Heterogeneous Graph Neural Networks for Keyphrase Generation

Jiacheng Ye[1]*, Ruijian Cai[1]*, Tao Gui[2]† and Qi Zhang[1]†
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[2]Institute of Modern Languages and Linguistics, Fudan University
{yejc19, 19210240253, tgui, qz}@fudan.edu.cn

EMNLP 2021

https://github.com/jiacheng-ye/kg_ gater

gesis
Leibniz-Institut für Sozialwissenschaften

Reported by Dongdong Hu

Introduction

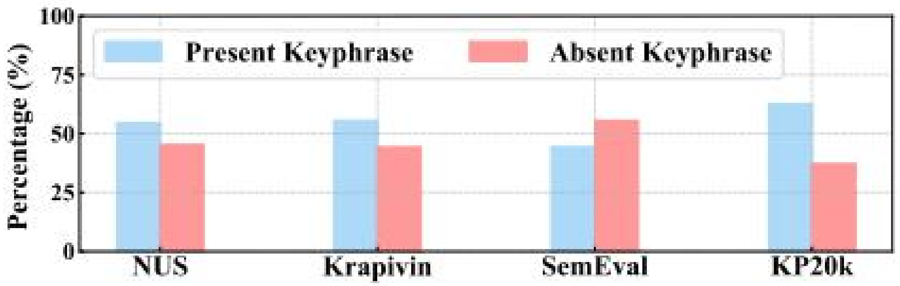


Figure 1: Proportion of present and absent keyphrases among four datasets. Although the previous methods for keyphrase generation have shown promising results on present keyphrase predictions, they are not yet satisfactory on the absent keyphrase predictions, which also occupy a large proportion.

Relying solely on the source document can result in generating uncontrollable and inaccurate absent keyphrases.

