SemSUM: Semantic Dependency Guided Neural Abstractive Summarization

Hanqi Jin,[1,2,3]* **Tianming Wang,**[1,3]* **Xiaojun Wan**[1,2,3]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Center for Data Science, Peking University
[3]The MOE Key Laboratory of Computational Linguistics, Peking University
{jinhanqi, wangtm, wanxiaojun}@pku.edu.cn

Task

Source	even some of his rivals in the roland garros locker-room are hoping that roger federer can create a bit of tennis history by winning all four grand slams .
Target	even fellow players keen for federer grand slam dream picture
Seq2seq	federer hoping to win at roland garros
SemSUM	even some rivals hope federer to make grand slam history

Table 1: Example summaries of a sentence with and without semantics guidance.

Task

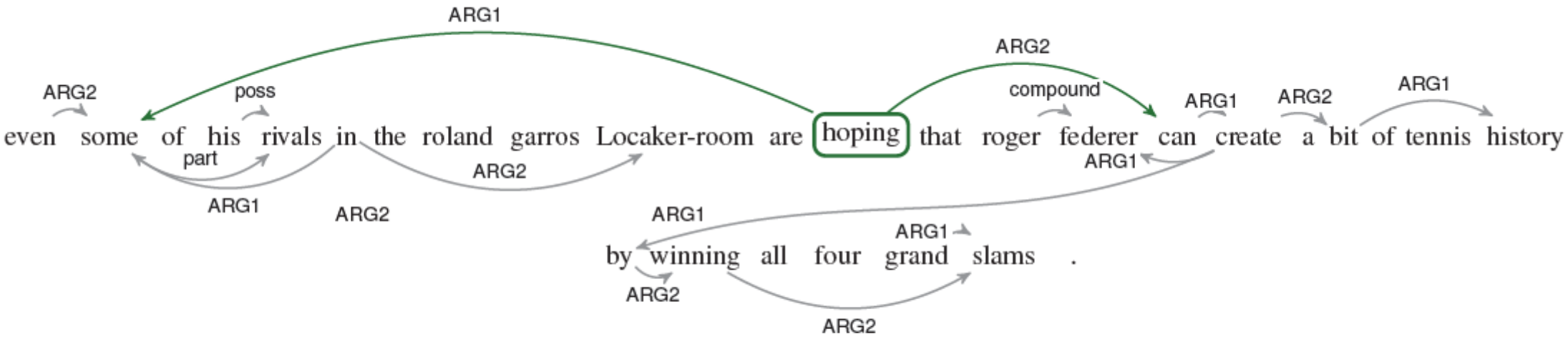


Figure 1: An example sentence annotated with a semantic dependency graphs. The green color represents the dependency of root node “hoping”. Some dependency edges are omitted for display.

