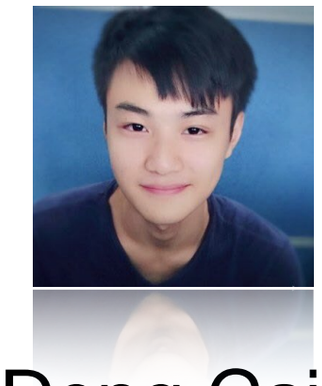
Graph Transformer for Graph-to-Sequence Learning



Deng Cai and Wai Lam

The Chinese University of Hong Kong

AAAI2020

Background

- Graphical structure in natural language processing (NLP)
    - Syntax
    - Semantics
    - Knowledge

Background

- Graphical structure in natural language processing (NLP)
  - **Syntax** **e.g., Dependency Tree**
  - Semantics
  - Knowledge

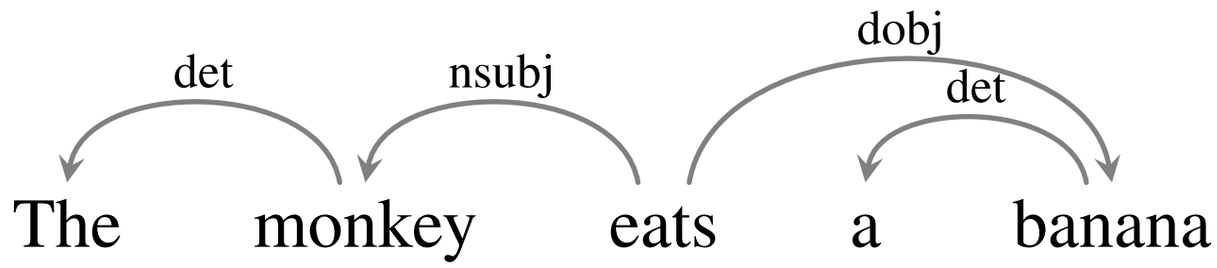
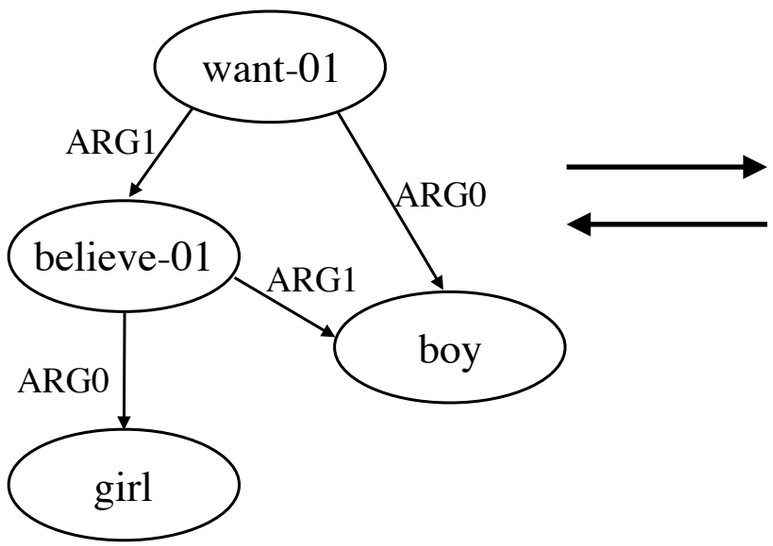


figure from Basting et al, 2017

Background

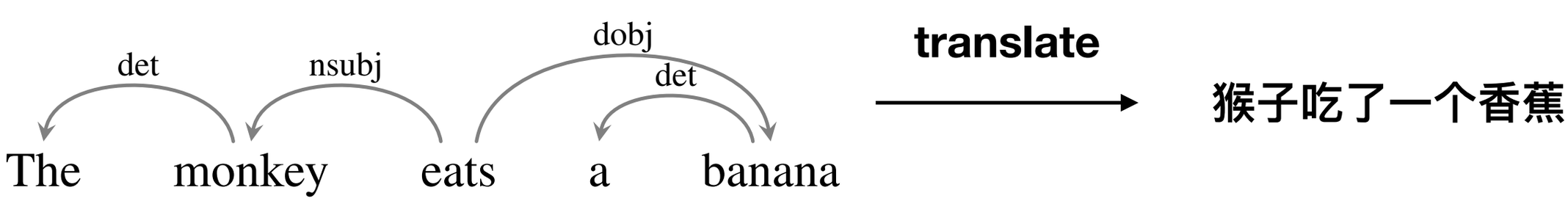
- Graphical structure in natural language processing (NLP)
  - Syntax
  - **Semantics e.g., Abstract Meaning Representation (AMR)**
  - Knowledge



The boy wants the girl to believe him.