Method

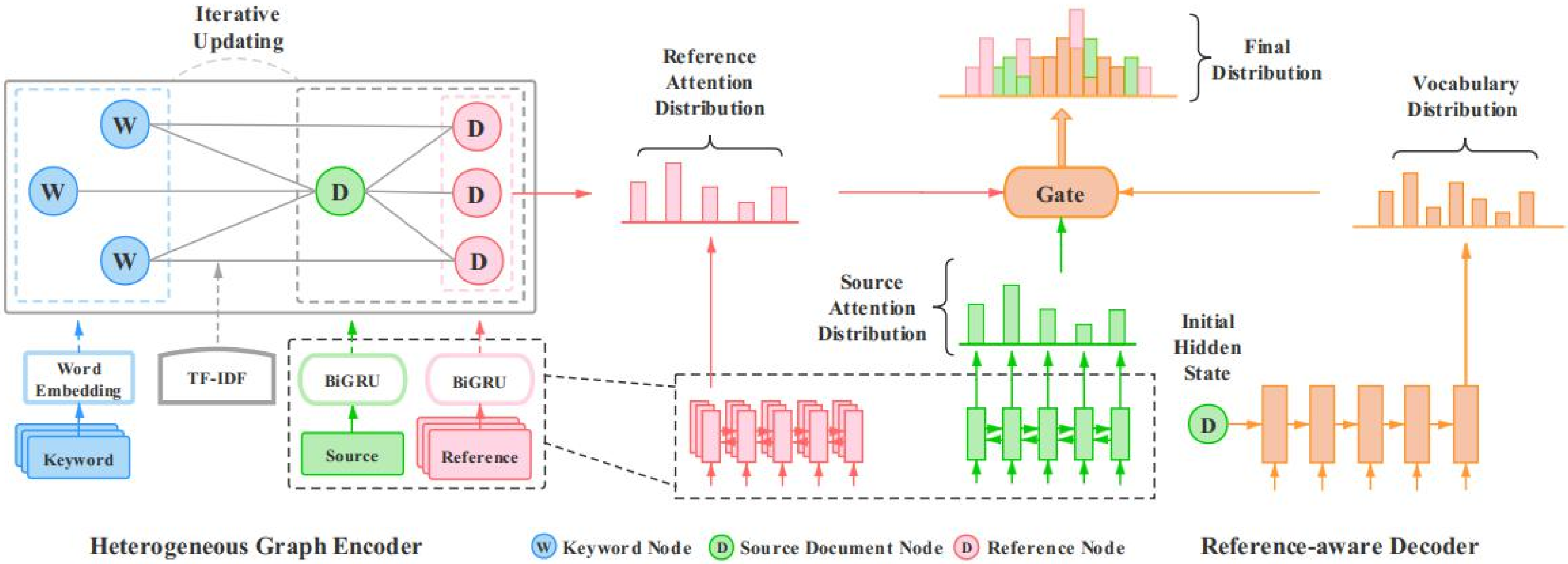


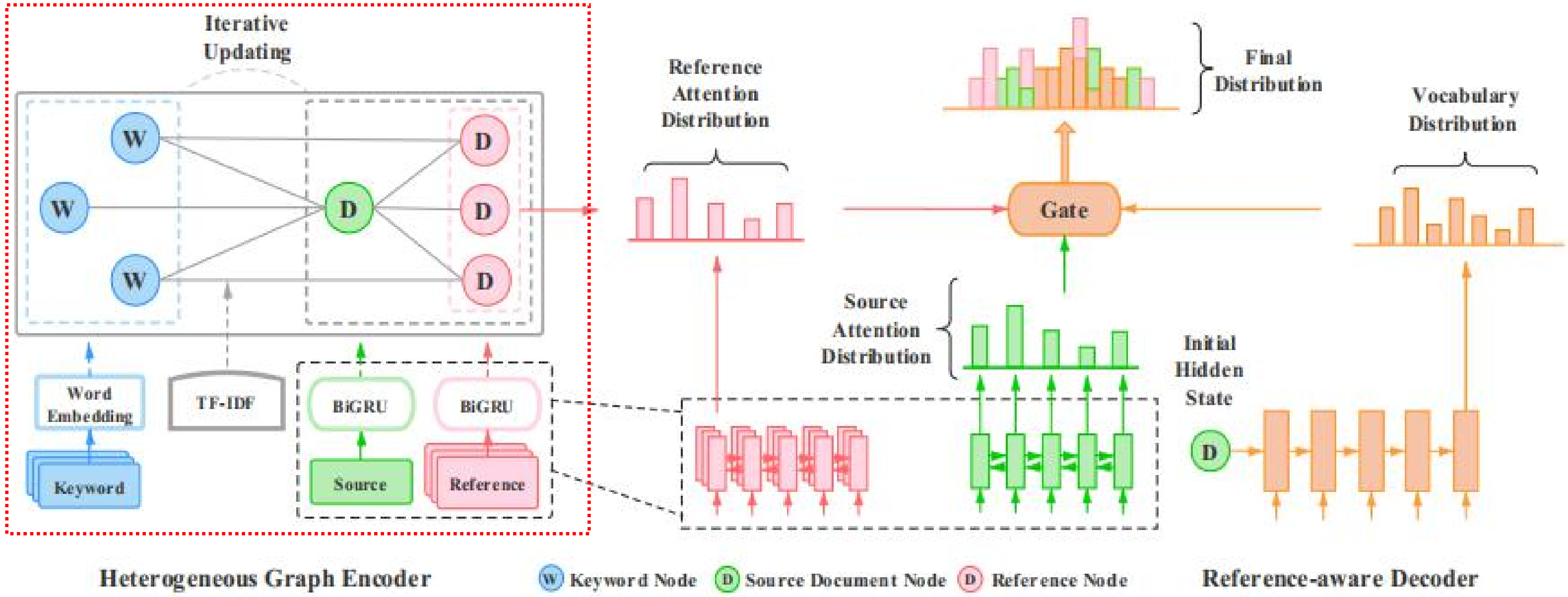
Figure 2: Graphical illustration of our proposed GATER. We first retrieve references using the source document, where each reference is the concatenation of document and keyphrases pair from the training set. Then we construct a heterogeneous graph and perform iterative updating. Finally, the source document node is extracted to decode the keyphrase sequence with a hierarchical attention and copy mechanism.

Method

we first represent the source document and all the reference candidates as TF-IDF weighted uni/bi-gram vectors. Then, the most similar K references $\mathcal{X}^r = \{\mathbf{x}^{r_i}\}_{i=1,\ldots,K}$ are retrieved by comparing the cosine similarities of the vectors of the source document and all the references.