Model

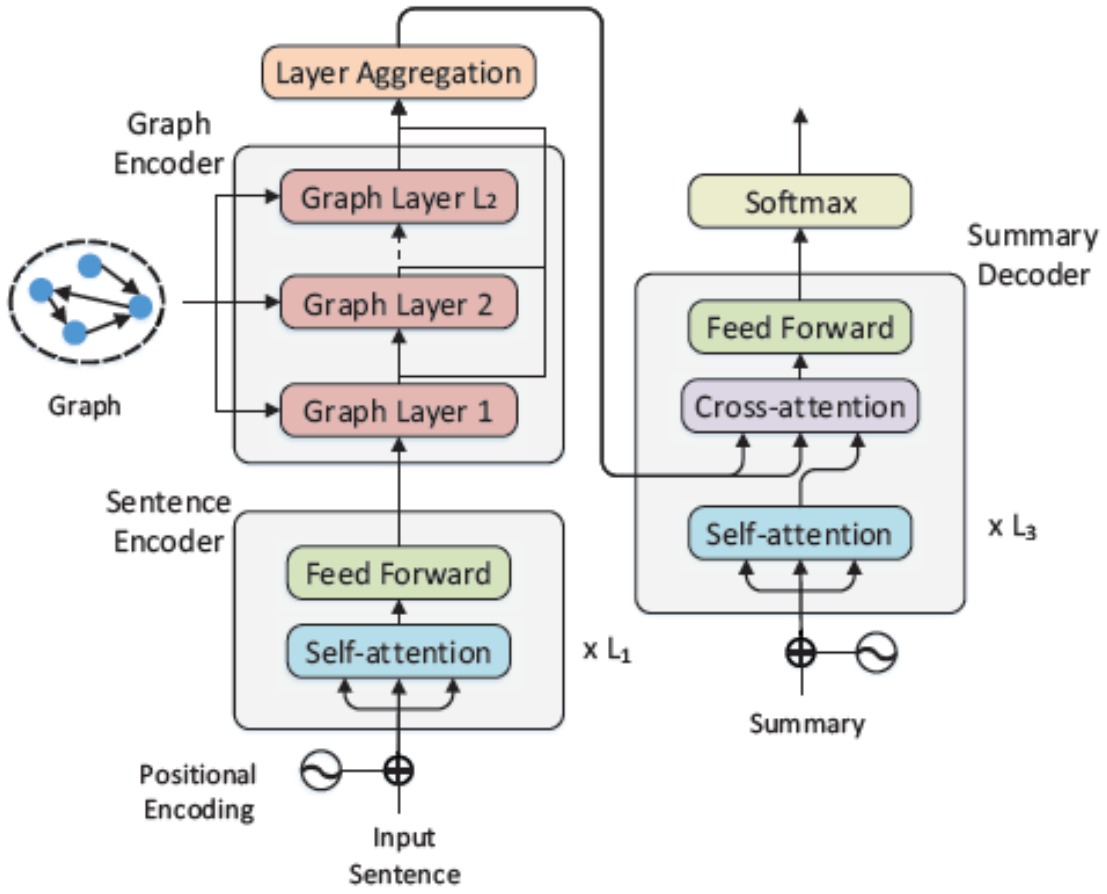


Figure 2: The overview of our SemSUM model

Graph Encoder

- **Graph attention mechanism**

$$u_{out} = \mathrm{ReLU}([g_{head} \parallel e_{type} \parallel g_{tail}]W_{out} + b_{out})$$
$$u_{in} = \mathrm{ReLU}([g_{head} \parallel e_{type} \parallel g_{tail}]W_{in} + b_{in})$$

$$\hat{g}_v = \sum_{u \in \mathcal{N}(v)} \alpha(u, g_v) u W^V$$

$$\alpha(u, g_v) = \frac{\exp\left(\left(uW^K\right)^\top g_v W^Q\right)}{\sum_{z \in \mathcal{N}(v)} \exp\left(\left(zW^K\right)^\top g_v W^Q\right)}$$

$$\mathrm{MHGAT}(g_v) = \left(\Big\Vert_{j=1}^{H} \hat{g}_v^j\right) W^O$$

- **Layer aggregation**
  - we adopt a bidirectional LSTM to aggregate the outputs of all graph layers:

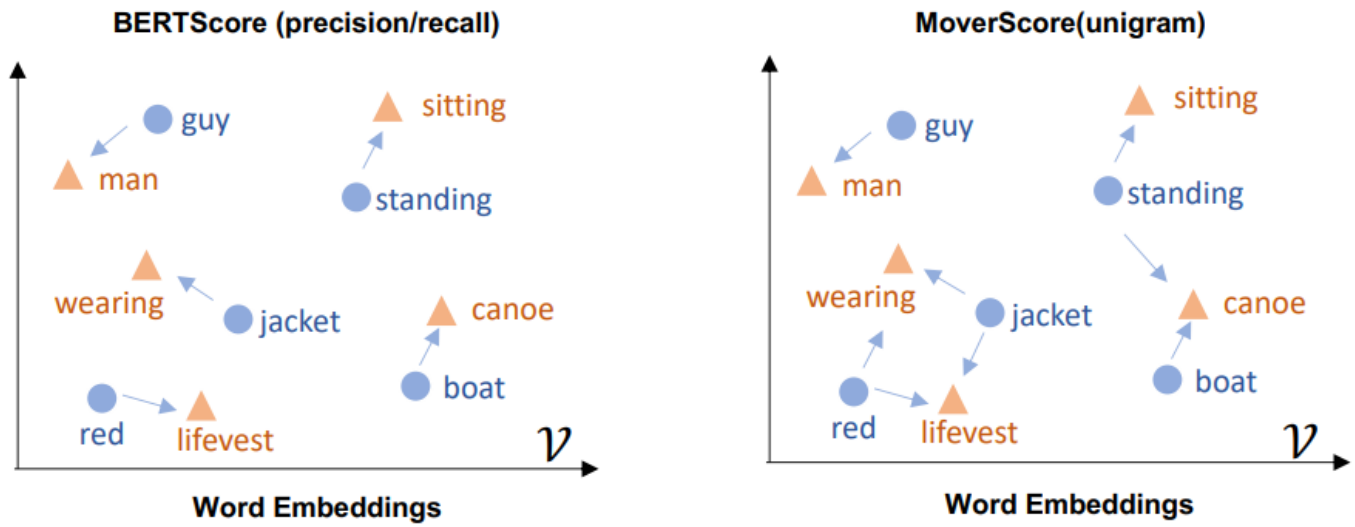
$$o = \overleftarrow{h}_1 + \overrightarrow{h}_N$$

Dataset

- **Gigaword**
  - 3.8M sentence-summary pairs for training and 189K pairs for development
  - 1951 sentence-summary pairs for test
- **DUC2004**
  - 500 input sentence with each sentence paired with 4 different human-written reference summaries for test.
- **MSR-ATC**
  - 785 input sentences with each sentence paired with 3-5 summaries for test

Metrics

- **Rough**
- **BERTScore**
- **MoverScore**



System x: A **guy** with a **red jacket** is **standing** on a **boat**

Ref y: A **man wearing** a **lifevest** is **sitting** in a **canoe**

Figure 1: An illustration of MoverScore and BERTScore.