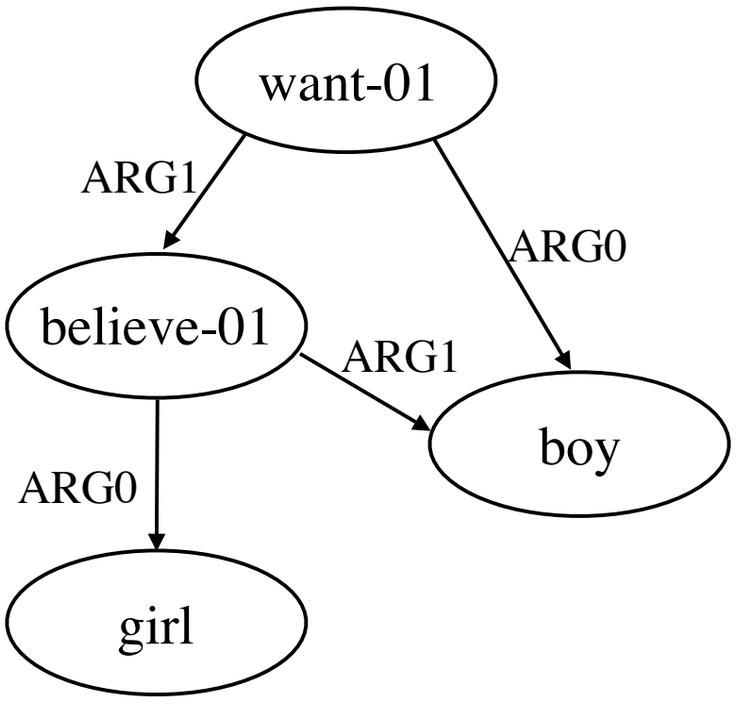
Graph-to-Sequence

- Syntactic machine translation



Graph-to-Sequence

- AMR-to-text generation



The boy wants the girl to believe him.

Existing Work

- **Grammar-based** methods (Jones et al., 2012; Flanigan et al., 2016; Song et al., 2016; Song et al., 2017)
  - use specialized graph-to-string transduction rules.
- **Seq2Seq-based** methods (Pourdamghani, Knight, and Hermjakob 2016; Konstas et al. 2017)
  - treat graph as sequence by linearizing input graphs.
- **Graph neural network-based** methods (Beck, Haffari, and Cohn 2018; Song et al., 2018; Basting et al, 2017; Damonte and Cohen, 2019; Guo et al., 2019; Koncel-Kedziorski et al 2019)
  - Directly and explicitly model the graph structure.

Existing Work

- Graph neural network based methods (Beck, Haffari, and Cohn 2018; Song et al., 2018; Basting et al, 2017; Damonte and Cohen, 2019; Guo et al., 2019; Koncel-Kedziorski et al, 2019)
  - previous SOTA
  - compute the representation of each node iteratively based on those of its adjacent nodes.

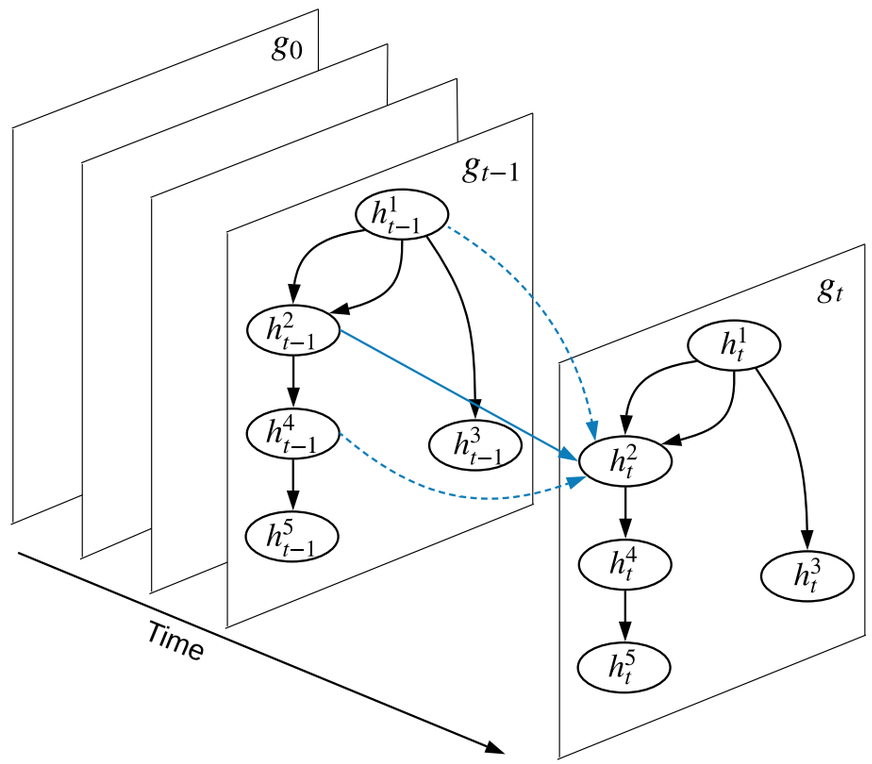
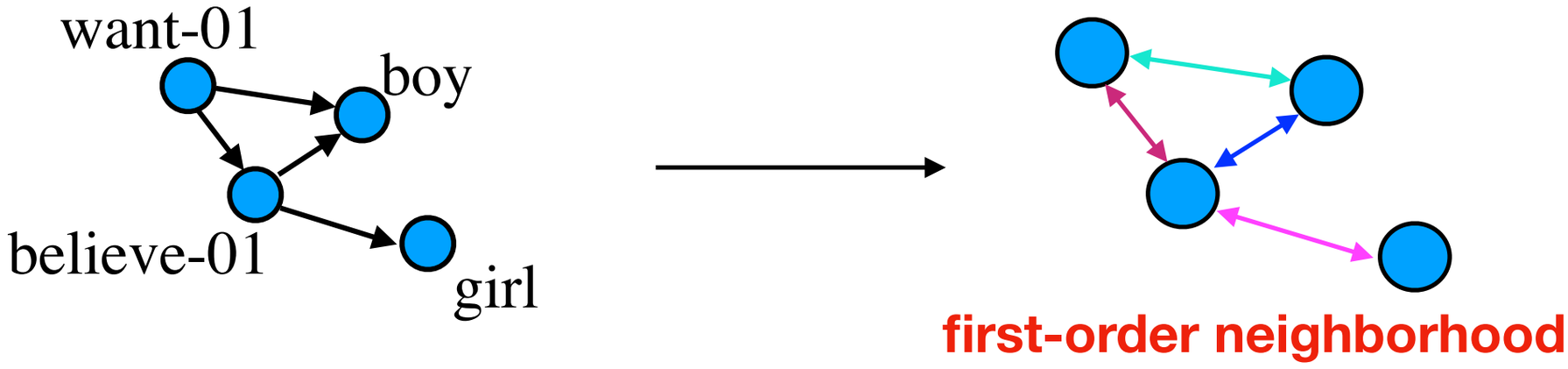