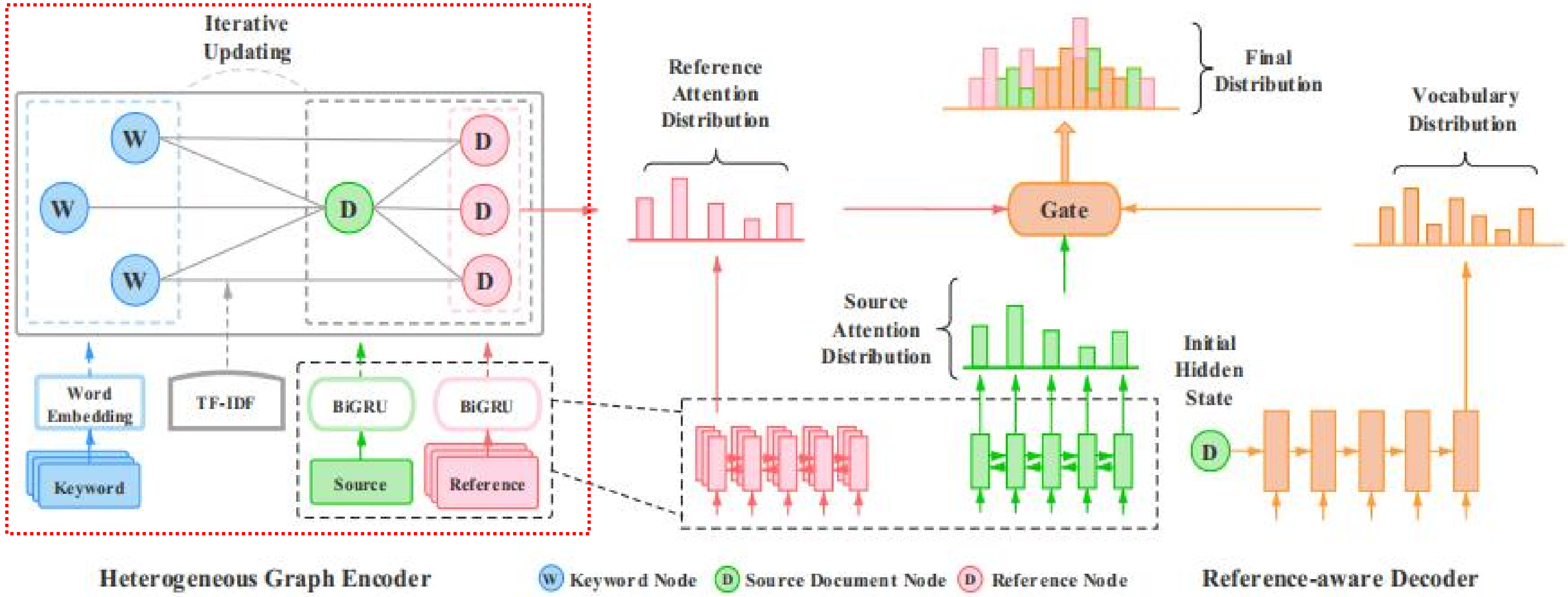
Graph Construction

$$G = \{V, E\}$$
$$V = V_w \cup V_d$$
$$V_w = \{w_i\}\ (i \in \{1, \ldots, m\})$$
$$V_d = \mathbf{x} \cup \mathcal{X}^r$$
$$E = E_{d2d} \cup E_{w2d}$$
$$E_{d2d} = \{e_k\}\ (k \in \{1, \ldots, K\})$$
$$E_{w2d} = \{e_{i,j}\}\ (i \in \{1, \ldots, m\}, j \in \{1, \ldots, K+1\})$$



Similarly, we also infuse TF-IDF values in the edge weights of E_{d2d} as a prior statistical n-gram similarity between documents.

Method



Graph Initializers

document node

$\mathbf{e}^w$

$\mathbf{d} = [\overrightarrow{\mathbf{m}}_1; \overleftarrow{\mathbf{m}}_{L_\mathbf{x}}]$ and $\mathbf{m}_i = [\overrightarrow{\mathbf{m}}_i; \overleftarrow{\mathbf{m}}_i]$