Result

Model	RG-1	RG-2	RG-L	WMD	BERT
ABS(Rush, Chopra, and Weston 2015)	29.55	11.32	26.42	-	-
SEASS(Zhou et al. 2017)	36.15	17.54	33.63	-	-
Re3Sum(Cao et al. 2018a)	37.04	19.03	34.46	-	-
FTSum(Cao et al. 2018b)	37.27	17.65	34.24	-	-
PostEnsemble(Kobayashi 2018)	37.52	18.55	34.86	-	-
Sun-Attention(Niu et al. 2019)	38.27	16.45	36.08	-	-
MASS(Song et al. 2019)	38.73	19.71	35.96	34.28	**61.56**
BiSET(Wang, Quan, and Wang 2019b)	**39.11**	**19.78**	**36.87**	33.79	61.24
Transformer	36.69	18.08	34.22	32.50	60.48
TFM&GCN	37.51	19.03	34.89	33.67	61.02
SemSUM	38.78	**19.75**	36.09	**34.39**	**61.56**

Table 2: ROUGE F1, WMD unigram and BERTScore F1 evaluation results on the Gigaword test set.

Model	RG-1	RG-2	RG-L	WMD	BERT
ABS(Rush, Chopra, and Weston 2015)	26.55	7.06	22.05	-	-
SEASS(Zhou et al. 2017)	29.21	9.56	25.51	-	-
ERAML(Li et al. 2018)	29.33	10.24	25.24	-	-
WACNNs(Yuan et al. 2019)	30.54	10.87	26.94	-	-
Transformer	29.78	9.61	25.85	22.95	56.36
TFM&GCN	30.24	10.44	26.32	24.80	57.21
SemSUM	**31.00**	**11.11**	**26.94**	**26.71**	**57.99**

Table 3: ROUGE recall, WMD unigram and BERTScore F1 evaluation results on the DUC2004 test set.

Model	RG-1	RG-2	RG-L	WMD	BERT
ABS(Rush, Chopra, and Weston 2015)	20.27	5.26	17.10	-	-
SEASS(Zhou et al. 2017)	25.75	10.63	22.90	-	-
Transformer	29.29	12.45	25.93	13.76	54.31
TFM&GCN	32.53	15.41	29.26	15.61	55.48
SemSUM	**33.82**	**17.08**	**30.62**	**17.14**	**56.19**

Table 4: ROUGE F1, WMD unigram and BERTScore F1 evaluation results on the MSR-ATC test set.

Result

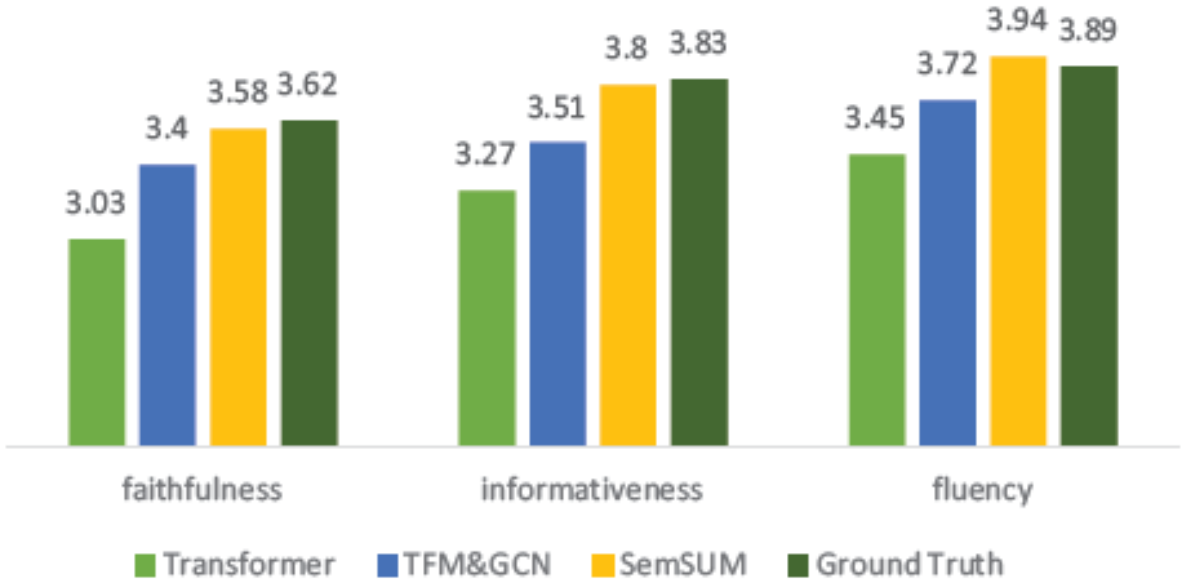


Figure 4: Human evaluation. They are rated on a Likert scale of 1(worst) to 5(best).

Ablation study

Model	RG-1	RG-2	RG-L
SemSUM	48.25	26.54	45.14
only sentence encoder	47.14	24.94	44.00
only graph encoder	47.28	25.21	44.17
only graph encoder*	47.50	25.51	44.39
without layer aggregation	48.02	25.83	44.82
without separated mappings	47.82	25.57	44.59

Table 5: ROUGE F1 evaluation results on the development set of ablation study. $\star$ denotes taking the adjacent relation as an edge type.