figure from Song et al., 2018

Existing Work

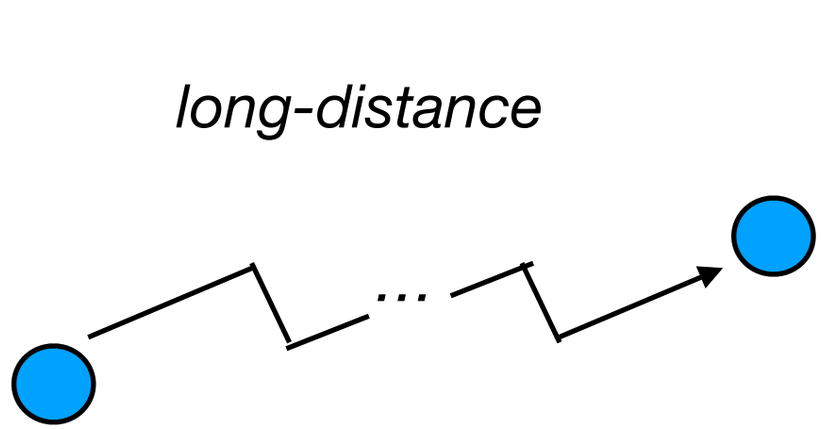
- Graph neural network based methods (Beck, Haffari, and Cohn 2018; Song et al., 2018; Basting et al, 2017; Damonte and Cohen, 2019; Guo et al., 2019; Koncel-Kedziorski et al, 2019)



- Different information passing schemes
  - Attention over adjacent neighbors (Koncel-Kedziorski et al, 2019)
  - Graph convolutional neural networks (Kipf and Welling 2017)
  - Gated graph neural network (Li et al, 2016; Song et al, 2018)

Motivation

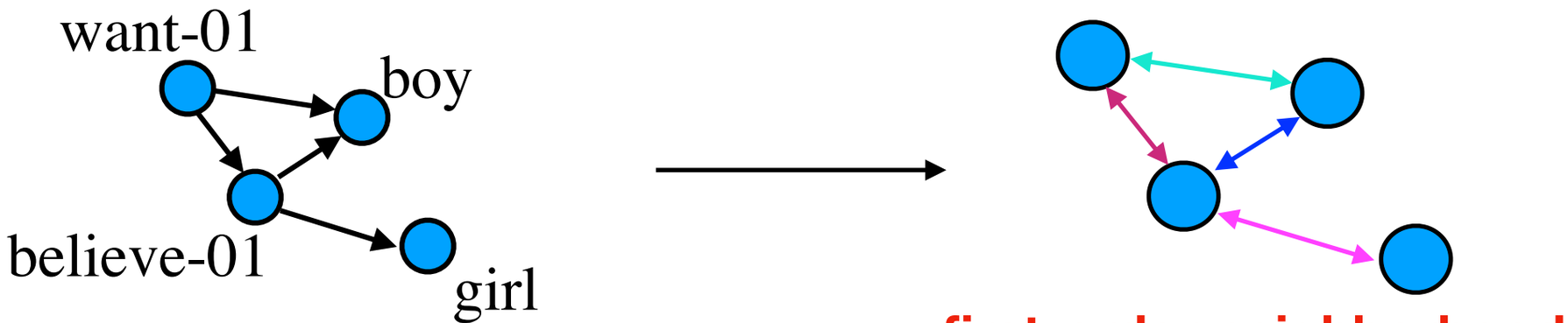
- Drawbacks of existing graph neural network based methods