keyword node

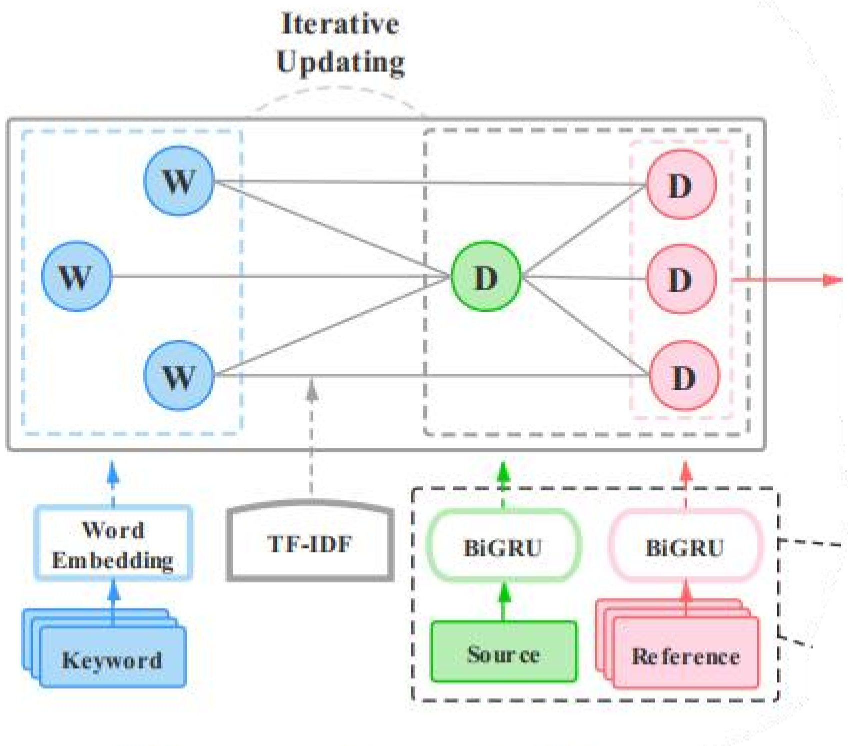
$\mathbf{w}_i = \mathbf{e}^w(w_i).$

edge initializers

E_{d2d} E_{d2w}

$\mathbf{e}^{d2d}$ and $\mathbf{e}^{w2d}$.

Method



Heterogeneous Graph Encoder

$$z_{ij} = \text{LeakyReLU}\left(\mathbf{w}_a^T\left[\mathbf{W}_q\mathbf{h}_i; \mathbf{W}_k\mathbf{h}_j; \mathbf{e}_{ij}\right]\right)$$

$$\alpha_{ij} = \text{softmax}_j\left(z_{ij}\right) = \frac{\exp\left(z_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(z_{ik}\right)}$$

$$\mathbf{u}_i = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}_v\mathbf{h}_j\right),$$

$\text{GAT}(\mathbf{H}, \mathbf{H}, \mathbf{H}, \mathbf{E})$ to denote the GAT aggregating layer

$\mathbf{H}$ is used for query, key, and value

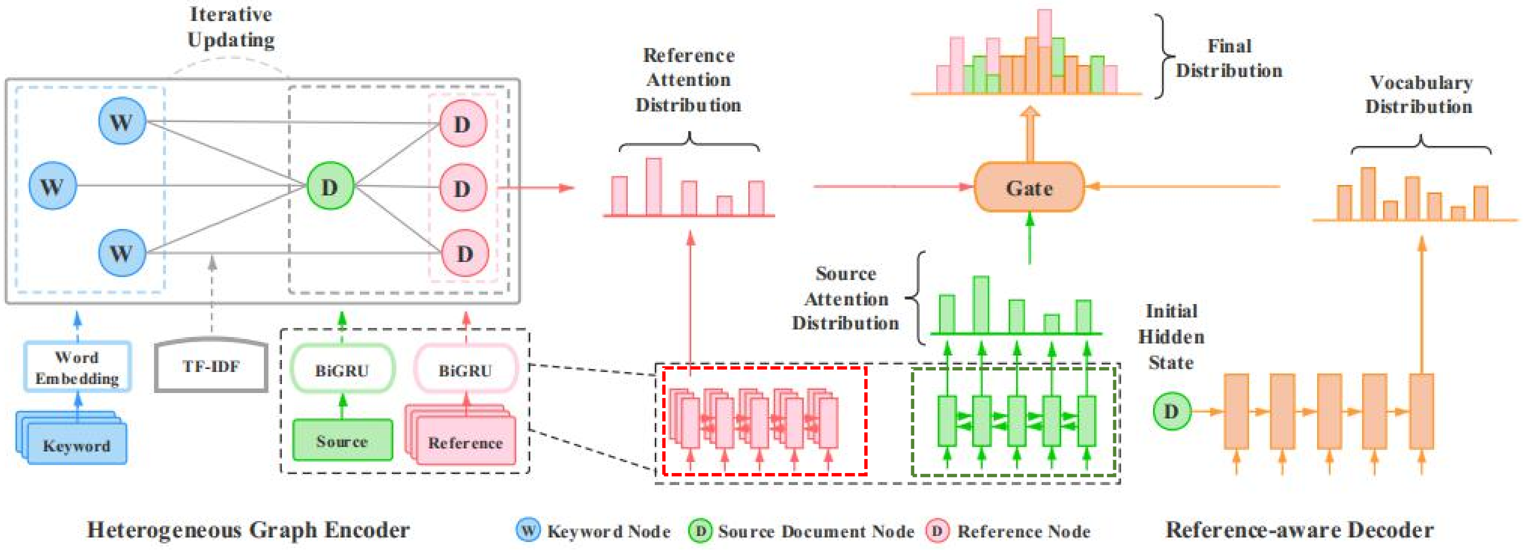
$$\mathbf{H}_w^1 = \text{FFN}\left(\text{GAT}\left(\mathbf{H}_w^0, \mathbf{H}_d^0, \mathbf{H}_d^0, \mathbf{E}_{w2d}\right) + \mathbf{H}_w^0\right)$$
$$\mathbf{H}_d^1 = \text{FFN}\left(\text{GAT}\left(\mathbf{H}_d^0, \mathbf{H}_w^1, \mathbf{H}_w^1, \mathbf{E}_{w2d}\right) + \mathbf{H}_d^0\right)$$
$$\mathbf{H}_d^1 = \text{FFN}\left(\text{GAT}\left(\mathbf{H}_d^1, \mathbf{H}_d^1, \mathbf{H}_d^1, \mathbf{E}_{d2d}\right) + \mathbf{H}_d^1\right), \quad (2)$$

