GNN & NLP

im0qianqian

East China Normal University
School of Computer Science and Technology

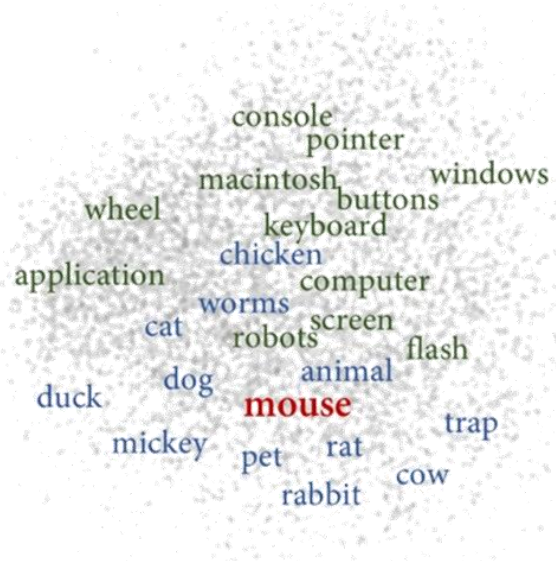
Outline

1. Introduction
2. Graph-Bert: Only Attention is Needed for Learning Graph Representations
3. Multi-Paragraph Reasoning with Knowledge-enhancedGraph Neural Network
4. GSN: A Graph-Structured Network for Multi-Party Dialogues
5. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation

Introduction

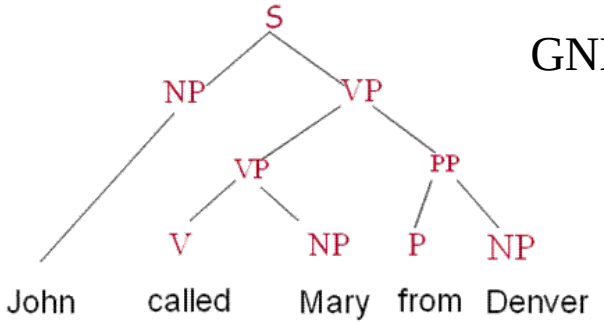
知识图谱

GNN



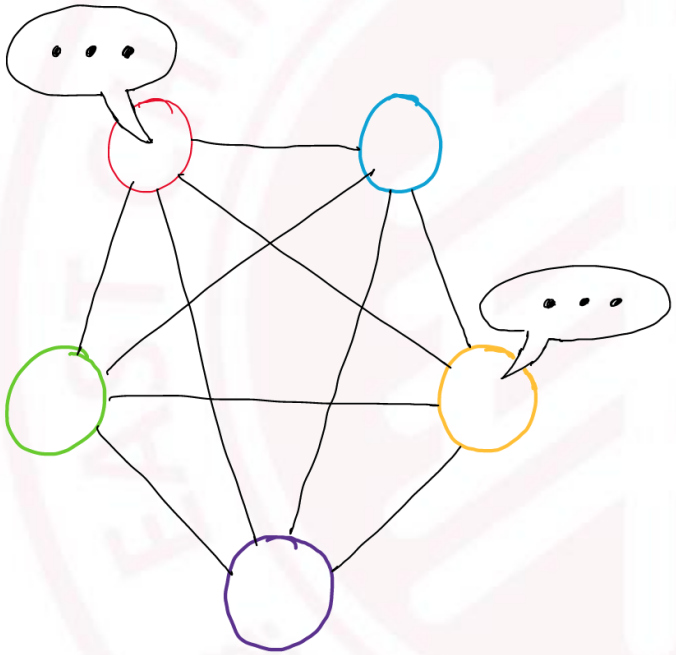
分布式表示

GNN



文本中的语法树

GNN



对话模型

Graph-Bert (Introduction)

背景：以往的 GNN 存在很多问题

- suspended animation
- over-smoothing
- 难以并行化
- 难以迁移到其他任务

Jiawei Zhang, Haopeng Zhang, Congying Xia, Li Sun. Graph-Bert: Only Attention is Needed for Learning Graph Representations. arXiv:2001.05140

Graph-Bert (Model architecture)

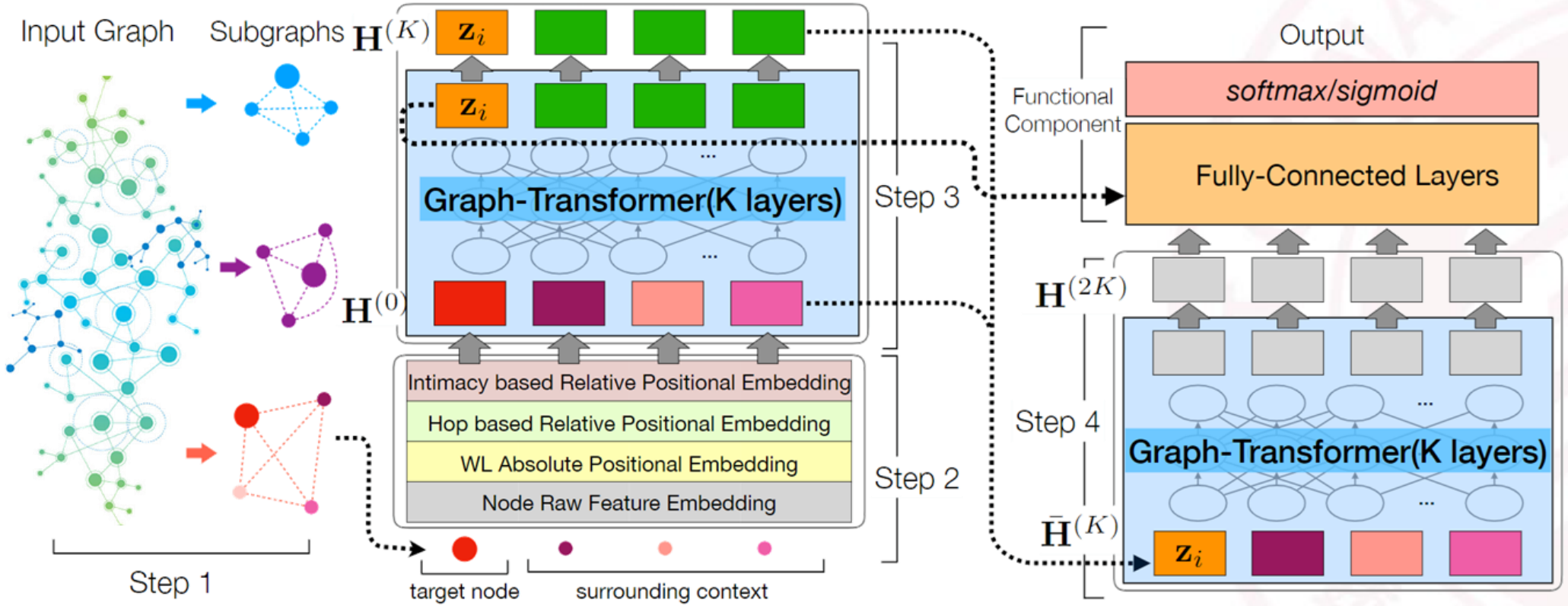


Figure 1: Architecture of the GRAPH-BERT Model. (Step 1: linkless subgraph batching; Step 2: node input vector embeddings; Step 3: graph transformer based encoder; Step 4: graph transformer based decoder. Depending on the target application task, the function component will generate different output. In the sampled subgraphs, the larger node denotes the target node and the remaining ones are its context.)

Jiawei Zhang, Haopeng Zhang, Congying Xia, Li Sun. Graph-Bert: Only Attention is Needed for Learning Graph Representations. arXiv:2001.05140

Graph-Bert (Linkless Subgraph Batching)

