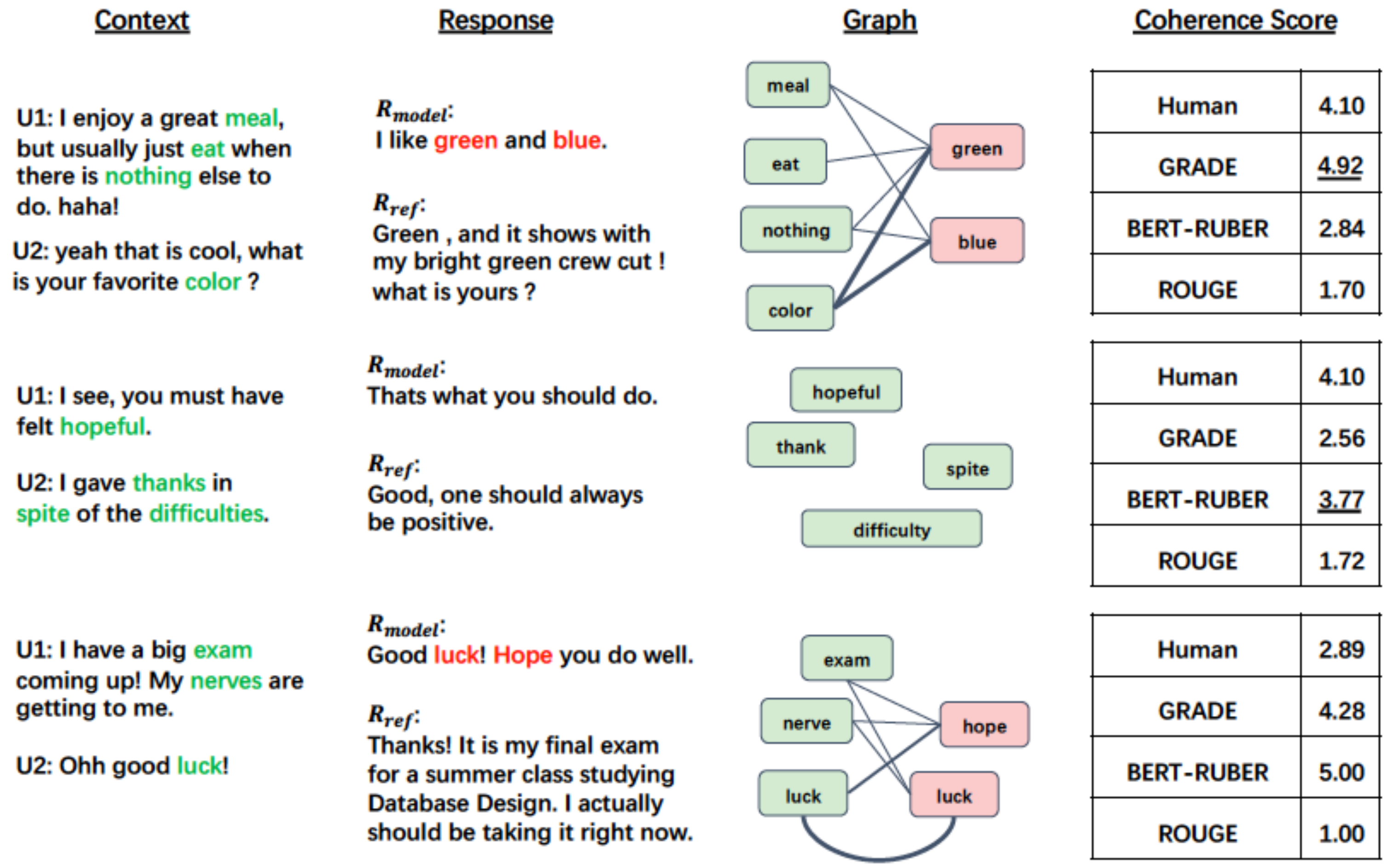


Figure 5: Visualization results of our GRADE, compared with two baseline metrics, ROUGE and BERT-RUBER. Keywords of the contexts and the model responses R_{model} are highlighted in green and red respectively. R_{ref} is the reference response. For comparison, the auto-metric scores are normalized to the range of human scores, i.e., [1,5].

GRADE—Conclusions

- First attempt to introduce **graph reasoning** into dialogue evaluation
- **SOTA** performance for dialogue coherence evaluation
- One limitation:
    - the **inconsistency** between the training objective (relative ranking) and the expected behavior (absolute scoring)

UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation

Jian Guan, Minlie Huang*
Department of Computer Science and Technology, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
j-guan19@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

UNION—Background

Leading Context
Jack was at the bar.