we introduce a residual connection and position-wise feed-forward (FFN) layer consisting of two linear transformations.

seperate $\mathbf{H}_d^I$ into $\mathbf{d}^s$ and $\mathbf{D}^r = \{\mathbf{d}^{r_i}\}_{i=1...,K}$

Experiments

$\mathbf{H}_d^l$ into $\mathbf{d}^s$ and $\mathbf{D}^r = \{\mathbf{d}^{r_i}\}_{i=1...,K}$



$$\mathbf{M}^r = \{\mathbf{M}^{r_i}\}_{i=1...,K}$$
$$\mathbf{M}^{r_i} = \{\mathbf{m}_j^{r_i}\}_{j=1...,L_{r_i}}$$

$$\mathbf{M}^s = \{\mathbf{m}_i^s\}_{i=1,...,L_\mathbf{x}}$$

$$\mathbf{h}_t = \mathrm{GRU}(\mathbf{e}^w(y_{t-1}), \mathbf{h}_{t-1})$$
$$\mathbf{c}_t = \mathrm{hier\_attn}(\mathbf{h}_t, \mathbf{M}^s, \mathbf{M}^r, \mathbf{D}^r) \quad (3)$$
$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_c[\mathbf{c}_t; \mathbf{h}_t]),$$

hier_attn

$$\mathbf{c}_t^s = \sum_{i=1}^{L_\mathbf{x}} a_{t,i}^s \mathbf{m}_i^s; \mathbf{c}_t^r = \sum_{i=1}^{K} \sum_{j=1}^{L_{\mathbf{x}^{r_i}}} a_{t,i}^r a_{t,j}^{r_i} \mathbf{m}_j^{r_i} \quad (4)$$
$$\mathbf{c}_t = g_{ref} \cdot \mathbf{c}_t^s + (1 - g_{ref}) \cdot \mathbf{c}_t^r,$$

where $\mathbf{a}_t^s$ is a word-level attention distribution over words from the source document using $\mathbf{M}^s$