Case study

Source	german parliament called on the international olympic committee on thursday to do more for women in sport .
Target	olympics told to help women
Transformer	german parliament calls for more women in sport
SemSUM	german parliament urges ioc to do more for women
Source	democrats in georgia and alabama , borrowing an idea usually advanced by conservative republicans , are promoting bible classes in the public schools .
Target	democrats in southern states push bills on bible study
Transformer	bible classes in public schools are promoting bible classes
SemSUM	democrats promote bible classes in public schools

Table 6: Case Study.

Encoding Syntactic Constituency Paths for Frame-Semantic Parsing with Graph Convolutional Networks

Emanuele Bastianelli, Andrea Vanzo, and Oliver Lemon
Interaction Lab, MACS, Heriot-Watt University, Edinburgh, UK
{e.bastianelli, a.vanzo, o.lemon}@hw.ac.uk

Task

- **Target identification**
- **Frame identification**
- **Argument identification**
- **Argument classification**

Task

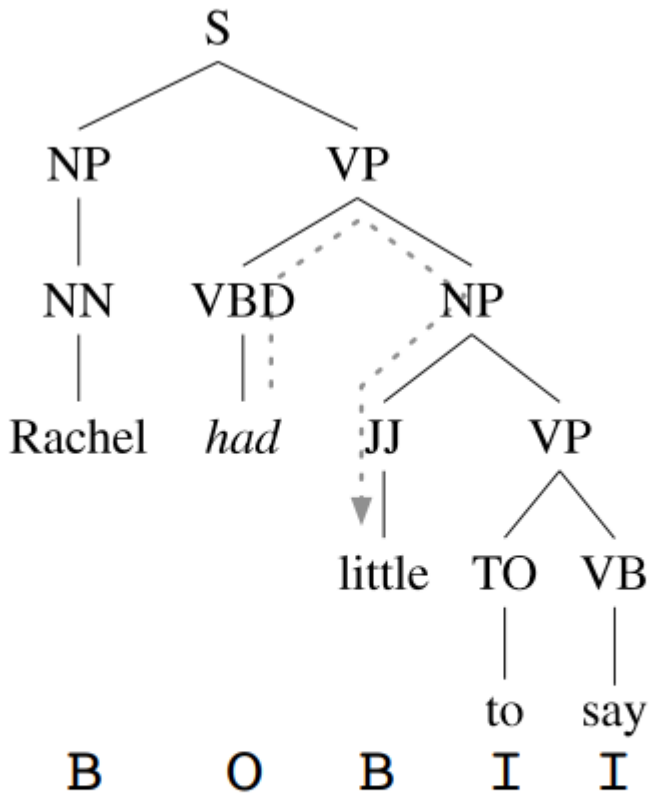


Figure 1: Constituency path between the target predicate *had* and the word *little*. At the bottom, the IOB tagging of the corresponding argument spans.