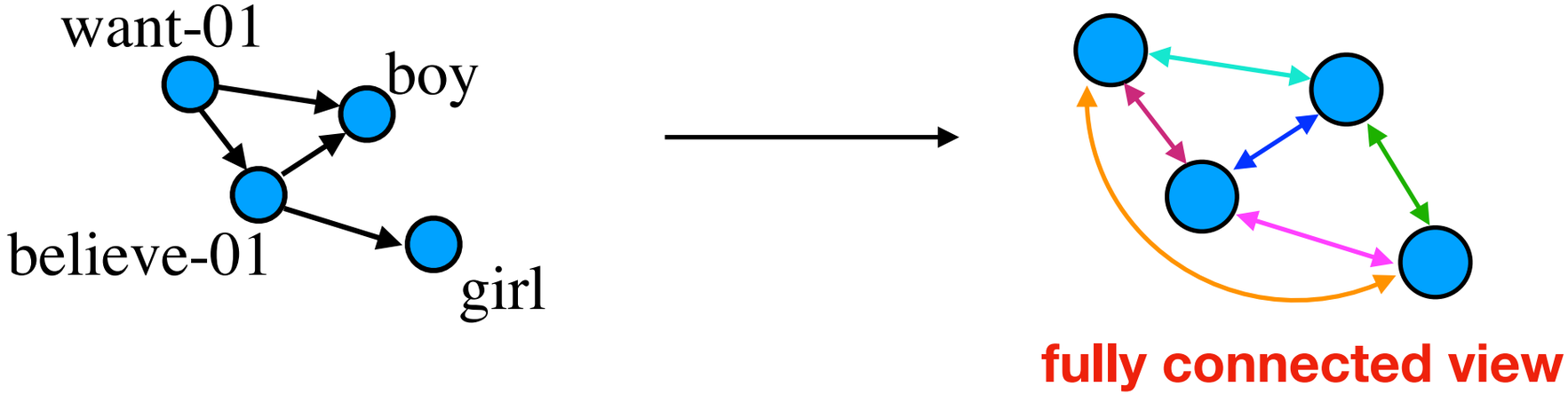
  - The local propagation nature precludes efficient global communication
  - The information may also be disrupted in the long journey.

Motivation

- **Graph Neural Network**



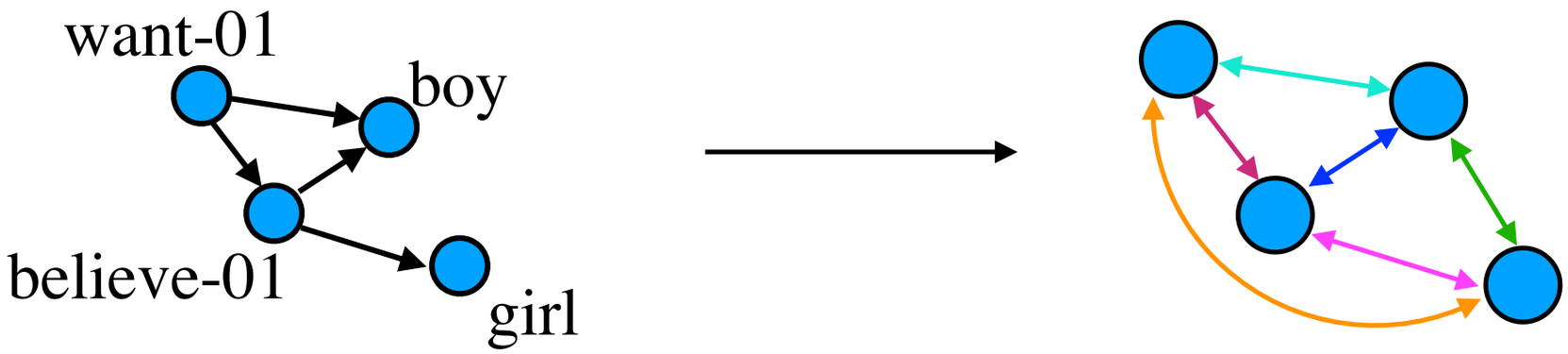
- **Graph Transformer (Ours)**



Graph-to-Sequence

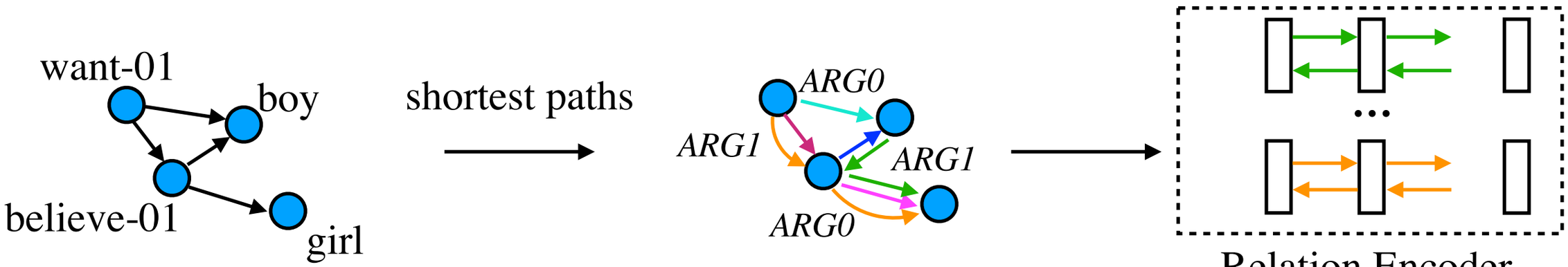
- Graph Encoder
    - responsible for transforming an input graph into a set of corresponding node embeddings.
- Sequence Decoder
    - responsible for yield the natural language sequence.