Reference By Human
He noticed a phone on the floor. He was going to take it to lost and found. But it started ringing on the way. Jack answered it and returned it to the owner's friends.

Sample 1 (Reasonable, B=0.29, M=0.49, U=1.00)
On the way out he noticed a phone on the floor. He asked around if anybody owned it. Eventually he gave it to the bartender. They put it into their lost and found box.

Sample 2 (Reasonable, B=0.14, M=0.27, U=1.00)
He had a drinking problem. He kept having more beers. After a while he passed out. When he waked up, he was surprised to find that he lost over a hundred dollars.

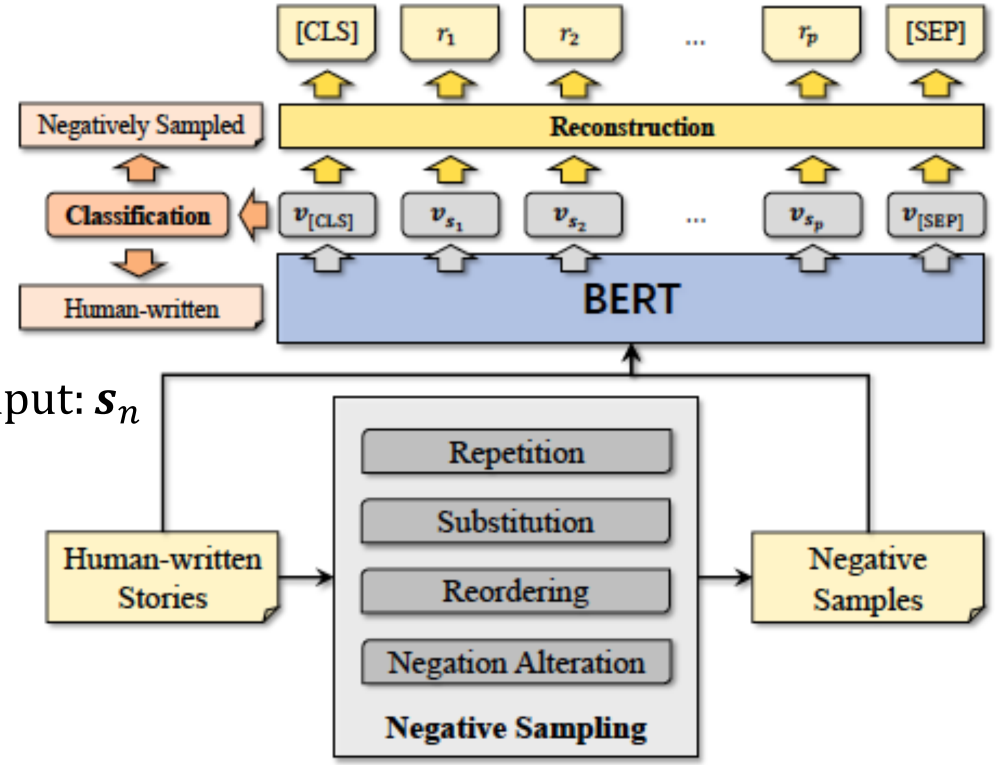
Sample 3 (Unreasonable, B=0.20, M=0.35, U=0.00)
He was going to get drunk and get drunk. The bartender told him it was already time to leave. Jack started drinking. Jack wound up returning but cops came on the way home.

B: BLEU

M: MoverScore

U: UNION, A **UN**referenced metr**I**c for evaluating **O**pen-e**N**ded story generation