Encoder

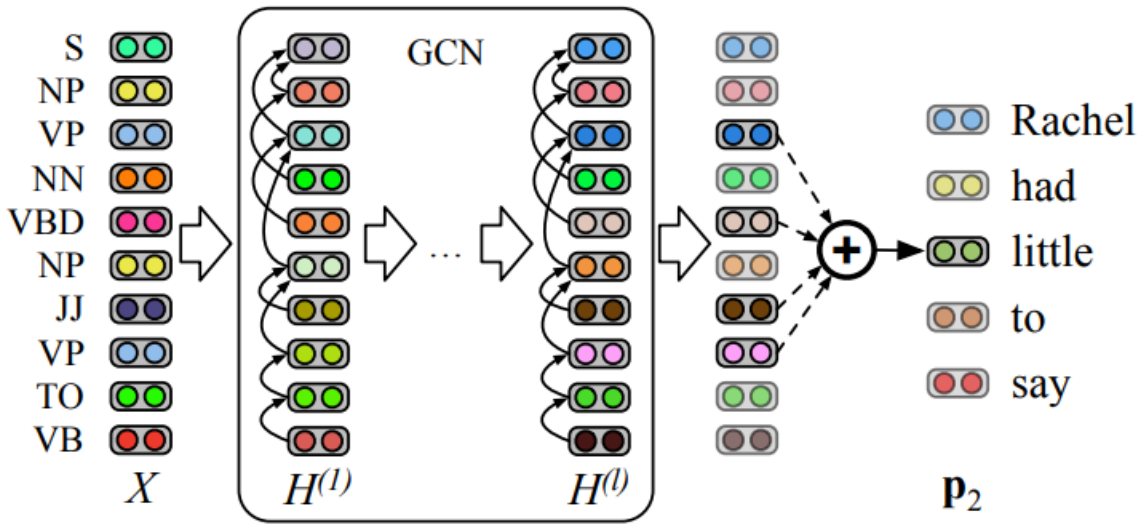
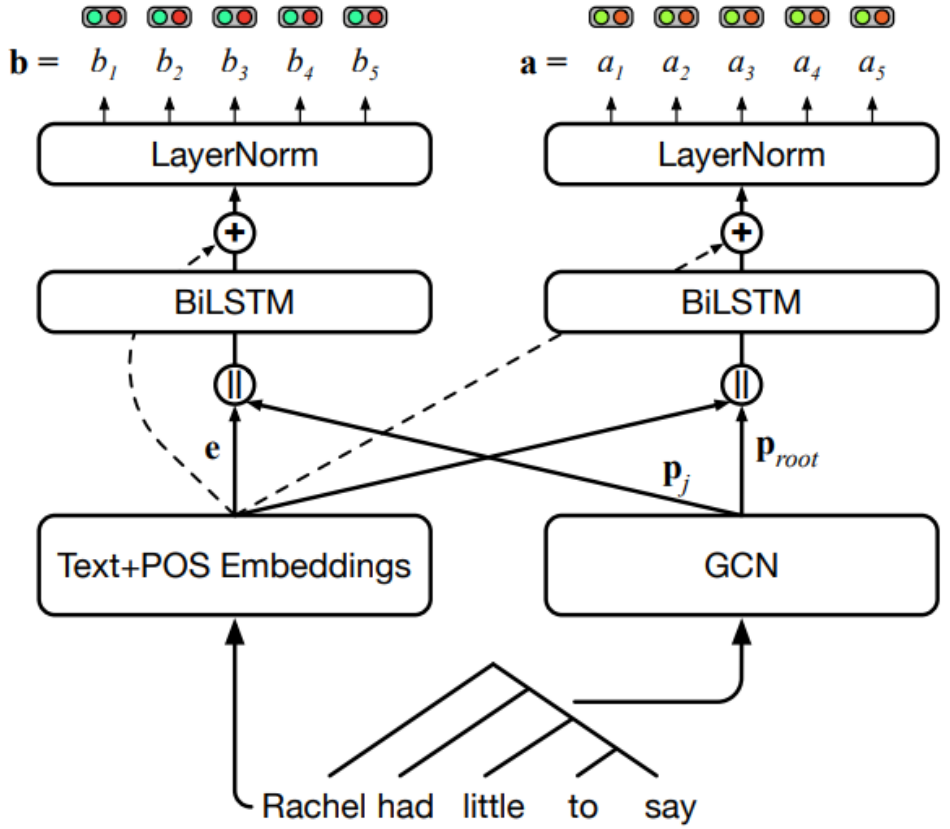


Figure 2: Process of learning constituent encodings and evaluating path features $\mathbf{p}_2$ from the node *had* over the tree in Figure 1. Detail of feature evaluation for the word *little*.