Graph Encoder



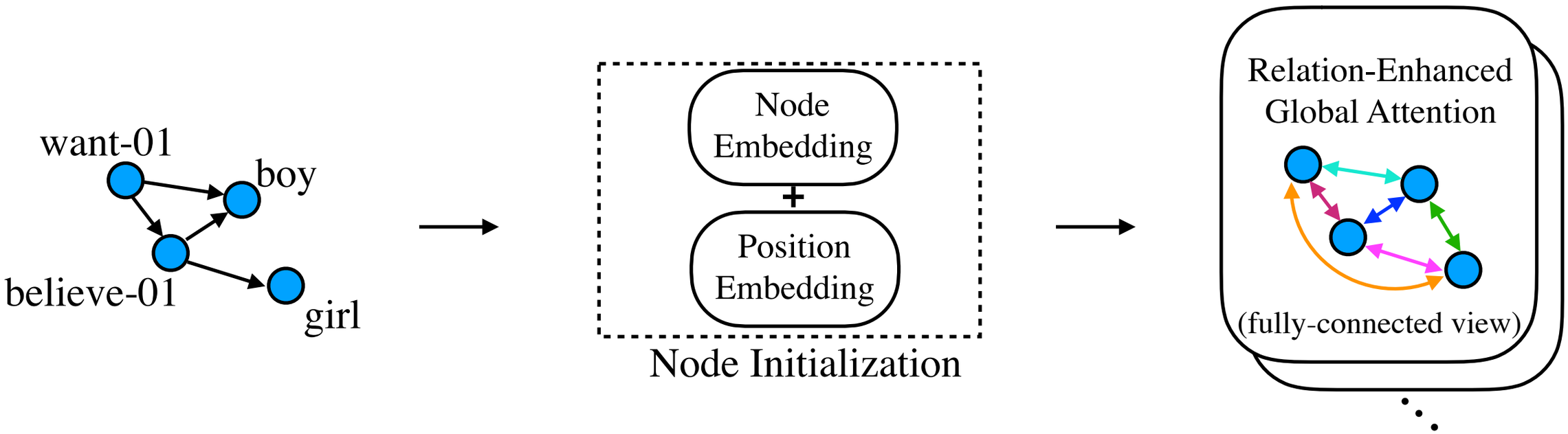
- **Explicit Relation Encoding**
- **Global Attention Network**

Explicit Relation Encoding



- **Shortest paths**
- **Recurrent neural networks**

Global Attention Network



- **Relation-enhanced Global Attention Mechanism**

Relation-enhanced Global Attention Mechanism

$$
\begin{aligned}
s_{ij} &= f(x_i, x_j) \\
&= x_i W_q^T W_k x_j
\end{aligned}
$$

Relation-enhanced Global Attention Mechanism

$$s_{ij} = f(x_i, x_j)$$
$$= x_i W_q^T W_k x_j$$

$$[r_{i\to j}; r_{j\to i}] = W_r r_{ij}$$

Relation-enhanced Global Attention Mechanism

$$s_{ij} = f(x_i, x_j)$$
$$= x_i W_q^T W_k x_j$$

$$[r_{i\to j}; r_{j\to i}] = W_r r_{ij}$$

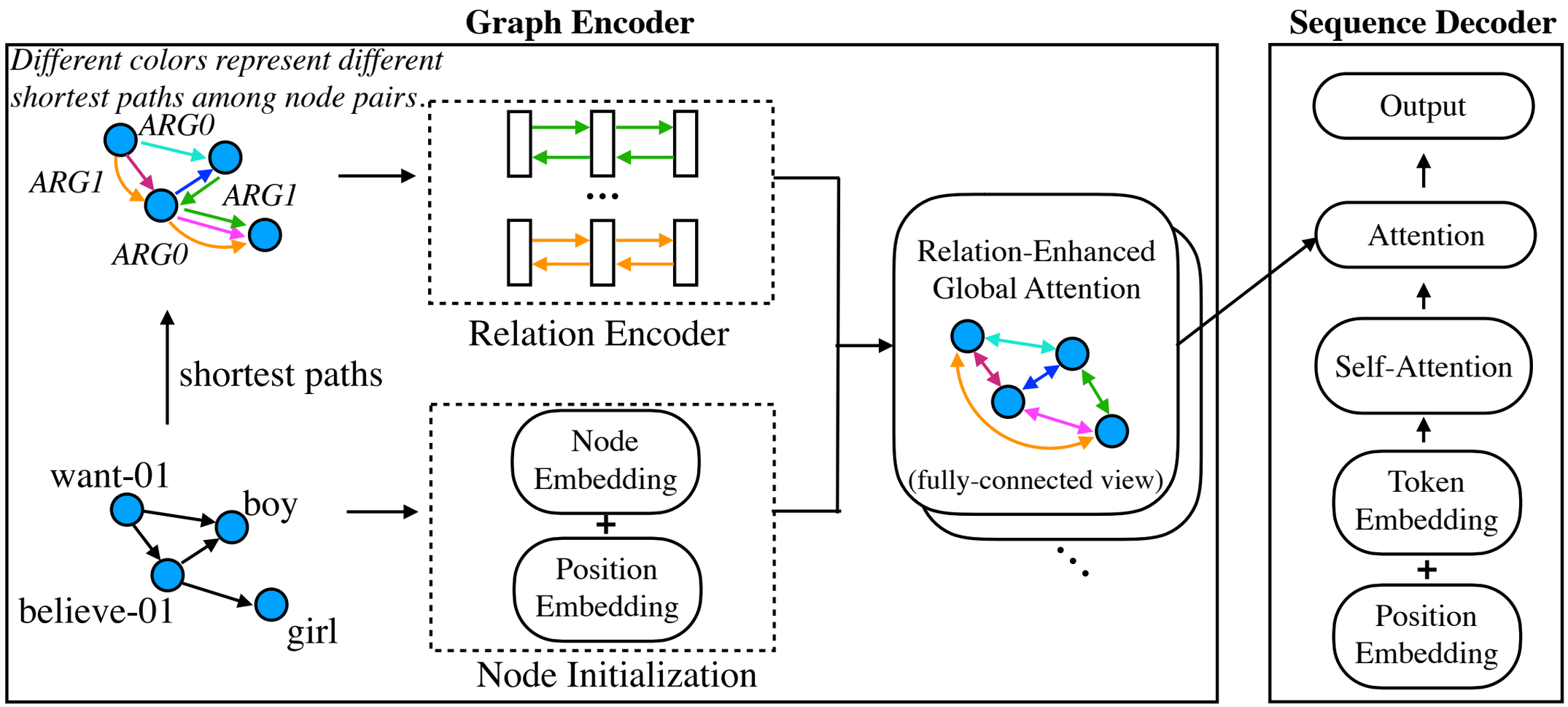
$$s_{ij} = g(x_i, x_j, r_{ij})$$
$$= (x_i + r_{i\to j}) W_q^T W_k (x_j + r_{j\to i})$$
$$= \underbrace{x_i W_q^T W_k x_j}_{(a)} + \underbrace{x_i W_q^T W_k r_{j\to i}}_{(b)}$$
$$+ \underbrace{r_{i\to j} W_q^T W_k x_j}_{(c)} + \underbrace{r_{i\to j} W_q^T W_k r_{j\to i}}_{(d)}$$