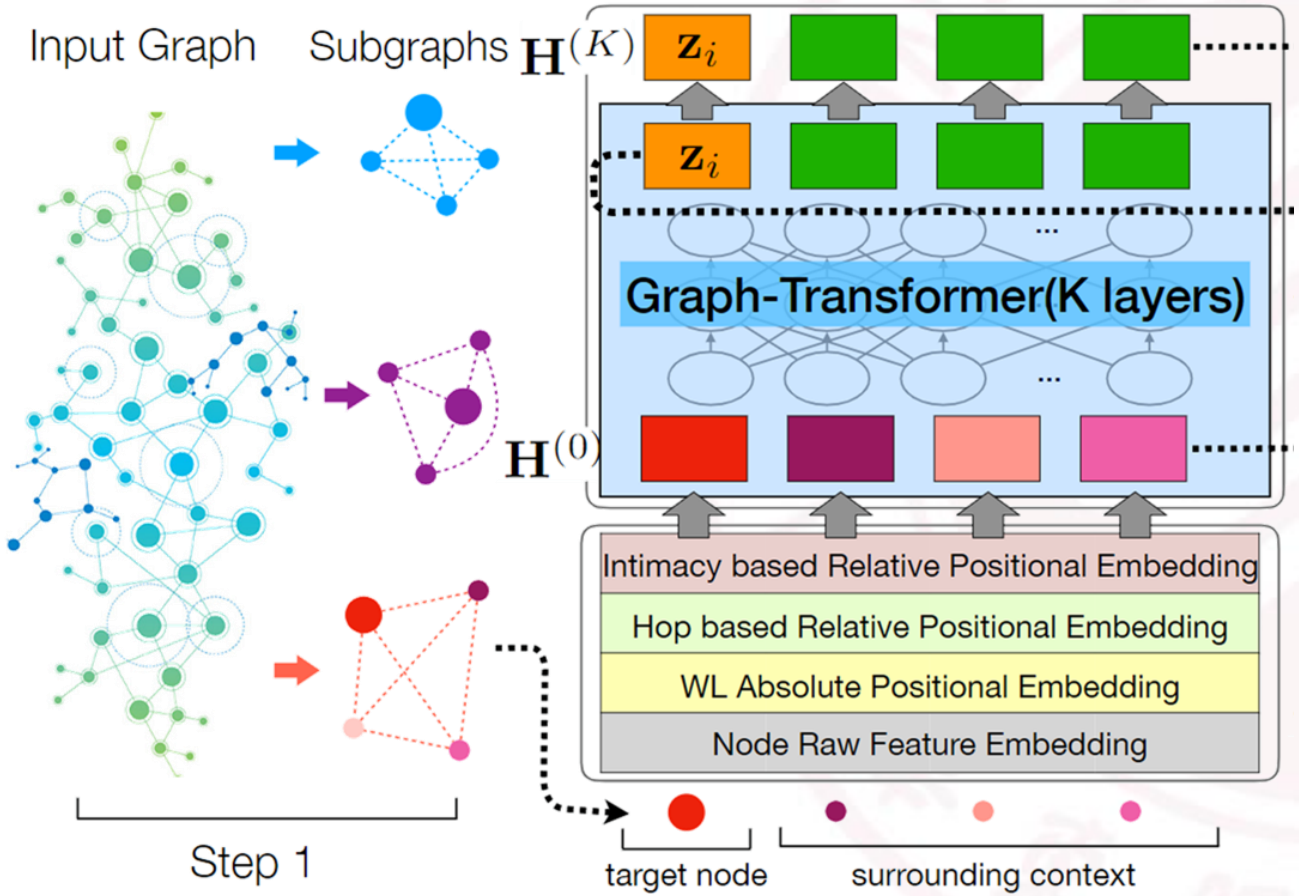
无边子图分解（采样）：

- 中心节点
- K个邻接节点

选取邻接节点（top-K）：

- 计算节点间关联程度
- 选取最大的K个节点作为邻接节点

$$\mathbf{S} = \alpha \cdot \left(\mathbf{I} - (1-\alpha)\cdot \bar{\mathbf{A}}\right)^{-1}$$
$$\bar{\mathbf{A}} = \mathbf{A}\mathbf{D}^{-1}$$



Jiawei Zhang, Haopeng Zhang, Congying Xia, Li Sun. Graph-Bert: Only Attention is Needed for Learning Graph Representations. arXiv:2001.05140

Graph-Bert (Node Input Vector Embeddings)

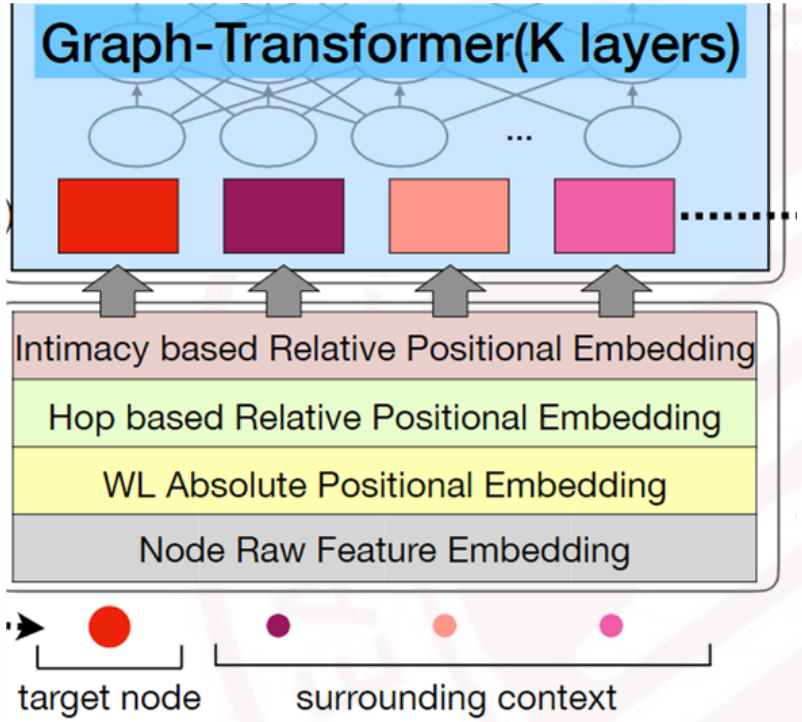
Raw Feature Vector Embedding

$$\mathbf{e}_j^{(x)} = \text{Embed}\left(\mathbf{x}_j; \mathbf{W}_x, \mathbf{b}_x\right), \text{ and } \mathbf{e}_j^{(x)} \in \mathbb{R}^{d_h \times 1}$$

Weisfeiler-Lehman Absolute Role Embedding

$$\mathbf{e}_j^{(r)} = \text{Position-Embed}\left(\text{WL}(v_j)\right)$$

$$= \left[sin\left(\frac{\text{WL}(v_j)}{10000^{\frac{2l}{d_h}}}\right), cos\left(\frac{\text{WL}(v_j)}{10000^{\frac{2l+1}{d_h}}}\right) \right]_{l=0}^{\lfloor \frac{d_h}{2} \rfloor}$$



Intimacy based Relative Positional Embedding

$$\mathbf{e}_j^{(p)} = \text{Position-Embed}\left(\text{P}(v_j)\right) \in \mathbb{R}^{d_h \times 1}$$

Hop based Relative Distance Embedding

$$\mathbf{e}_j^{(d)} = \text{Position-Embed}\left(\text{H}(v_j; v_i)\right) \in \mathbb{R}^{d_h \times 1}$$

Jiawei Zhang, Haopeng Zhang, Congying Xia, Li Sun. Graph-Bert: Only Attention is Needed for Learning Graph Representations. arXiv:2001.05140

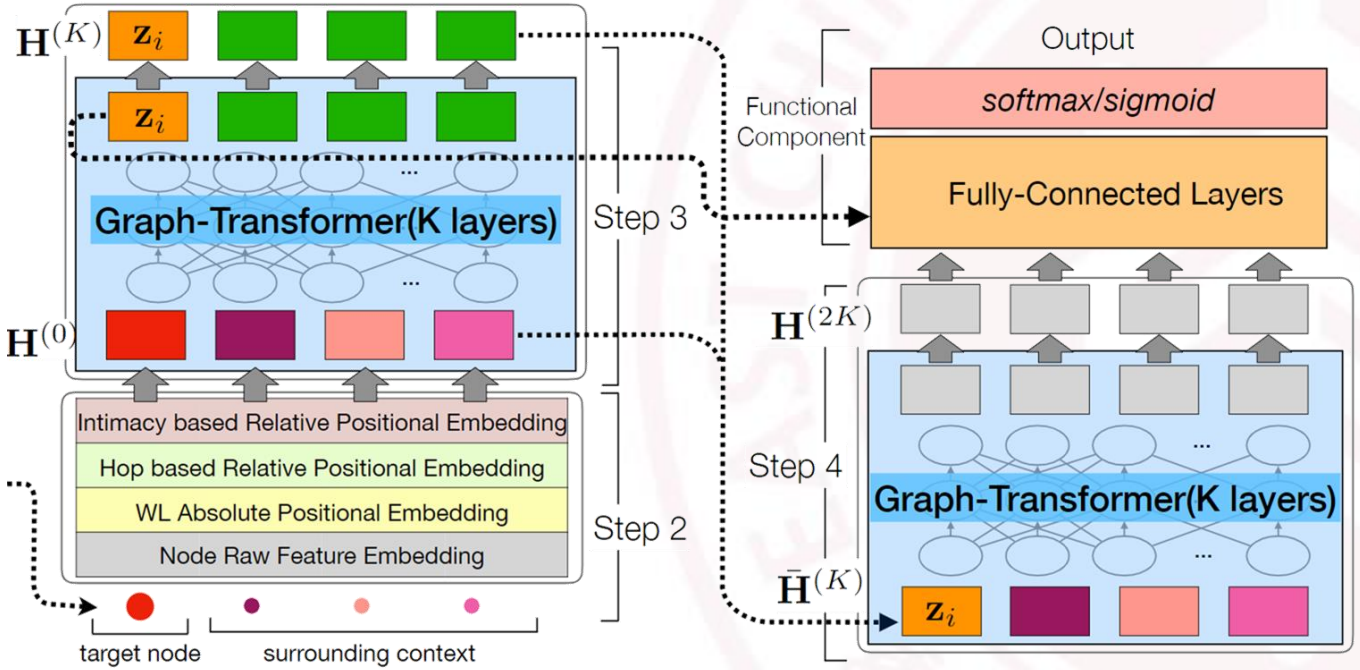
Graph-Bert (Graph Transformer based Encoder)

Transformer Input：

$$\mathbf{h}_j^{(0)} = \text{Aggregate}(\mathbf{e}_j^{(x)}, \mathbf{e}_j^{(r)}, \mathbf{e}_j^{(p)}, \mathbf{e}_j^{(d)})$$