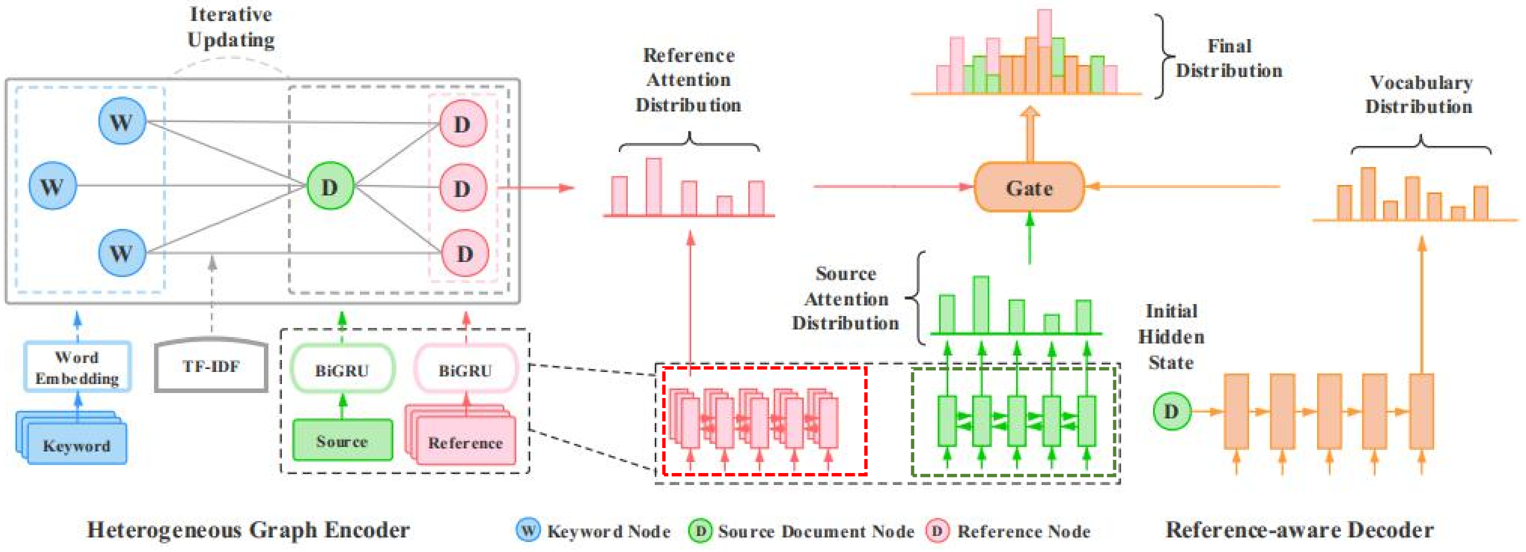
$\mathbf{a}_t^r$ is an attention distribution over references using $\mathbf{D}^r$.

$\mathbf{a}_t^{r_i}$ is a word-level attention distribution over words from i-th reference using $\mathbf{M}^{r_i}$.

$$g_{ref} = \mathrm{sigmoid}(\mathbf{w}_{ref}[\mathbf{c}_t^s; \mathbf{c}_t^r])$$

Method

$\mathbf{H}_d^l$ into $\mathbf{d}^s$ and $\mathbf{D}^r = \{\mathbf{d}^{r_i}\}_{i=1\ldots,K}$



$$\mathbf{M}^r = \{\mathbf{M}^{r_i}\}_{i=1\ldots,K}$$
$$\mathbf{M}^{r_i} = \{\mathbf{m}_j^{r_i}\}_{j=1\ldots,L_{r_i}}$$

$$\mathbf{M}^s = \{\mathbf{m}_i^s\}_{i=1,\ldots,L_{\mathbf{x}}}$$

hierarchical copy mechanism

$$P_{\mathcal{V}'}(y_t) = p_1 P_{\mathcal{V}}(y_t) + p_2 P_{\mathcal{V}_{\mathbf{x}}}(y_t) + p_3 P_{\mathcal{V}_{\mathcal{X}^r}}(y_t)$$

$$P_{\mathcal{V}}(y_t) = \operatorname{softmax}(\operatorname{MLP}([\mathbf{h}_t; \tilde{\mathbf{h}}_t]))$$

$$P_{\mathcal{V}_{\mathbf{x}}}(y_t) = \sum\nolimits_{i:x_i=y_t} a_{t,i}^s$$

$$P_{\mathcal{V}_{\mathcal{X}^r}}(y_t) = \sum\nolimits_i \sum\nolimits_{j:x_j^{r_i}=y_t} a_{t,j}^{r_i}$$

$$\mathbf{p} = \operatorname{softmax}(\mathbf{W}_p[\mathbf{h}_t; \mathbf{h}_t; \mathbf{e}^w(y_{t-1})]) \in \mathbb{R}^3$$

loss

$$\mathcal{L}_{\text{ONE2ONE}}(\theta) = -\sum_{i=1}^{|\mathcal{Y}|} \sum_{t=1}^{L_{\mathbf{y}_i}} \log P_{\mathcal{V}'}(y_{i,t} \mid \mathbf{y}_{i,1:t-1}, \mathbf{x}; \theta),$$

$$\mathcal{L}_{\text{ONE2SEQ}}(\theta) = -\sum_{t=1}^{L_{\mathbf{y}^\star}} \log P_{\mathcal{V}'}(y_t^\star \mid \mathbf{y}^\star{}_{1:t-1}, \mathbf{x}; \theta)$$