UNION—Model



y_n
classification loss

input: $\boldsymbol{s}_n$

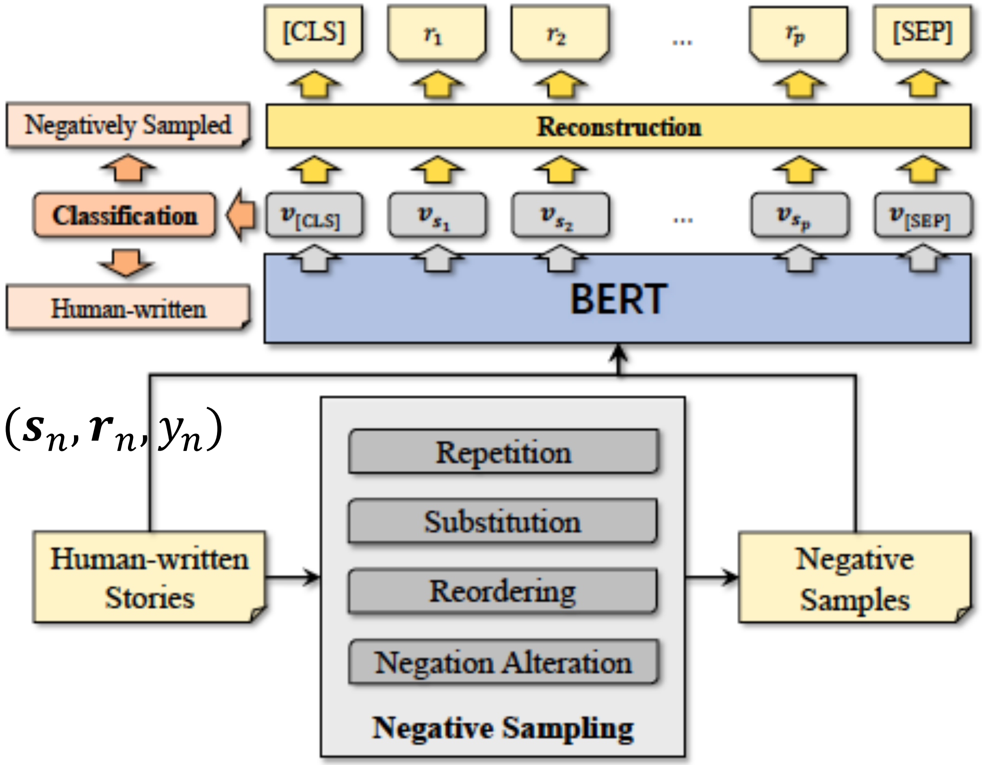
$\boldsymbol{r}_n$, reconstruction loss

Figure 1: Overview of the UNION metric. UNION is trained to distinguish the human-written stories from the negative samples constructed by four negative sampling techniques, as well as to reconstruct the original human-written stories.

- $\boldsymbol{s}_n$ is a human-written story or an auto-constructed negative sample
- $\boldsymbol{r}_n$ is the corresponding original human-written story of $\boldsymbol{s}_n$
- $y_n \in \{0, 1\}$ indicates whether $\boldsymbol{s}_n$ is written by human

UNION—Model

y_n
classification loss



input: $(\boldsymbol{s}_n, \boldsymbol{r}_n, y_n)$

Figure 1: Overview of the UNION metric. UNION is trained to distinguish the human-written stories from the negative samples constructed by four negative sampling techniques, as well as to reconstruct the original human-written stories.

$\boldsymbol{r}_n$, reconstruction loss

Classification loss

$$\boldsymbol{v}_{[\mathrm{CLS}]}, \boldsymbol{v}_{s_1}, \cdots, \boldsymbol{v}_{s_p}, \boldsymbol{v}_{[\mathrm{SEP}]} = \mathrm{BERT}(\boldsymbol{s}_n), \quad (1)$$

$$\hat{y}_n = \mathrm{sigmoid}(\mathbf{W}_c \boldsymbol{v}_{[\mathrm{CLS}]} + \mathbf{b}_c), \quad (2)$$

$$\mathcal{L}_n^C = -y_n \log \hat{y}_n - (1 - y_n) \log (1 - \hat{y}_n). \quad (3)$$

Reconstruction loss

$$P(\hat{r}_i | \boldsymbol{s}_n) = \mathrm{softmax}(\mathbf{W}_r \boldsymbol{v}_{s_i} + \mathbf{b}_r), \quad (4)$$

$$\mathcal{L}_n^R = -\frac{1}{p} \sum_{i=1}^{p} \log P(\hat{r}_i = r_i | \boldsymbol{s}_n), \quad (5)$$