Attention：

$$\mathbf{H}^{(l)} = \text{softmax}\left(\mathbf{M}^{(l)} \otimes \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \forall l \in \{1, \cdots, 2K\}$$



Jiawei Zhang, Haopeng Zhang, Congying Xia, Li Sun. Graph-Bert: Only Attention is Needed for Learning Graph Representations. arXiv:2001.05140

Graph-Bert (Pre-training & Transfer & Fine-tuning)

Pre-training Task

- 节点原始特征重建

  类似于 auto-encoder

- 图结构重建

  保证学习到了图的结构信息

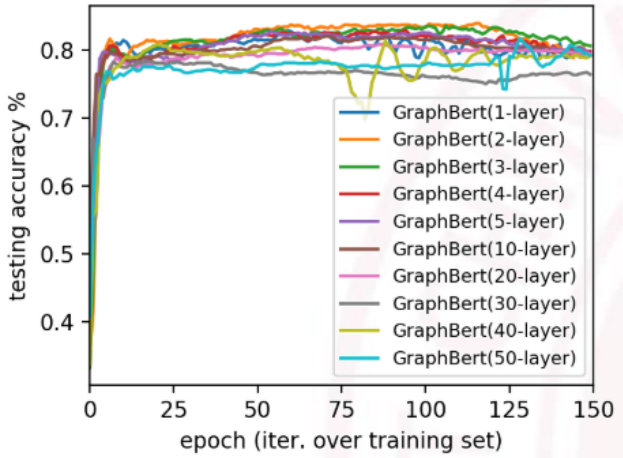
Transfer & Fine-tuning：

- 节点分类

  直接新增 FC 层，以及 softmax 分类

- 图聚类

  直接在学习到的节点特征中聚类即可



Methods	Datasets (Accuracy)		
	Cora	**Citeseer**	**Pubmed**
LP ([Zhu *et al.*, 2003])	0.680	0.453	0.630
ICA ([Lu and Getoor, 2003])	0.751	0.691	0.739
ManiReg ([Belkin *et al.*, 2006])	0.595	0.601	0.707
SemiEmb ([Weston *et al.*, 2008])	0.590	0.596	0.711
DeepWalk ([Perozzi *et al.*, 2014b])	0.672	0.432	0.653
Planetoid ([Yang *et al.*, 2016])	0.757	0.647	0.772
MoNet ([Monti *et al.*, 2016])	0.817	-	0.788
GCN ([Kipf and Welling, 2016])	0.815	0.703	**0.790**
GAT ([Veličković *et al.*, 2018])	**0.830**	**0.725**	**0.790**
LOOPYNET ([Zhang, 2018])	**0.826**	**0.716**	**0.792**
GRAPH-BERT	**0.843**	**0.712**	**0.793**

Jiawei Zhang, Haopeng Zhang, Congying Xia, Li Sun. Graph-Bert: Only Attention is Needed for Learning Graph Representations. arXiv:2001.05140

KGNN (Introduction)

任务背景：

目前多段推理下的 OpenQA 很少被关注

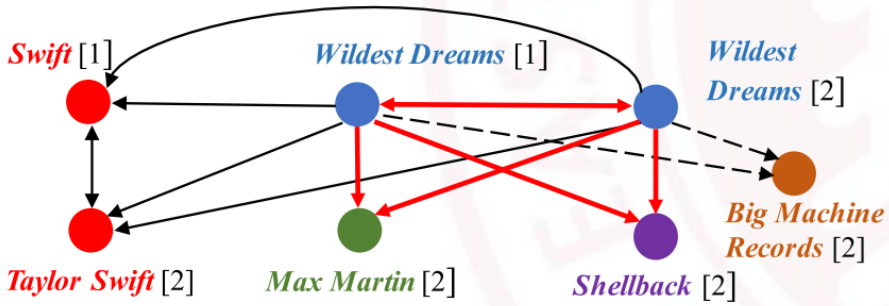
现有的处理方案：

- 将每个段落单独用来处理
- 将所有段落连接成一个长文本

数据集：

HotpotQA

[1] The 2015 MTV Video Music Awards were held on August 30, 2015. … ***Swift***'s "***Wildest Dreams***" music video premiered during the pre-show. …
[2] "***Wildest Dreams***" is a song recorded by American singer-songwriter ***Taylor Swift*** for her fifth studio album, "1989". The song was released to radio by ***Big Machine Records*** on August 31, 2015, as the album's fifth single. **Swift** co-wrote the song with its producers ***Max Martin*** and ***Shellback***. …



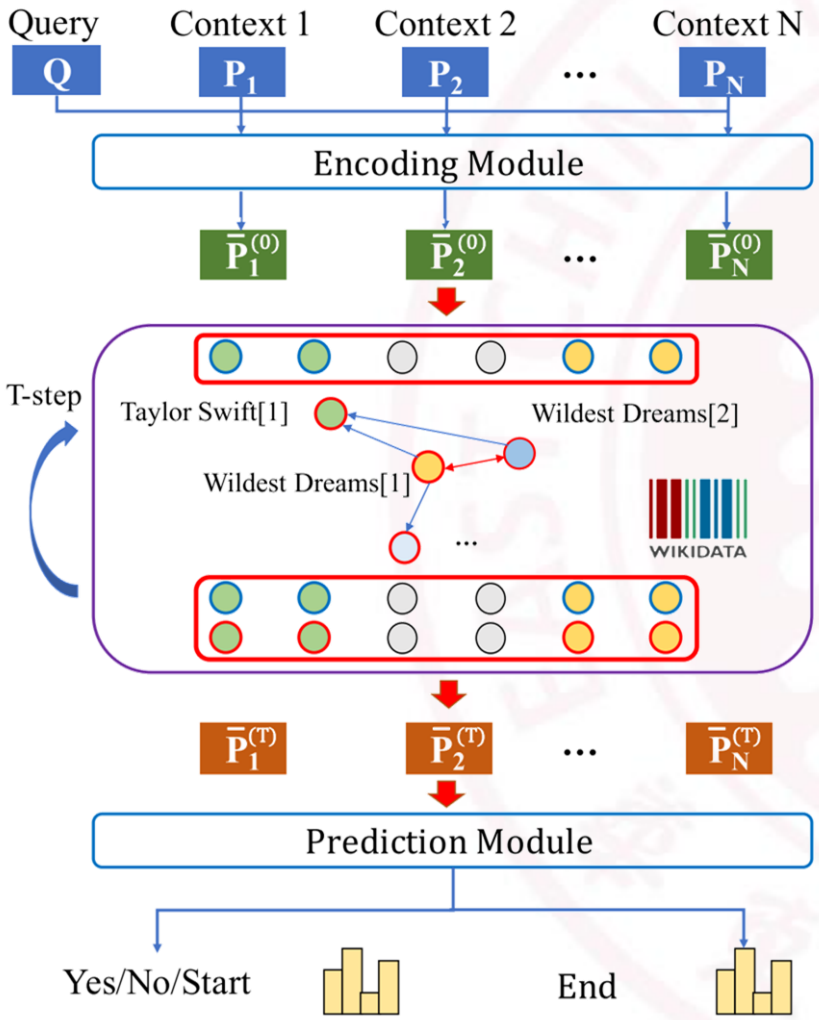
Question: Who co-wrote Taylor Swift's song that had its music video premiere during the pre-show of the 2015 MTV Video Music Awards?
Answer: ***Max Martin*** and ***Shellback***

Deming Ye, Yankai Lin et al. Multi-Paragraph Reasoning with Knowledge-enhanced Graph Neural Network. arXiv:1911.02170

KGNN (Model architecture)

模型结构：

- Encoding module
- Reasoning module
  - 构建实体图
  - 关系推理
- Prediction module
  - 预测答案所在的范围（span）
  - 预测是否包含关键信息（Yes/No）



Deming Ye, Yankai Lin et al. Multi-Paragraph Reasoning with Knowledge-enhanced Graph Neural Network. arXiv:1911.02170

KGNN (Encoding module)

Input：Q 代表问题，P_i代表第 i 个段落

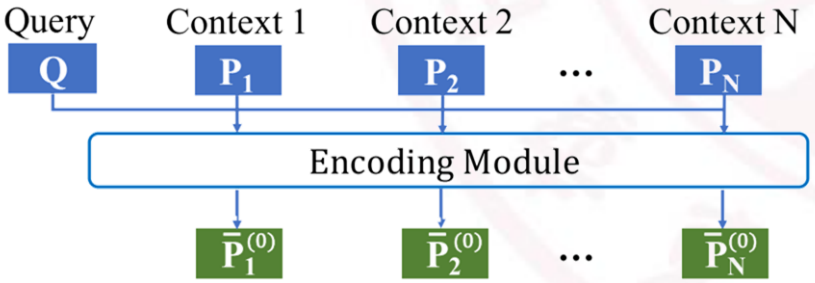
$$\boldsymbol{Q} = \texttt{Self-Att}(\texttt{Char-Enc}(Q)),$$
$$\boldsymbol{P}_i = \texttt{Self-Att}(\texttt{Char-Enc}(P_i)),$$

$$\bar{\boldsymbol{P}}_i^{(0)} = \texttt{Bi-Att}(\boldsymbol{Q}, \boldsymbol{P}_i),$$

Output：$\bar{P}_i^{(0)}$代表与问题相关的段落表示

[1] The 2015 MTV Video Music Awards were held on August 30, 2015. … ***Swift***'s "***Wildest Dreams***" music video premiered during the pre-show. …
[2] "***Wildest Dreams***" is a song recorded by American singer-songwriter ***Taylor Swift*** for her fifth studio album, "1989". The song was released to radio by ***Big Machine Records*** on August 31, 2015, as the album's fifth single. **Swift** co-wrote the song with its producers ***Max Martin*** and ***Shellback***. …

Question: Who co-wrote Taylor Swift's song that had its music video premiere during the pre-show of the 2015 MTV Video Music Awards?