Experiments

Model	NUS				SemEval				KP20k			
	Present		Absent		Present		Absent		Present		Absent	
	$F1@5$	$F1@10$	$R@10$	$R@50$	$F1@5$	$F1@10$	$R@10$	$R@50$	$F1@5$	$F1@10$	$R@10$	$R@50$
CopyRNN (Meng et al., 2017)	0.311	0.266	0.058	0.116	0.293	0.304	0.043	0.067	0.333	0.262	0.125	0.211
CorrRNN (Chen et al., 2018)	0.318	0.278	0.059	-	0.320	0.320	0.041	-	-	-	-	-
TG-Net (Chen et al., 2019b)	0.349	0.295	0.075	0.137	0.318	0.322	0.045	0.076	0.372	0.315	0.156	0.268
KG-KE-KR-M (Chen et al., 2019a)	0.344	0.287	0.123	**0.193**	0.329	0.327	0.049	0.090	0.400	**0.327**	0.177	0.278
CopyRNN-GATER (Ours)	$\mathbf{0.374}_4$	$\mathbf{0.304}_4$	$\mathbf{0.126}_3$	$\mathbf{0.193}_2$	$\mathbf{0.366}_3$	$\mathbf{0.340}_4$	$\mathbf{0.056}_1$	$\mathbf{0.092}_2$	$\mathbf{0.402}_1$	0.324_1	$\mathbf{0.186}_0$	$\mathbf{0.285}_1$

Table 1: Keyphrase prediction results of all the models trained under ONE2ONE paradigm. The best results are bold. The subscript are corresponding standard deviation (e.g., 0.285_1 means 0.285 ± 0.001).

Experiments

Model	NUS				SemEval				KP20k			
	Present		Absent		Present		Absent		Present		Absent	
	$F1@5$	$F1@M$	$F1@5$	$F1@M$	$F1@5$	$F1@M$	$F1@5$	$F1@M$	$F1@5$	$F1@M$	$F1@5$	$F1@M$
catSeq (Yuan et al., 2020)	0.323	0.397	0.016	0.028	0.242	0.283	0.020	0.028	0.291	0.367	0.015	0.032
catSeqD (Yuan et al., 2020)	0.321	0.394	0.014	0.024	0.233	0.274	0.016	0.024	0.285	0.363	0.015	0.031
catSeqCorr (Chan et al., 2019)	0.319	0.390	0.014	0.024	0.246	0.290	0.018	0.026	0.289	0.365	0.015	0.032
catSeqTG (Chan et al., 2019)	0.325	0.393	0.011	0.018	0.246	0.290	0.019	0.027	0.292	0.366	0.015	0.032
SenSeNet (Luo et al., 2020)	**0.348**	0.403	0.018	0.032	0.255	0.299	0.024	0.032	**0.296**	0.370	0.017	0.036
catSeq-GATER (Ours)	0.337_4	$\mathbf{0.418}_4$	$\mathbf{0.033}_3$	$\mathbf{0.054}_4$	$\mathbf{0.257}_3$	$\mathbf{0.309}_4$	$\mathbf{0.026}_4$	$\mathbf{0.035}_5$	0.295_2	$\mathbf{0.384}_1$	$\mathbf{0.030}_1$	$\mathbf{0.060}_2$

Table 2: Keyphrase prediction results of all the models trained under ONE2SEQ paradigm. The best results are bold. The subscript are corresponding standard deviation (e.g., 0.060_2 means 0.060 ± 0.002).

Experiments

Model	Present $F_1@5$	Present $F_1@M$	Absent $F_1@5$	Absent $F_1@M$
catSeq-GATER	**0.295**	**0.384**	**0.030**	**0.060**
Input Reference				
- retrieved documents	0.293	0.377	0.026	0.052
- retrieved keyphrases	0.291	0.369	0.018	0.037
- both	0.291	0.367	0.015	0.032
Heterogeneous Graph Encoder				
- $d2d$ edge	0.294	0.379	0.024	0.049
- $w2d$ edge	0.294	0.379	0.026	0.052
- both	0.293	0.371	0.020	0.041
Reference-aware Decoder				
- hierarchical copy	0.293	0.373	0.022	0.042
- hierarchical attention	0.291	0.368	0.018	0.036

Table 3: Ablation study of catSeq-GATER on **KP20k** dataset. All references are ignored in graph encoder when removing $d2d$ edge and the heterogeneous graph becomes homogeneous graph when removing $w2d$ edge.