Decoder

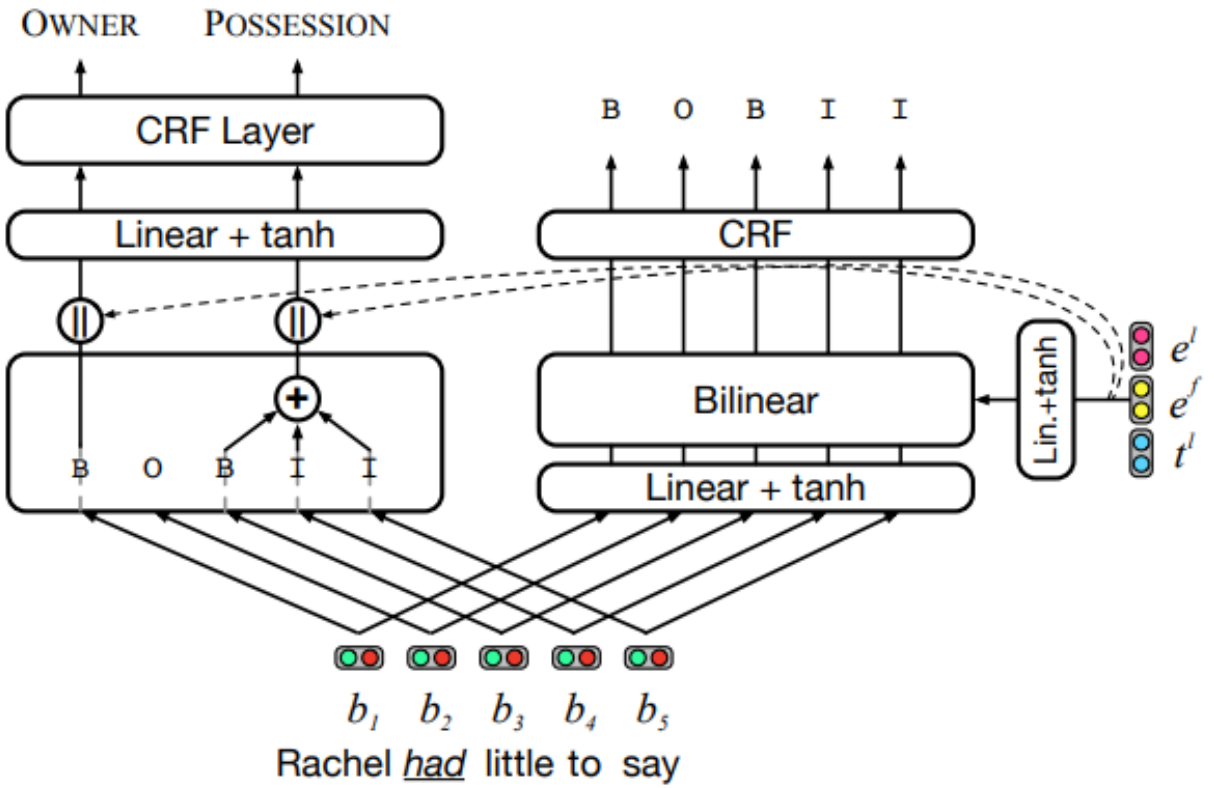


Figure 4: AI and AC sub-networks on the right and on the left, respectively.

Model

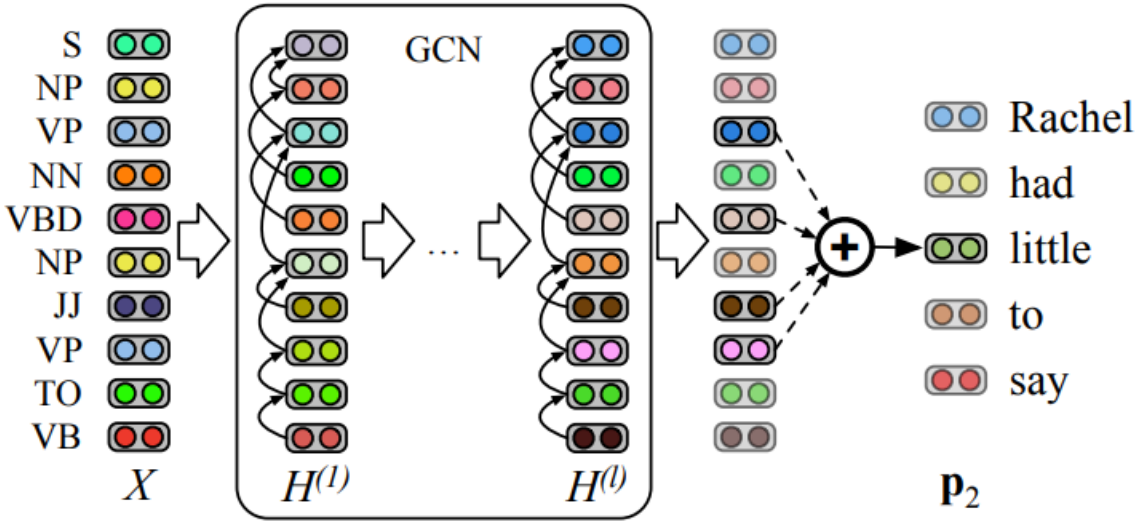
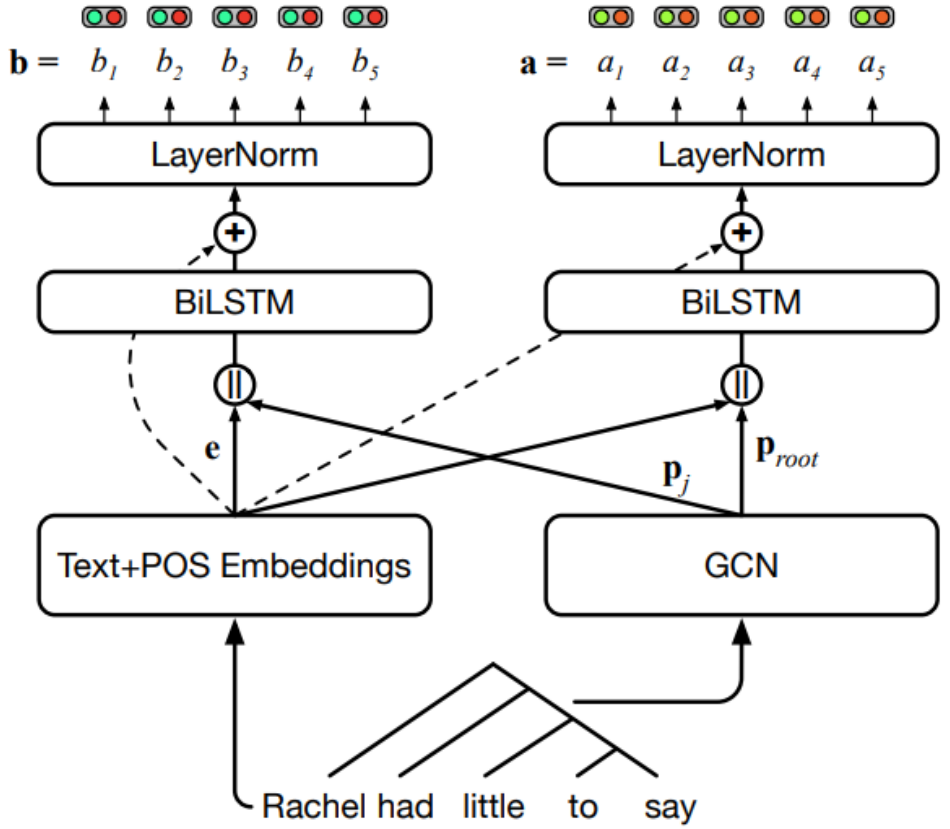


Figure 2: Process of learning constituent encodings and evaluating path features $\mathbf{p}_2$ from the node *had* over the tree in Figure 1. Detail of feature evaluation for the word *little*.