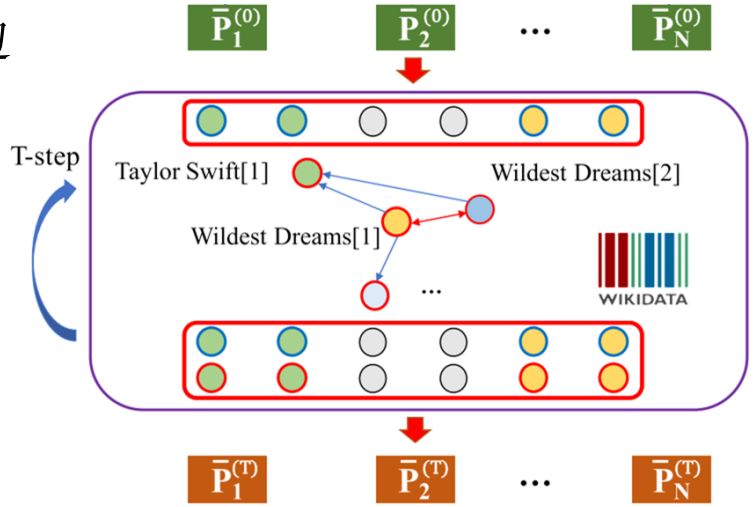


Deming Ye, Yankai Lin et al. Multi-Paragraph Reasoning with Knowledge-enhanced Graph Neural Network. arXiv:1911.02170

KGNN (Reasoning module)

构建实体图

- **选点：**所有段落中出现的实体作为图的节点
- **连边：**
  - 若两个节点表示同一实体，则建边
  - 若两个节点在知识库中有关系 r，则建立关系为 r 的边

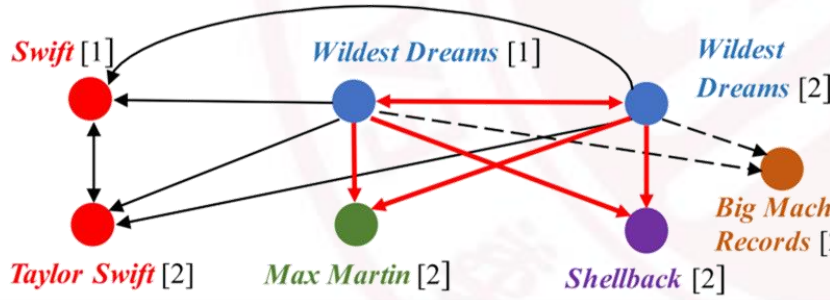


关系推理

异质图 GNN

refine 实体表示

[1] The 2015 MTV Video Music Awards were held on August 30, 2015. … *Swift*'s "*Wildest Dreams*" music video premiered during the pre-show. …
[2] "*Wildest Dreams*" is a song recorded by American singer-songwriter *Taylor Swift* for her fifth studio album, "1989". The song was released to radio by *Big Machine Records* on August 31, 2015, as the album's fifth single. **Swift** co-wrote the song with its producers *Max Martin* and *Shellback*. …



Deming Ye, Yankai Lin et al. Multi-Paragraph Reasoning with Knowledge-enhanced Graph Neural Network. arXiv:1911.02170

KGNN (Experiments)

Model	Setting	Ans		Sup Fact		Joint	
		EM	F1	EM	F1	EM	F1
Yang et al. (2018)	distractor	45.60	59.02	20.32	64.49	10.83	40.16
KGNN	distractor	**50.81**	**65.75**	**38.74**	**76.79**	**22.40**	**52.82**
Yang et al. (2018)	full wiki	23.95	32.89	3.86	37.71	1.85	16.15
KGNN	full wiki	**27.65**	**37.19**	**12.65**	**47.19**	**7.03**	**24.66**

Table 1: Results on HotpotQA test set for *distractor* and *full wiki* settings.

Model	EM	F1
Yang et al. (2018)-split	11.87	41.87
Yang et al. (2018)	18.14	50.72
KGNN (#Layer=1)	22.26	53.50
KGNN (#Layer=2)	**22.41**	**54.05**
KGNN (#Layer=3)	22.24	53.49

Table 2: Effect of Layer Number on joint metrics.

Model	10	20	30
Yang et al. (2018)	20.77	20.88	21.52
KGNN	**22.20**	**25.02**	**25.12**

Table 3: Effect of paragraph number on joint F1.

Deming Ye, Yankai Lin et al. Multi-Paragraph Reasoning with Knowledge-enhanced Graph Neural Network. arXiv:1911.02170

GSN (Introduction)

背景：

现有的对话问答生成模型假定话语是按顺序组织的

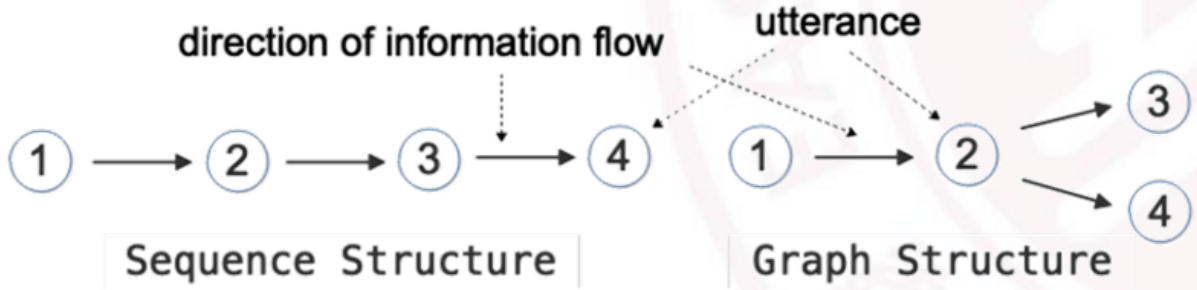
数据集：

Ubuntu Dialogue Corpus

任务：

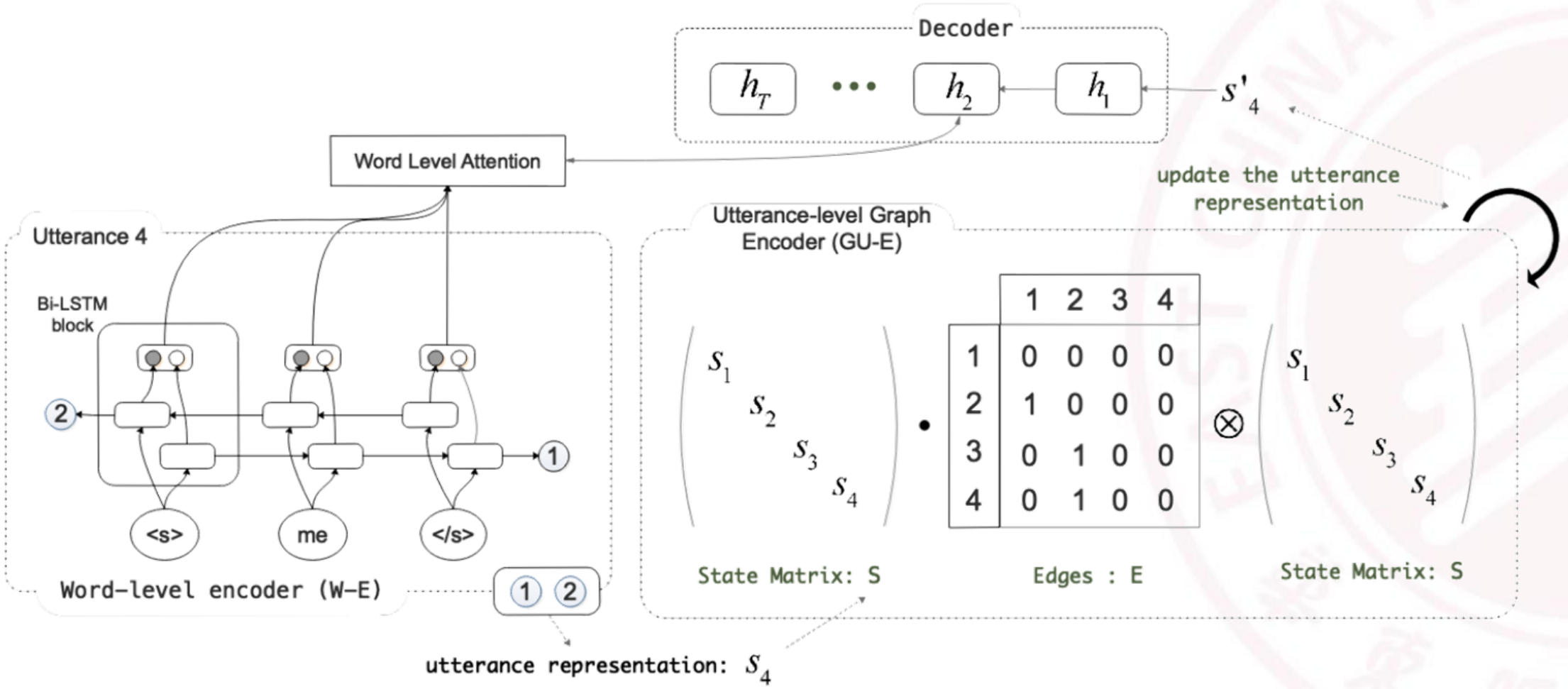
在多轮对话后生成下一句话

utterance 1 (p_1): When the screen goes blank and won't display any login page.
utterance 2 (p_2): I don't know if its a hardware problem or an os.
utterance 3 (p_1): Did you do any upgrade recently?
utterance 4 (p_3): If it works for one user it's probably not a hardware issue.



Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

GSN (Model architecture)



Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

GSN (Word-level Encoder)

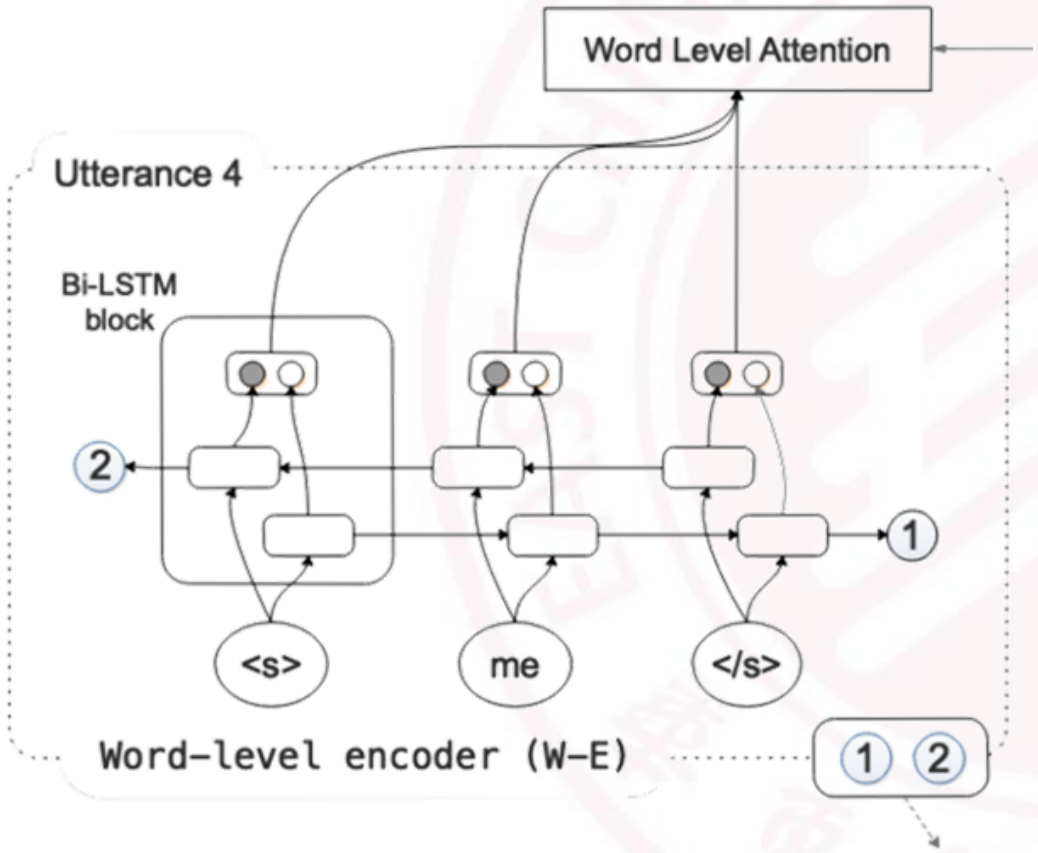
Input: $i = (w_{i,1}, w_{i,2}, \ldots, w_{i,n})$ 代表一个句子

Bi-LSTM:

$$\overrightarrow{\mathbf{s}_{i,t}} = \overrightarrow{LSTM}(\mathbf{e\_w}_{i,t}, \overrightarrow{\mathbf{s}_{i,t-1}})$$
$$\overleftarrow{\mathbf{s}_{i,t}} = \overleftarrow{LSTM}(\mathbf{e\_w}_{i,t}, \overleftarrow{\mathbf{s}_{i,t-1}})$$

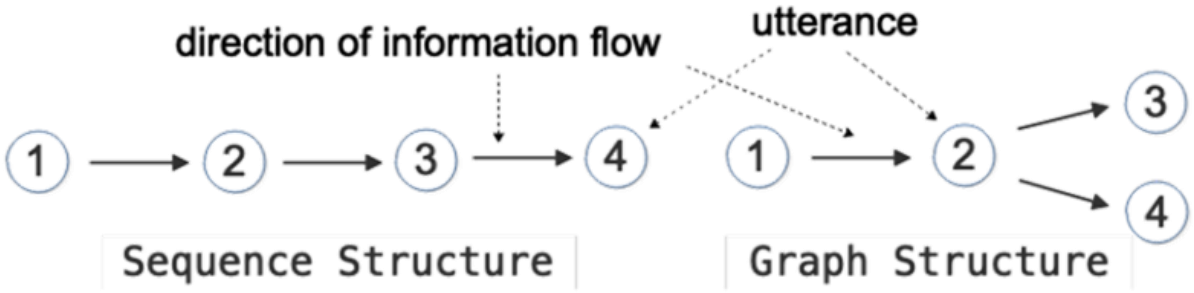
$$S_i = [\overrightarrow{S_{i,n}} ; \overleftarrow{S_{i,n}}]$$

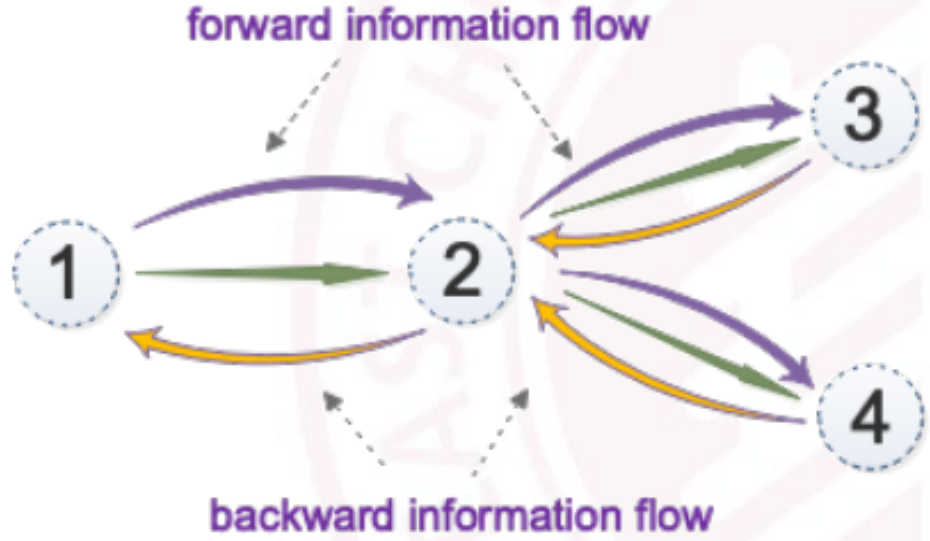
Output: $S = \{s_i, i \in \{1, \ldots, m\}\}$



Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

GSN (Utterance-level Graph-Structured Encoder)





(a) Bi-directional information flow.

- 每个节点只接收之前邻接的几个节点中的信息
- 为了让后续节点辅助前面节点生成增加反向边

Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

GSN (Utterance-level Graph-Structured Encoder)

Input：$S = \{s_i, i \in \{1, \ldots, m\}\}$

$$\mathbf{s}_i^l = \mathbf{s}_i^{l-1} + \eta \cdot \Delta\mathbf{s}_{I|i}^{l-1}$$

$$\Delta\mathbf{s}_{I|i}^{l-1} = \sum_{i' \in \varphi} \Delta\mathbf{s}_{i'|i}^{l-1}$$

$$\Delta\mathbf{s}_{i'|i}^{l-1} = \mathbf{s}_{i'}^{l-1} \otimes \mathbf{s}_i^{l-1}$$