Relation-enhanced Global Attention Mechanism

$$
\begin{aligned}
s_{ij} &= g(x_i, x_j, r_{ij}) \\
&= (x_i + r_{i\to j}) W_q^T W_k (x_j + r_{j\to i}) \\
&= \underbrace{x_i W_q^T W_k x_j}_{(a)} + \underbrace{x_i W_q^T W_k r_{j\to i}}_{(b)} \\
&+ \underbrace{r_{i\to j} W_q^T W_k x_j}_{(c)} + \underbrace{r_{i\to j} W_q^T W_k r_{j\to i}}_{(d)}
\end{aligned}
$$

- The term (a) captures purely content-based addressing
- The term (b) represents a source-dependent relation bias.
- The term (c) governs a target-dependent relation bias.
- The term (d) encodes the universal relation bias.

Graph Transformer

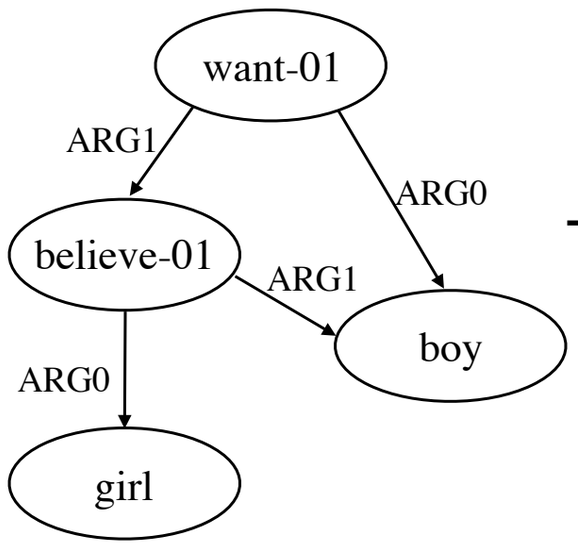


Experiments

- **AMR-to-text Generation**
- **Syntax-based Machine Translation**

Experiments

- **AMR-to-text Generation**



The boy wants the girl to believe him.

AMR-to-text Generation

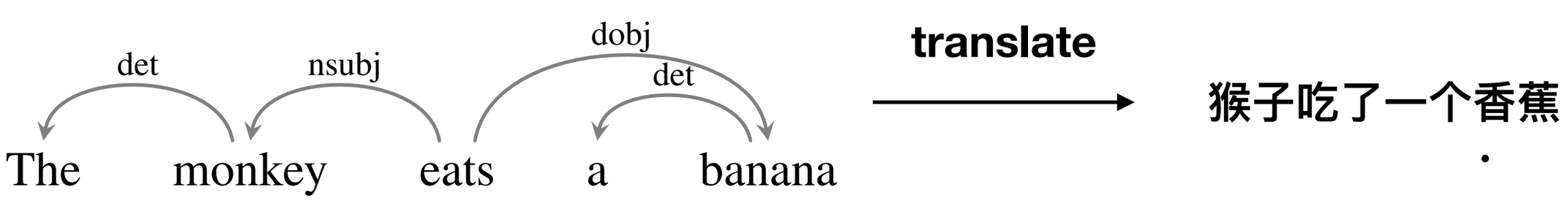
Dataset	#train	#dev	#test	#edge types	#node types	avg #nodes	avg #edges	avg diameter
LDC2015E86	16,833	1,368	1,371	113	18735	17.34	17.53	6.98
LDC2017T10	36,521	1,368	1,371	116	24693	14.51	14.62	6.15