Result

Model	P	R	F1
Das et al. (2014)	37.5	57.5	45.4
Swayamdipta et al. (2017)	–	–	73.2
OURS-ELMo	69.84	*78.88*	74.08
OURS-ELMo-GCN	*70.08*	78.72	*74.15*
OURS-BERT	71.71	78.51	74.96
OURS-BERT-GCN	70.66	**84.12**	**76.80**
OURS-BERT-GCN-JL	**72.59**	79.43	75.86

Table 1: Precision, Recall and F1 of the Target Identification task.

Model	Acc
Das et al. (2014)	83.6
Hermann et al. (2014)	88.4
Hartmann et al. (2017)	87.6
Yang and Mitchell (2017)	88.2
Peng et al. (2018)	89.9
OURS-ELMo	88.89
OURS-ELMo-GCN	88.82
OURS-BERT	89.90
OURS-BERT-GCN	89.83
OURS-BERT-GCN-JL	**90.10**

Table 2: Frame Identification results using gold targets, in terms of Accuracy.

Model	P	R	F1
Das et al. (2014)†	65.6	53.8	59.1
Kshirsagar et al. (2015)†	66.0	60.4	63.1
Yang and Mitchell (2017)†	70.2	60.2	65.5
Swayamdipta et al. (2018)	69.2	69.0	69.1
Marcheggiani and Titov (2019)	69.8	68.8	69.3
OURS-ELMo	63.89	67.36	65.58
OURS-ELMo-GCN	*72.02*	*73.70*	*72.85*
OURS-BERT	71.19	74.26	72.69
OURS-BERT-GCN	74.23	**76.94**	**75.56**
OURS-BERT-GCN-JL	**74.56**	74.43	74.50

Table 3: Semantic Role Labeling results with gold targets and frames.

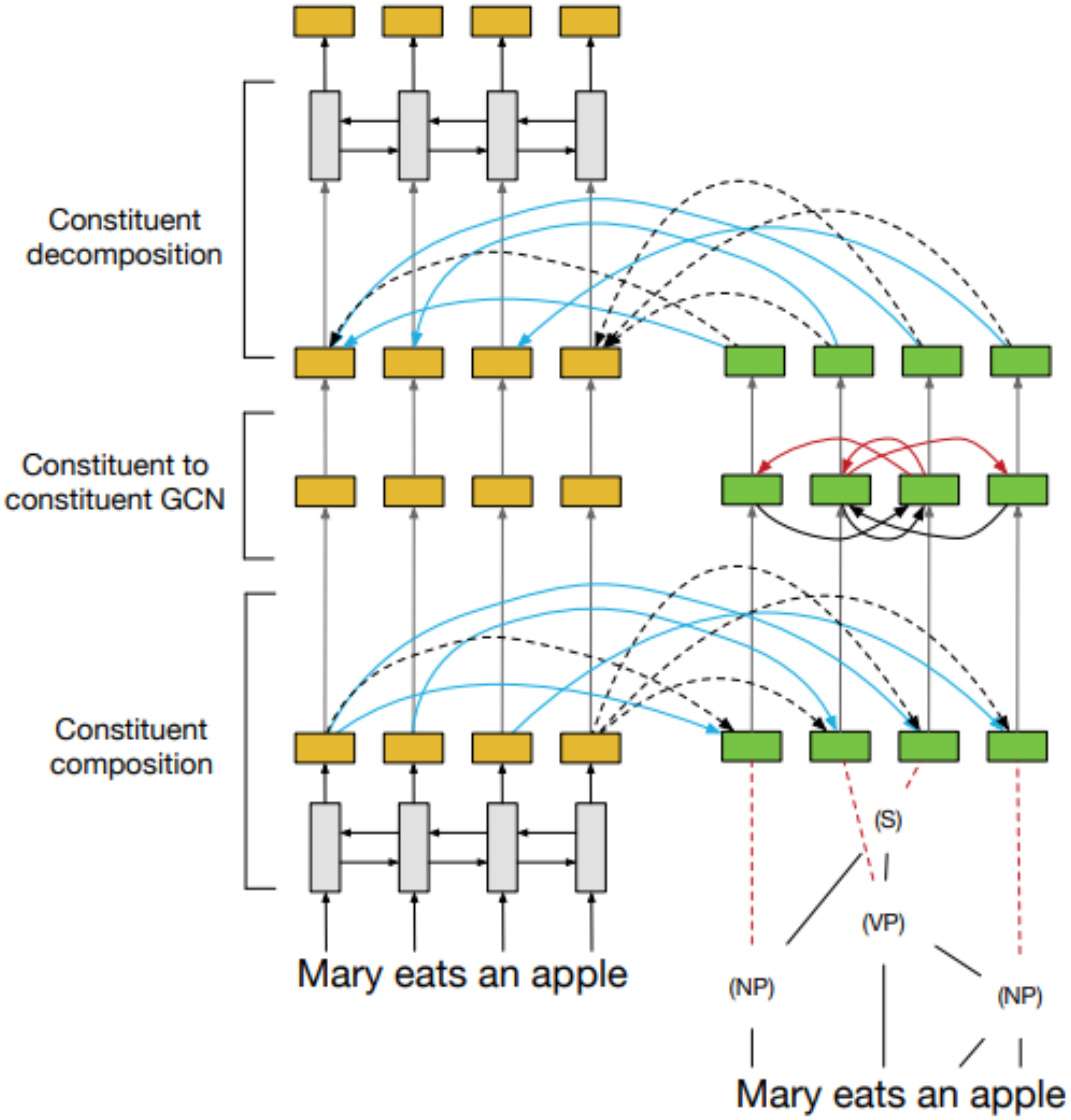
	SRL		
Input	No-GCN	GCN	gain
GloVe	49.13	68.23	+19.1
ELMo	65.58	72.85	+7.27
BERT	72.69	75.56	+2.87