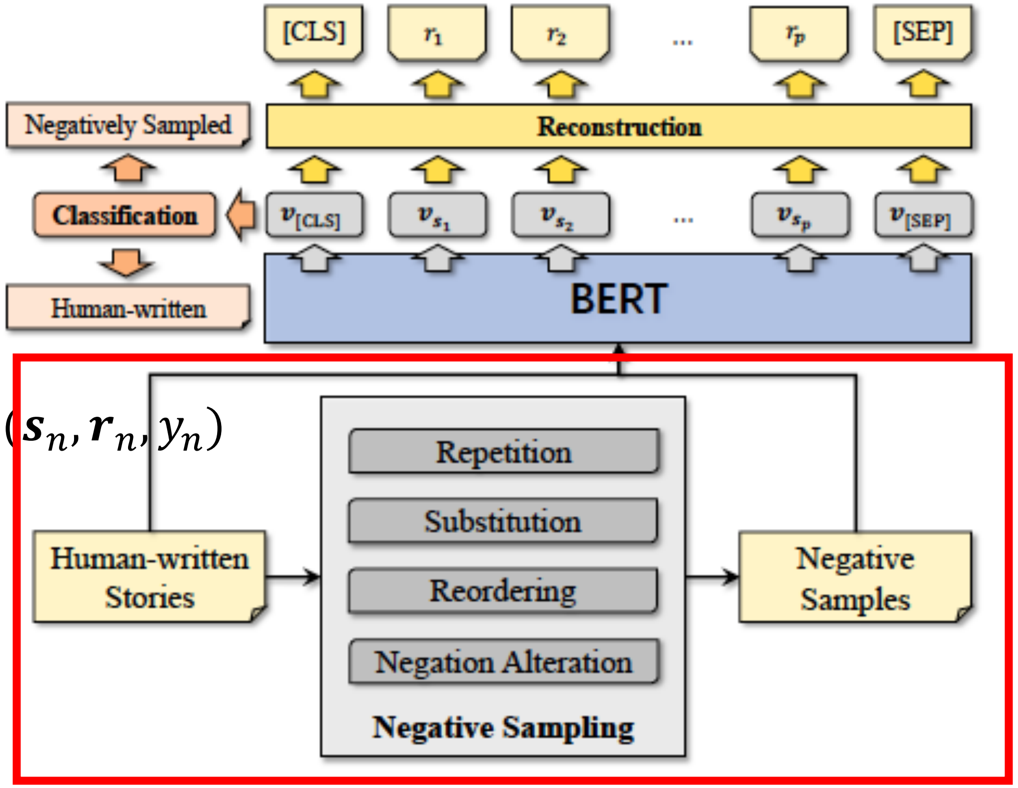
Total loss

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} (\mathcal{L}_n^C + \lambda \mathcal{L}_n^R), \quad (6)$$

UNION—Negative Sampling

y_n
classification loss

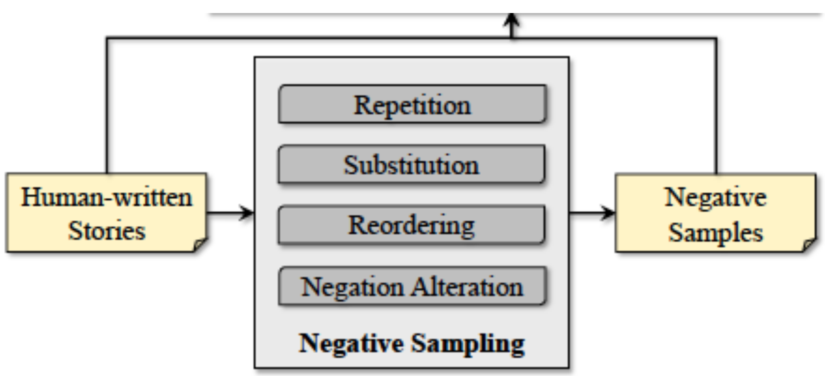
input: $(\boldsymbol{s}_n, \boldsymbol{r}_n, y_n)$



$\boldsymbol{r}_n$, reconstruction loss

Figure 1: Overview of the UNION metric. UNION is trained to distinguish the human-written stories from the negative samples constructed by four negative sampling techniques, as well as to reconstruct the original human-written stories.

UNION—Negative Sampling

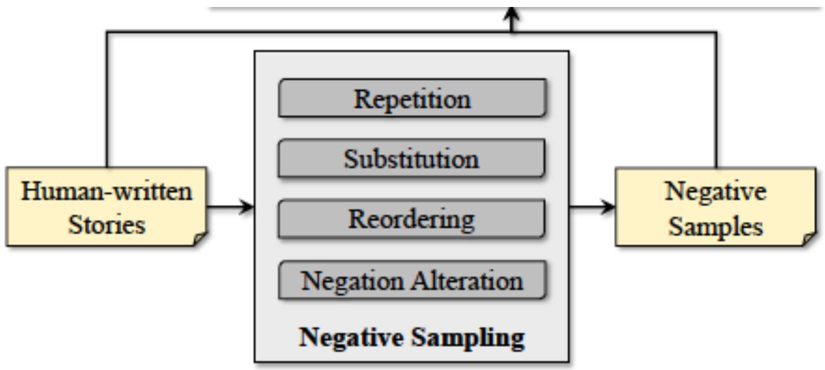


- Motivation is from empirical observations of major errors that make a story unreasonable