	Model	LDC2015E86			LDC2017T10		
		BLEU	CHRF++	METEOR	BLEU	CHRF++	METEOR
statistical methods	Song et al.(2016)†	22.4	-	-	-	-	-
	Flanigan et al.(2016)†	23.0	-	-	-	-	-
	Pourdamghani, Knight, and Hermjakob(2016)†	26.9	-	-	-	-	-
	Song et al.(2017)†	25.6	-	-	-	-	-
Seq2seq (Neural)	Konstas et al.(2017)	22.0	-	-	-	-	-
	Cao and Clark(2019)‡	23.5	-	-	26.8	-	-
GNN based	Song et al.(2018)	23.3	-	-	24.9	-	-
	Beck, Haffari, and Cohn(2018)	-	-	-	23.3	50.4	
	Damonte and Cohen(2019)	24.4	-	23.6	24.5	-	24.1
	Guo et al.(2019)	25.7	54.5*	31.5*	27.6	57.3	34.0*
	Ours	**27.4**	**56.4**	**32.9**	**29.8**	**59.4**	**35.1**

- The first neural model that surpasses the strong non-neural baseline established by Pourdamghani, Knight, and Hermjakob(2016)
- improving over the state-of-the-art sequence-to-sequence model by 3 BLEU points and the state-of-the-art GNN-based model by BLEU 2.2 points

Experiments

- **Syntax-based Machine Translation**



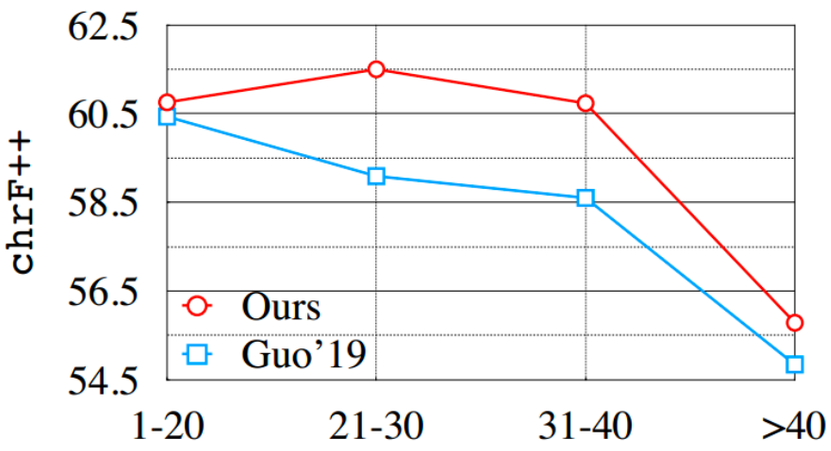
Syntax-based Machine Translation

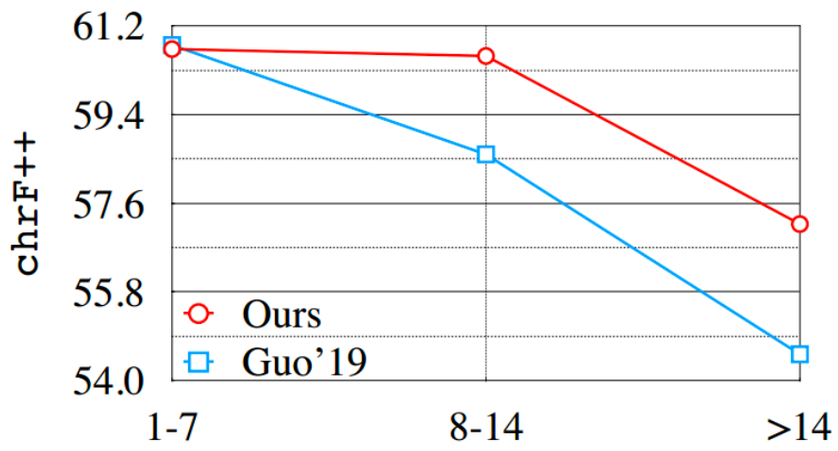
Dataset	#train	#dev	#test	#edge types	#node types	avg #nodes	avg #edges	avg diameter
English-Czech	181,112	2,656	2,999	46	78017	23.18	22.18	8.36
English-German	226,822	2,169	2,999	46	87219	23.29	22.29	8.42