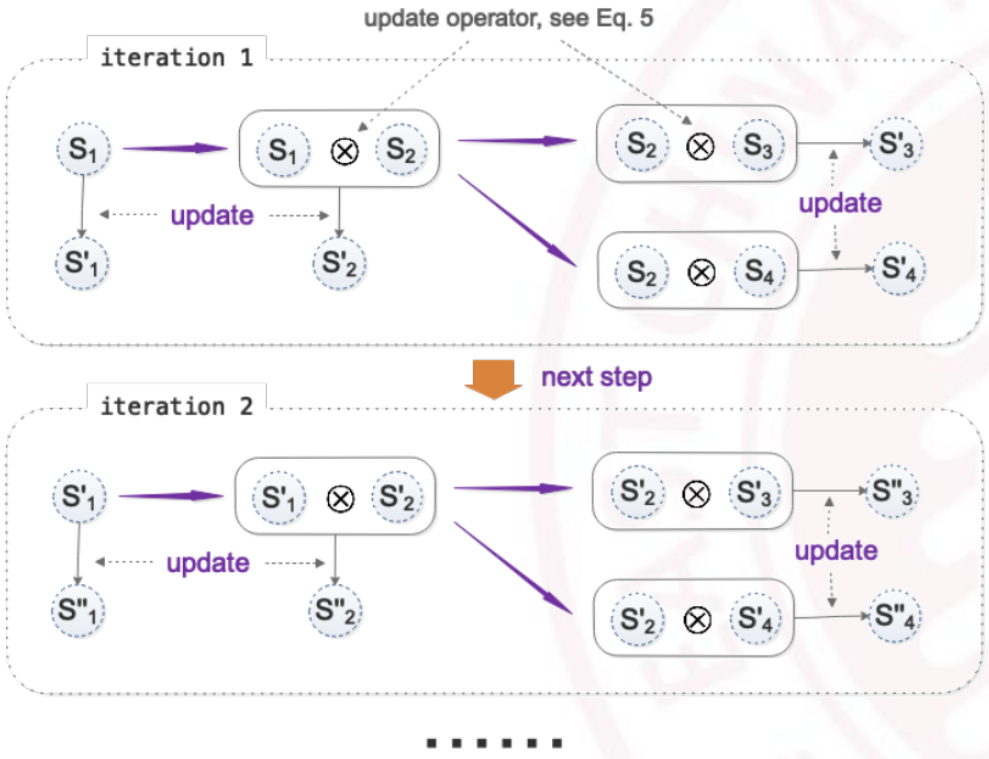
$$\eta = \mathrm{SQH}(\Delta\mathbf{s}_{I|i}^{l-1}) = \frac{\alpha + ||\Delta\mathbf{s}_{I|i}^{l-1}||}{1 + ||\Delta\mathbf{s}_{I|i}^{l-1}||}$$

$$\Delta\mathbf{s}_{i'|i}^{l-1} = (1 - \mathbf{x}_i) * \mathbf{s}_{i'}^{l-1} + \mathbf{x}_i * \tilde{\mathbf{h}}_i$$

$$\tilde{\mathbf{h}}_i = \tanh(\mathbf{W} \cdot [\mathbf{r}_i * \mathbf{s}_{i'}^{l-1}, \mathbf{s}_i^{l-1}])$$

$$\mathbf{x}_i = \sigma(\mathbf{W}_x \cdot [\mathbf{s}_{i'}^{l-1}, \mathbf{s}_i^{l-1}]$$

$$\mathbf{r}_i = \sigma(\mathbf{W}_r \cdot [\mathbf{s}_{i'}^{l-1}, \mathbf{s}_i^{l-1}]$$



φ：当前节点在信息流方向上的前一节点的集合

η：指示有多少信息会从 φ 中融合而来

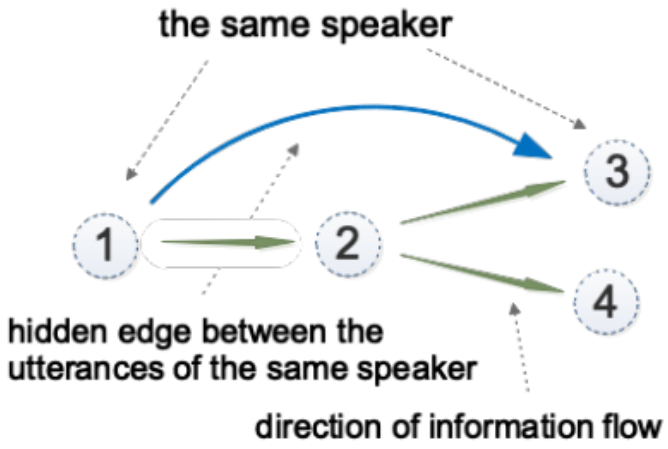
Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

GSN (Utterance-level Graph-Structured Encoder)

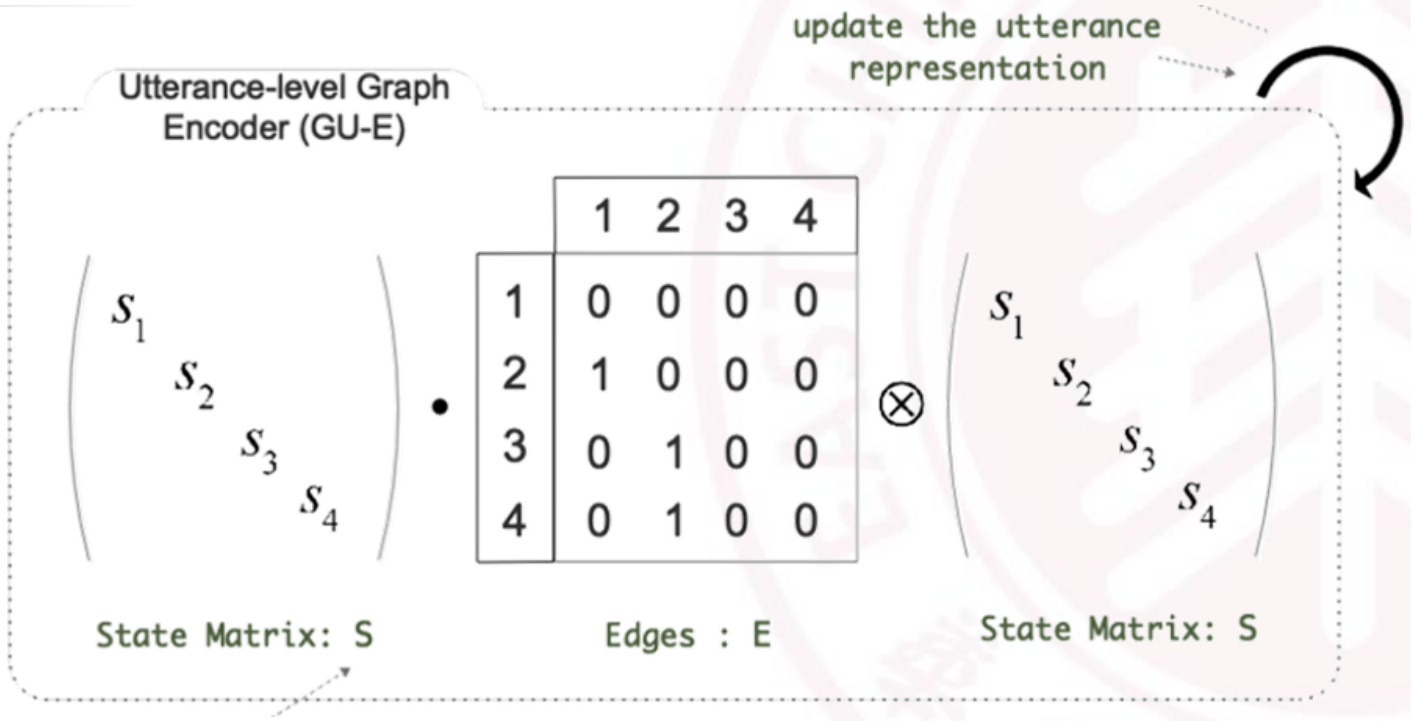
$$\Delta \mathbf{s}'^{l-1}_{i'|i} = \mathbf{s}^{l-1}_{i'} \circledast \mathbf{s}^{l-1}_{i}$$

$$\mathbf{s}^{l}_{i} = \mathbf{s}^{l-1}_{i} + \eta \cdot \Delta \mathbf{s}^{l-1}_{I|i} + \lambda \cdot \Delta \mathbf{s}'^{l-1}_{I|i}$$

$$\Delta \mathbf{s}'^{l-1}_{I|i} = \sum_{i' \in \varphi} \Delta \mathbf{s}'^{l-1}_{i'|i}$$