Type	**Repe**	**Cohe**	**Conf**	**Chao**	**Others**
Prop (%)	44.1	56.2	67.5	50.4	12.9

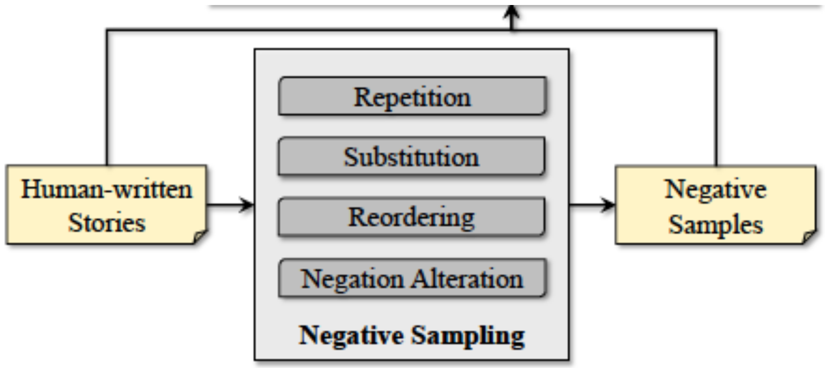
Table 2: Error type **Prop**ortions of 381 unreasonable stories, including **Repe**ated plots/poor **Cohe**rence/**Conf**licting logic/**Chao**tic scenes/**Others**.

UNION—Negative Sampling



- **Repetition**:
  - N-gram (N=1,2,3,4) in a random sentence
  - or randomly select a sentence to repeat and remove the following sentence
- **Substitution**:
  - *word-level*: replace random 15% keywords in a story with their corresponding antonyms, otherwise with another random keyword sampled from all the keywords of the same part-ofspeech (POS), according to the mention frequency
  - *sentence-level*: randomly replace a sentence in a story with another one sampled from the rest of stories in the dataset
- **Reordering**:
  - randomly reorder the sentences in a story to create negative stories with conflicting plot
- **Negation Alteration**:
  - adding or removing negation words using rules for different types of verbs

UNION—Negative Sampling



- **Step 1**: sample the number (n) of techniques from {1,2,3,4} with a distribution {50%, 20%, 20%, 10%}
- **Step 2**: sample a technique without replacement from {repetition, substitution, reordering, negation alteration} with a distribution {10%, 30%, 40%, 20%} until the total number of techniques (n) is reached
- **Step 3**: apply the sampled techniques on a human-written story to obtain a negative sample

- **Repetition**:
  - N-gram (N=1,2,3,4) in a random sentence
  - or randomly select a sentence to repeat and remove the following sentence
- **Substitution**:
  - *word-level*: replace random 15% keywords in a story with their corresponding antonyms, otherwise with another random keyword sampled from all the keywords of the same part-of-speech (POS), according to the mention frequency
  - *sentence-level*: randomly replace a sentence in a story with another one sampled from the rest of stories in the dataset

- **Reordering**:
  - randomly reorder the sentences in a story to create negative stories with conflicting plot
- **Negation Alteration**:
  - adding or removing negation words using rules for different types of verbs

UNION—Experiments: Datasets

Split	Metrics	ROC	WP	NS
Train/ Validate	Perplexity DisScore $\mathbf{RUBER}_u$ UNION	88,344/ 4,908	272,600/ 15,620	✗ ✓ ✓ ✓
	BLEURT	$360^{\dagger}/40^{\dagger}$	$180^{\dagger}/20^{\dagger}$	✗
Test	All metrics	$400^{\dagger}$	$200^{\dagger}$	N/A

Table 4: Data statistics. $\mathbf{RUBER}_u$ is short for $\mathbf{RUBER}_u$**-BERT**. **NS** (Negative Sampling) means whether a metric requires negative samples for training/validation. † means the stories are generated by NLG models and manually annotated.