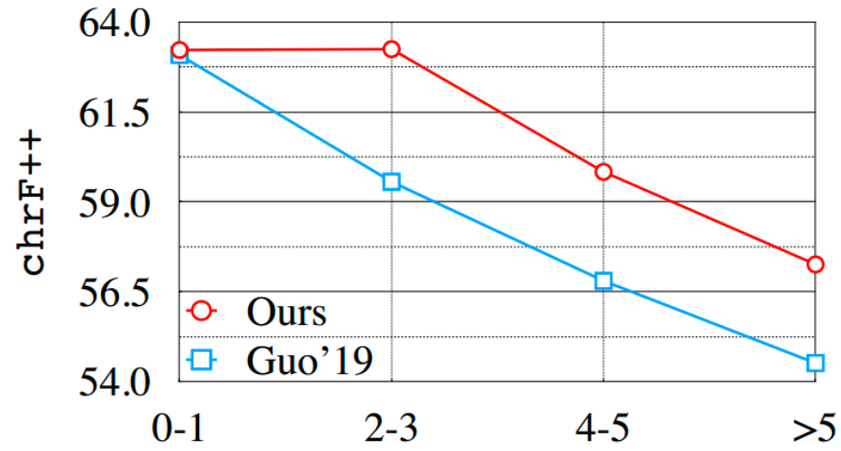
Model	Type	English-German		English-Czech	
		BLEU	CHRF++	BLEU	CHRF++
Bastings et al.(2017)	Single	16.1	-	9.6	-
Beck, Haffari, and Cohn(2018)	Single	16.7	42.4	9.8	33.3
Guo et al.(2019)	Single	19.0	44.1	12.1	37.1
Beck, Haffari, and Cohn(2018)	Ensemble	19.6	45.1	11.7	35.9
Guo et al.(2019)	Ensemble	20.5	45.8	13.1	37.8
Ours	Single	**21.3**	**47.9**	**14.1**	**41.1**

- Even better than previous state- of-the-art models that use ensembling!

More Analysis



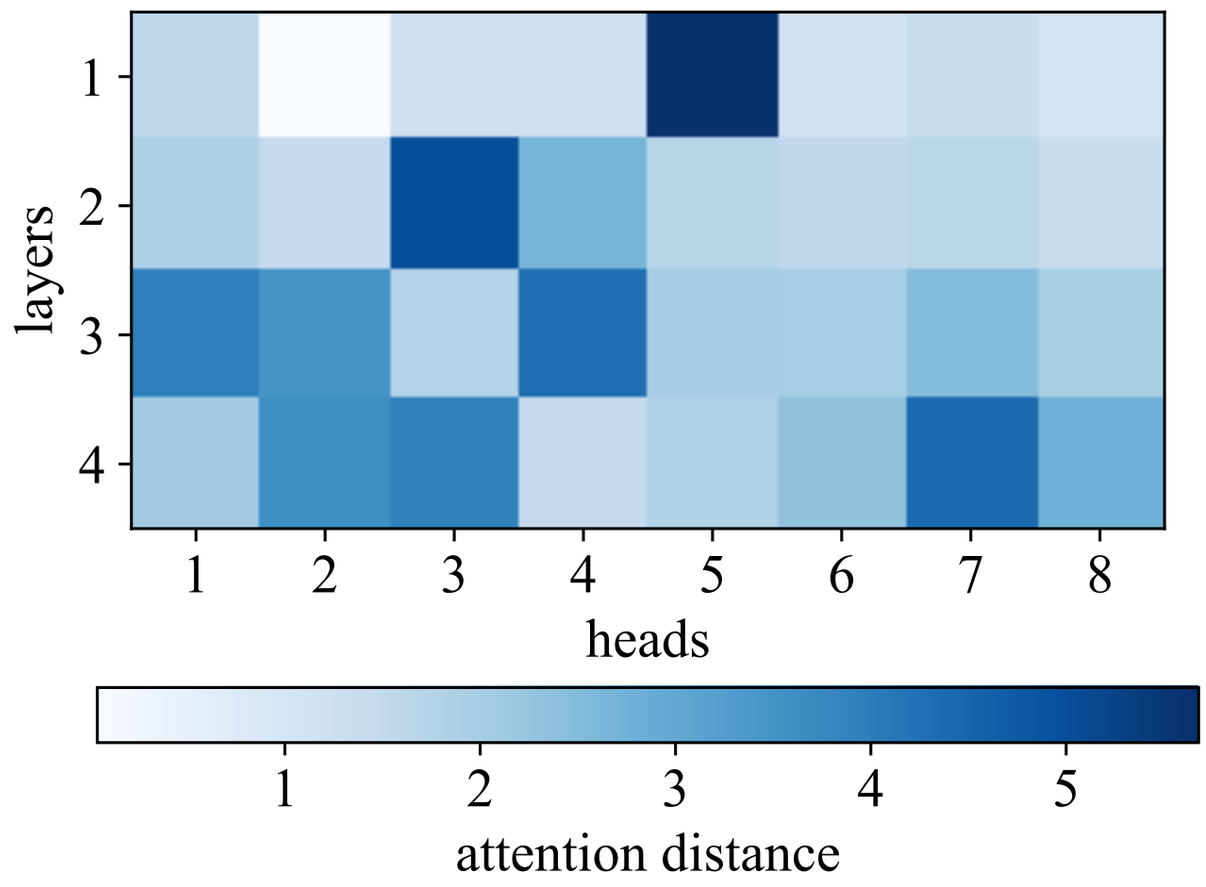




(on the test set of LDC2017T10)

- Graph size: Our model has better ability for dealing with complicated graphs.
- Graph diameter: Our model is superior in featuring long-distance dependencies
- Graph reentrancies: Our model is consistently better than the GNN-based model when there are more than one reentrancies

How Far Does Attention Look At



- The number of these far-sighted heads generally increases as layers go deeper.
- Interestingly, the longest-reaching head (layer1-head5) and the shortest-sighted head (layer1-head2) coexist in the very first layer.

Graph Transformer for Graph-to-Sequence Learning

Thanks!

Deng Cai and Wai Lam
The Chinese University of Hong Kong

https://github.com/jcyk/gtos
thisisjcykcd@gmail.com