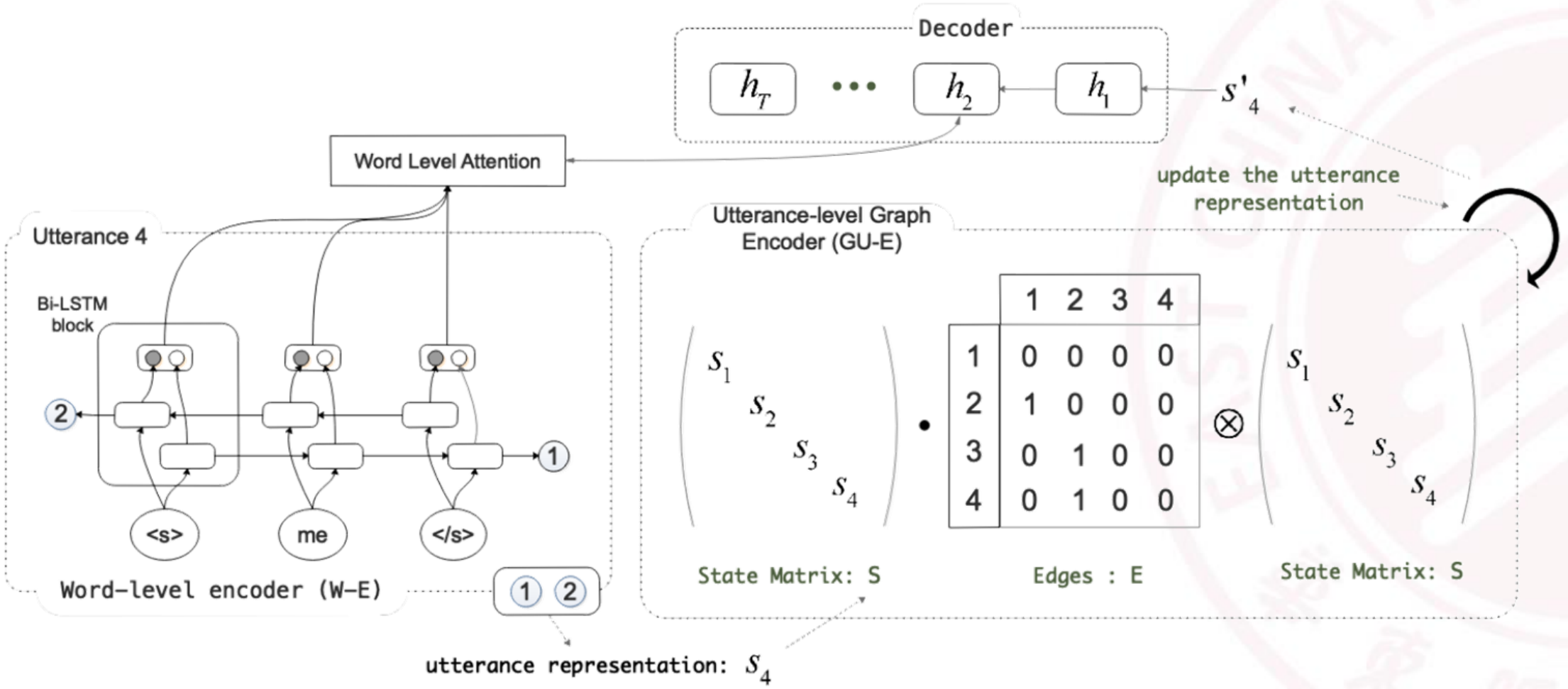


(b) Speaker information modeling.



Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

GSN (Decoder)



Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

GSN (Experiments)

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	ROUGE$_L$
seq2seq	10.45	4.13	2.08	1.02	3.43	9.67
seq2seq W-speaker	10.70	4.98	2.20	1.55	3.92	9.42
Seq2seq (last utte)	9.85	3.04	1.38	0.67	3.98	8.34
HRED [Serban *et al.*, 2016]	10.80	4.60	2.54	1.42	4.38	10.23
HRED W-speaker	11.23	4.82	3.06	1.64	4.36	10.98
GSN No-speaker (1-iter)	9.42	3.05	1.61	0.95	3.74	7.63
GSN No-speaker (2-iter)	12.06	4.87	2.80	1.70	4.32	10.09
GSN No-speaker (3-iter)	12.77▲	5.37▲	3.17	1.99▲	4.53	10.80
GSN W-speaker (1-iter)	10.31	4.06	2.34	1.45	3.88	9.96
GSN W-speaker (2-iter)	12.77	4.93	2.61	1.46	4.79	11.34
GSN W-speaker (3-iter)	**13.50**▲	**5.63**▲	**3.24**▲	**1.99**▲	**4.85**▲	**11.36**▲

Table 2: Experimental results, conducted in different settings, including sequential data and graph data using different models based on automated evaluation. 'Seq2seq (last utte)' is trained by using only the last utterance before the final response of the session as the input (all utterances before are ignored). 'n-iter' means that the results are obtained after n iterations. 'No-speaker' is our proposed GSN model without speaker information flow while 'W-speaker' has it. ▲denotes the p-value < 0.01 in paired t-test against the best baseline (shaded row).

Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

GSN (Experiments)

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	ROUGE$_L$
HRED [Serban *et al.*, 2016] (sequential)	9.61	3.48	1.86	1.01	4.08	8.22
GSN No-speaker (2-iter sequential)	11.39	4.55	2.68	1.71	4.40	9.74
GSN W-speaker (1-iter sequential)	8.69	3.1	1.78	1.19	3.67	9.19
GSN W-speaker (2-iter sequential)	12.72	4.84	2.59	1.59	4.70	11.41
GSN W-speaker (3-iter sequential)	12.03	4.92	2.94	1.97	4.31	10.1
HRED [Serban *et al.*, 2016] (graph)	12.16	4.90	2.68	1.49	4.42	10.90
GSN No-speaker (2-iter graph)	12.35	5.17	3.08	1.81	4.43	10.42
GSN W-speaker (1-iter graph)	10.66	4.36	2.52	1.50	3.97	10.10
GSN W-speaker (2-iter graph)	12.76	5.23	2.94	1.75	4.80	11.33
GSN W-speaker (3-iter graph)	**13.85**	**5.83**	**3.33**	**1.98**	**5.10**	**11.66**

Table 3: Experimental results of using *sequential data* (with only 2 interlocutors, the first five rows of the results) or *graph data* only (with more than 2 interlocutors, the last five rows of the results). The symbol string 'n-iter', 'No-speaker' and 'W-speaker' in the table have the same meaning as those in Table 2. The result of GSN No-speaker 3-iter isn't given as it performs worse.

Hu, Wenpeng and Chan et al. GSN: A Graph-Structured Network for Multi-Party Dialogues. arXiv:1905.13637

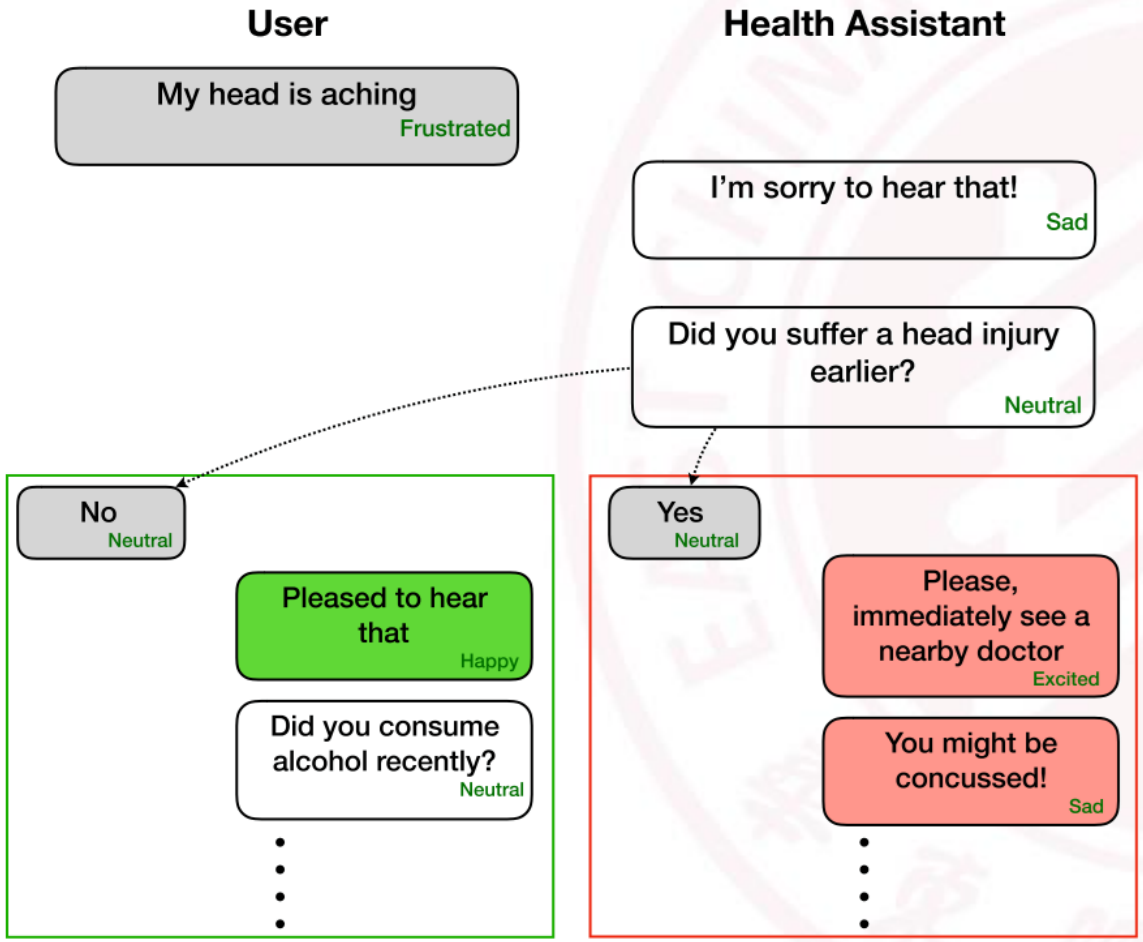
DialogueGCN (Introduction)