UNION—Experiments: Main Results

Metrics		ROC			WP		
		r	ρ	τ	r	ρ	τ
Referenced	BLEU	0.0299	0.0320	0.0231	0.1213	0.0941	0.0704
	MoverScore	0.1538*	0.1535*	0.1093*	0.1613	0.1450	0.1031
	RUBER$_r$-BERT	0.0448	0.0517	0.0380	0.1502	0.1357	0.0986
Unreferenced	Perplexity	0.2464*	0.2295*	0.1650*	-0.0705	-0.0479	-0.0345
	RUBER$_u$-BERT	0.1477*	0.1434*	0.1018*	0.1613	0.1605	0.1157
	DisScore	0.0406	0.0633	0.0456	0.0627	-0.0234	-0.0180
	UNION	**0.3687***	**0.4599***	**0.3386***	**0.3663***	**0.4493***	**0.3293***
	-Recon	0.3101*	0.4027*	0.2927*	0.3292*	0.3786*	0.2836*
Hybrid	RUBER-BERT	0.1412*	0.1395*	0.1015*	0.1676	0.1664	0.1194
	BLEURT	0.2310*	0.2353*	0.1679*	0.2229*	0.1602	0.1180

Table 5: Correlation with human judgments on ROC and WP datasets. $r/\rho/\tau$ indicates the Pearson/Spearman/Kendall correlation, respectively. The best performance is highlighted in **bold**. The correlation scores marked with * indicate the result significantly correlates with human judgments (p-value<0.01).

UNION—Experiments: Dataset Drift Setting

Metrics	r	ρ	τ
Training: WP Test: ROC			
Perplexity	-0.0015	0.0149	0.0101
RUBER$_u$-BERT	-0.0099	-0.0162	-0.0110
BLEURT	0.1326*	0.1137*	0.0828*
UNION	**0.1986***	**0.2501***	**0.1755***
-Recon	0.1704*	0.2158*	0.1523*
Training: ROC Test: WP			
Perplexity	0.0366	0.0198	0.0150
RUBER$_u$-BERT	0.1392	0.1276	0.0912
BLEURT	0.1560	0.1305	0.0941
UNION	**0.2872***	**0.2935***	**0.2142***
-Recon	0.2397*	0.2712*	0.1971*

Table 6: Correlation results in the dataset drift setting where the metrics are trained on one dataset and then used for the other one.

UNION—Experiments: Quality Drift Setting

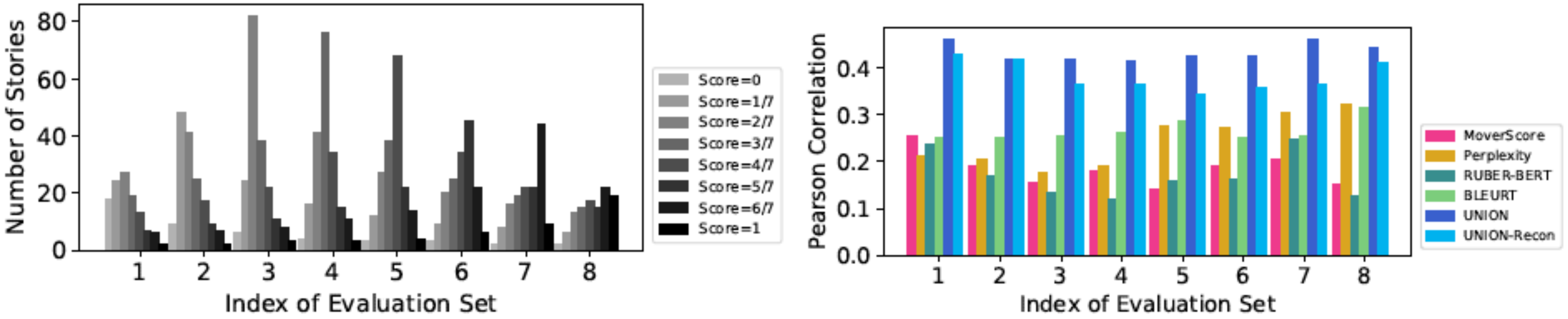


Figure 2: Generalization over different biased test sets. Left: distribution of stories of different annotation scores in different test sets. Right: the Pearson correlation of different metrics with human judgments on different test sets, where UNION-Recon denotes UNION without the reconstruction task.

UNION—Conclusions

- a learnable metric UNION for evaluating open-ended story generation to **alleviate the one-to-many issue** of referenced metrics
- **SOTA** performance and **better generalization ability** to data drift and quality drift