Experiments

Model	Present		Absent	
	$F1@5$	$F1@M$	$F1@5$	$F1@M$
catSeqD	0.285	0.363	0.015	0.031
+ GATER	**0.294**	**0.381**	**0.025**	**0.051**
catSeqCorr	0.289	0.365	0.015	0.032
+ GATER	**0.296**	**0.384**	**0.030**	**0.060**
catSeqTG	0.292	0.366	0.015	0.032
+ GATER	**0.293**	**0.380**	**0.025**	**0.052**

Table 4: Results of applying our GATER to other baseline models on **KP20k** test set. The best results are bold.

Experiments

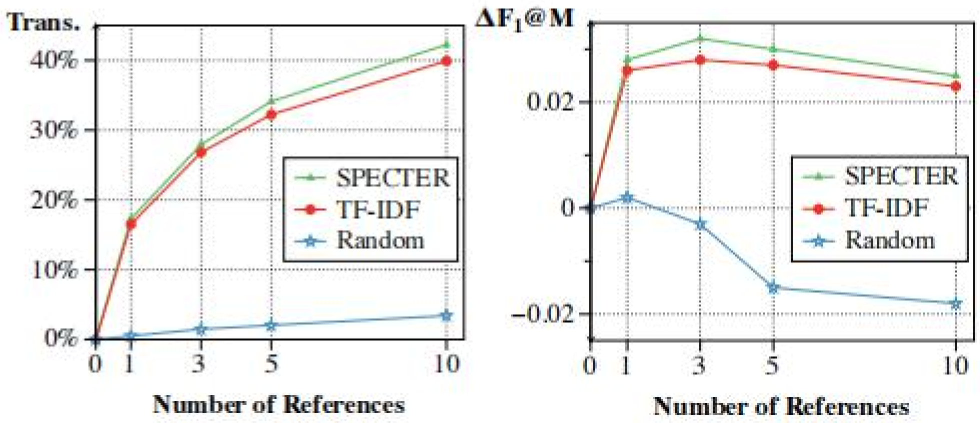


Figure 3: Transforming rate and $\Delta F_1@M$ for absent keyphrases under different types of retrievers on **KP20k** dataset for catSeq-GATER. We study a random retriever, a sparse retriever based on TF-IDF and a dense retriever based on SPECTER.

Experiments

Document: **natural convection** in porous annular domains **mimetic scheme** and **family of steady states**. **natural convection** of the incompressible fluid in the porous media based on the darcy hypothesis (lapwood convection) gives an intriguing branching off of one parameter family of steady patterns. this scenario may be suppressed in computations when governing equations are approximated by schemes which do not preserve the **cosymmetry** property ...	
Present Keyphrases	natural convection; mimetic scheme; family of steady states; cosymmetry
CopyRNN	**natural convection**; porous media; polar coordinates; mimetic; annular porous domain; porous domain; finite difference; steady states; **mimetic scheme**; ...
KG-KE-KR-M	**natural convection**; porous media; **mimetic scheme**; mimetic; polar coordinates; ...
CopyRNN-GATER (Ours)	**natural convection**; porous media; **mimetic scheme**; **cosymmetry**; mimetic; darcy hypothesis; finite difference; polar coordinates; ...
Absent Keyphrases	darcy law; porous medium; finite difference method
CopyRNN	convective convection; annular porous media; mimetic method; **finite difference method**; ...
KG-KE-KR-M	cosymmetry convection; mimetic method; **darcy law**; convective patterns; lapwood property; annular porous media; **finite difference method**; ...
CopyRNN-GATER (Ours)	**darcy law**; convective patterns; **porous medium**; multicomponent fluid; **finite difference method**; family convection; cosymmetry convection; staggered grids; darcy formulation; ...

Figure 4: Example of generated keyphrases by different models. The top 10 predictions are compared and some incorrect predictions are omitted for simplicity. The correct predictions are in bold blue and bold red for present and absent keyphrase, respectively. The absent predictions that appear in the references are highlighted in yellow, where only the keyphrases of retrieved documents are considered as references for KG-KE-KR-M.

Thanks