背景：

与 GSN 相似，现有的模型（基于 RNNs）难以捕获对话中的 turn 信号

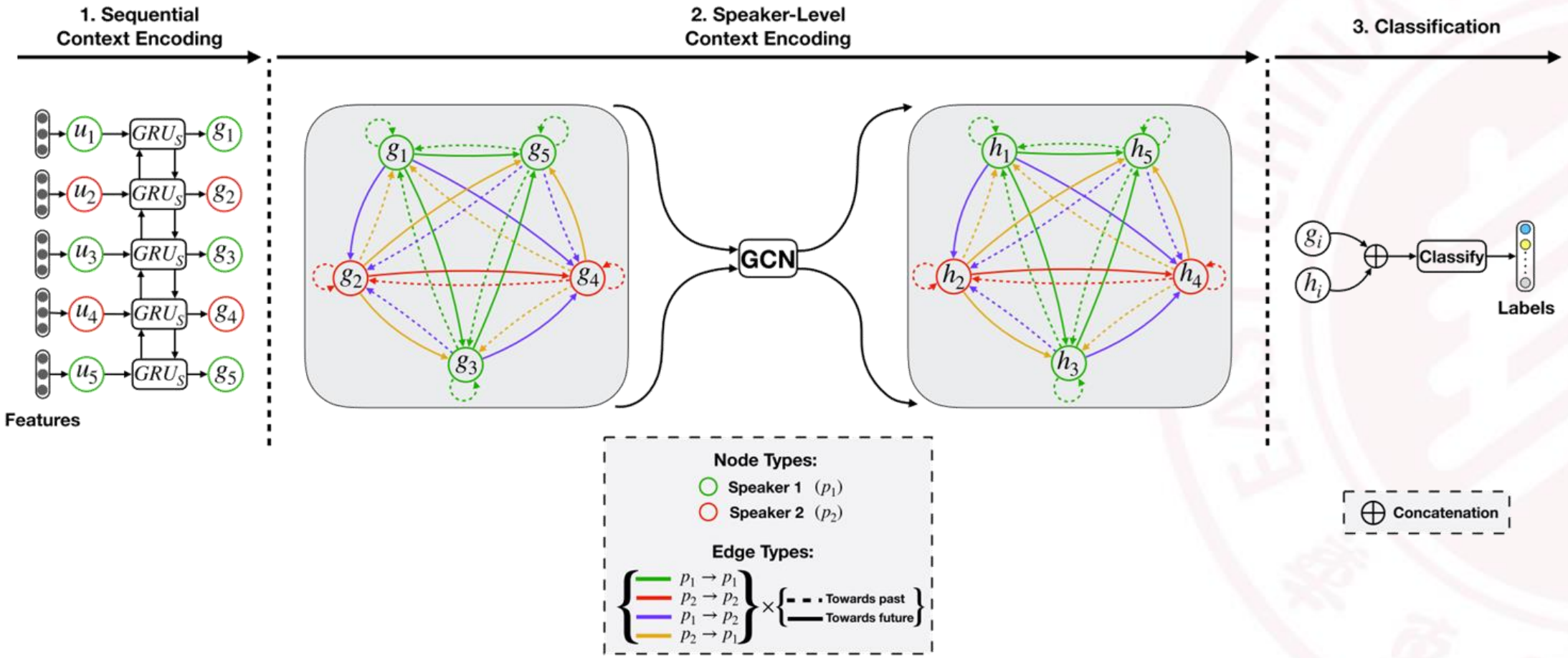
任务：

预测多人多轮对话中每一句话的情绪标签（快乐、悲伤等）



Deepanway Ghosal, Navonil Majumder et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. arXiv:1908.11540

DialogueGCN (Model architecture)



Deepanway Ghosal, Navonil Majumder et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. arXiv:1908.11540

DialogueGCN (Sequential Context Encoder)

顺序上下文语境：

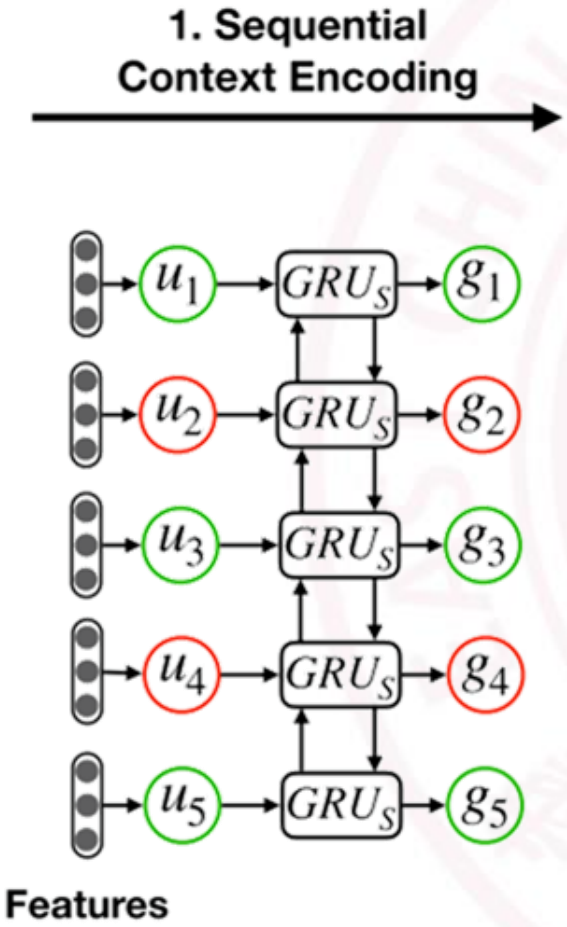
- 处理过去的句子如何影响未来
- 处理句子间语义/句法的特征

Input：u_i代表第 i 句话

Bi-GRU:

$$g_i = \overleftrightarrow{GRU_S}(g_{i(+,-)1}, u_i)$$

Output：g_i代表带上下文感知的话语表示



Deepanway Ghosal, Navonil Majumder et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. arXiv:1908.11540

DialogueGCN (Speaker-Level Context Encoder)

对话者级别语境：

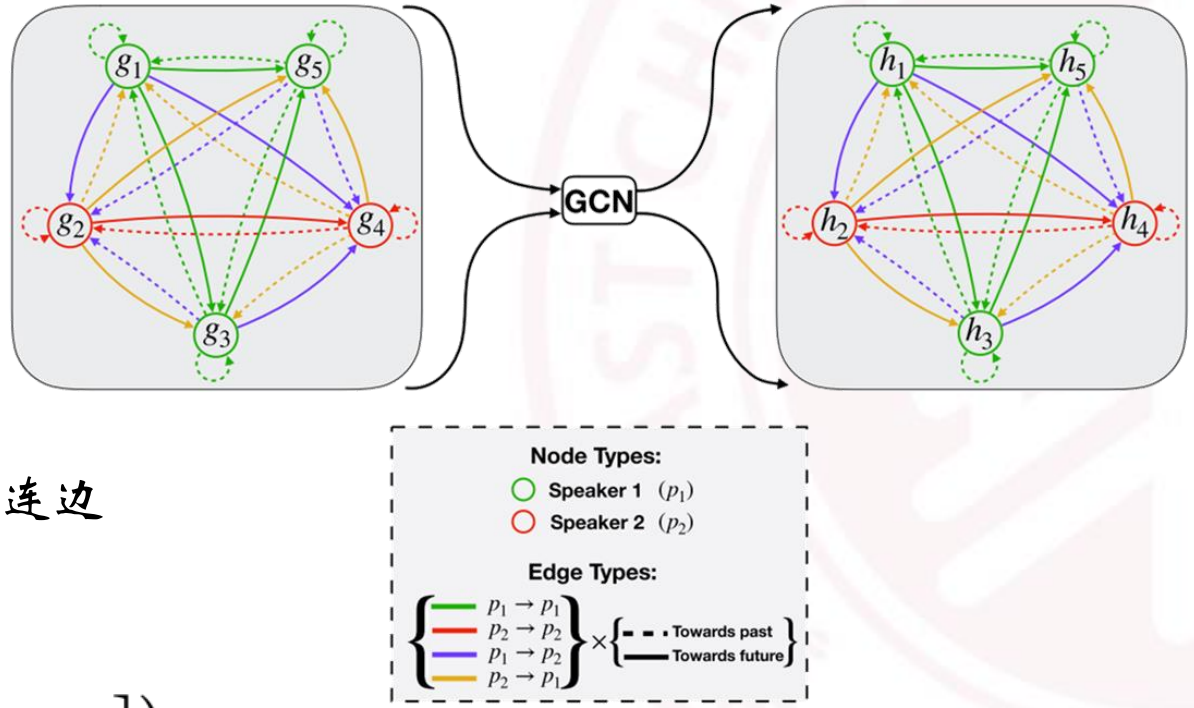
- inter-dependency：外界影响
- intra-dependency：自身影响

Input：g_i代表带上下文感知的话语表示

建图（每句话作为节点）：

- 每个节点往过去 p 个，未来 f 个节点连边
- 每个节点自环

边权：$\alpha_{ij} = \mathrm{softmax}(g_i^T W_e [g_{i-p}, \ldots, g_{i+f}])$



Deepanway Ghosal, Navonil Majumder et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. arXiv:1908.11540

DialogueGCN (Experiments)

Methods	IEMOCAP													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		**Average(w)**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
Memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
bc-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
bc-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13
ICON	22.22	29.91	58.78	64.57	62.76	57.38	64.71	63.04	58.86	63.42	67.19	60.81	59.09	58.54
DialogueRNN	25.69	33.18	75.10	78.80	58.59	59.21	64.71	**65.28**	80.27	**71.86**	61.15	58.91	63.40	62.75
DialogueGCN	40.62	**42.75**	89.14	**84.54**	61.92	**63.54**	67.53	64.19	65.46	63.08	64.18	**66.99**	65.25	**64.18**

Table 3: Comparison with the baseline methods on IEMOCAP dataset; Acc. = Accuracy; bold font denotes the best performances. Average(w) = Weighted average.

Deepanway Ghosal, Navonil Majumder et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. arXiv:1908.11540

Acknowledgement

Thanks!