Table 4: Ablation on the use of constituency features with different input layers for the SRL on FN15.

Graph Convolutions over Constituent Trees for Syntax-Aware Semantic Role Labeling

Diego Marcheggiani[1*] **Ivan Titov**[2,3]
[1]Amazon
[2]ILCC, School of Informatics, University of Edinburgh
[3]ILLC, University of Amsterdam
marchegg@amazon.es ititov@inf.ed.ac.uk

Model


Constituent decomposition
Constituent to constituent GCN
Constituent composition
Mary eats an apple
(S)
(VP)
(NP)
(NP)
Mary eats an apple

Result

Model	P	R	F_1
Yang and Mitchell (2017) (SEQ)	63.4	66.4	64.9
Yang and Mitchell (2017) (ALL)	70.2	60.2	65.5
Swayamdipta et al. (2018)†‡	69.2	69.0	69.1
SpanGCN†	69.8	68.8	69.3

Table 4: Results on FrameNet 1.5 test set using gold frames. † indicates syntactic models and ‡ indicates multi-task learning models.

	WSJ Test		
Single	P	R	F_1
He et al. (2017)	83.1	83.0	83.1
He et al. (2018a)	84.2	83.7	83.9
Tan et al. (2018)	84.5	85.2	84.8
Ouchi et al. (2018)	84.7	82.3	83.5
Strubell et al. (2018)(LISA)†‡	84.7	84.6	84.6
SpanGCN†	85.8	85.1	85.4
Single / Context. Emb.			
He et al. (2018a)(ELMo)	-	-	87.4
Li et al. (2019)(ELMo)	87.9	87.5	87.7
Ouchi et al. (2018)(ELMo)	88.2	87.0	87.6
Wang et al. (2019)(ELMo)†	-	-	88.2
SpanGCN (ELMo)†	87.5	87.9	87.7
SpanGCN (RoBERTa)†	87.7	88.1	87.9

	Brown Test		
Single	P	R	F_1
He et al. (2017)	72.9	71.4	72.1
He et al. (2018a)	74.2	73.1	73.7
Tan et al. (2018)	73.5	74.6	74.1
Ouchi et al. (2018)	76.0	70.4	73.1
Strubell et al. (2018)(LISA)†‡	74.8	74.3	74.6
SpanGCN†	76.2	74.7	75.5
Single / Context. Emb.			
He et al. (2018a)(ELMo)	-	-	80.4
Li et al. (2019)(ELMo)	80.6	80.4	80.5
Ouchi et al. (2018)(ELMo)	79.9	77.5	78.7
Wang et al. (2019)(ELMo)†	-	-	79.3
SpanGCN(ELMo)†	79.4	79.6	79.5
SpanGCN(RoBERTa)†	80.5	80.7	80.6

Table 2: Precision, recall and F_1 on the CoNLL-2005 test sets. † indicates syntactic models and ‡ indicates multi-task learning models.

	Test		
Single	P	R	F_1
He et al. (2017)	81.7	81.6	81.7
He et al. (2018a)	-	-	82.1
Tan et al. (2018)	81.9	83.6	82.7
Ouchi et al. (2018)	84.4	81.7	83.0
Swayamdipta et al. (2018)†‡	85.1	81.2	83.8
SpanGCN†	84.5	84.3	84.4
Single / Context. Emb.			
Peters et al. (2018a)(ELMo)	-	-	84.6
He et al. (2018a)(ELMo)	-	-	85.5
Li et al. (2019)(ELMo)	85.7	86.3	86.0
Ouchi et al. (2018)(ELMo)	87.1	85.3	86.2
Wang et al. (2019)(ELMo)†	-	-	86.4
SpanGCN (ELMo)†	86.3	86.8	86.5
SpanGCN (RoBERTa)†	86.5	87.1	86.8

Table 3: Precision, recall and F_1 on the CoNLL-2012 test set. † indicates syntactic models and ‡ indicates multi-task learning models.