DiffuSum: Generation Enhanced Extractive Summarization with Diffusion

Haopeng Zhang*, Xiao Liu*, Jiawei Zhang
IFM Lab, Department of Computer Science, University of California, Davis, CA, USA
haopeng,xiao,jiawei@ifmlab.org

（ACL-2023)

code： None

gesis
Leibniz-Institut
für Sozialwissenschaften

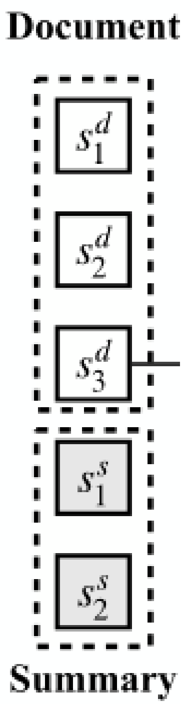
Reported by Zhaoze Gao

1. Introduction

2. Approach

3. Experiments

Introduction



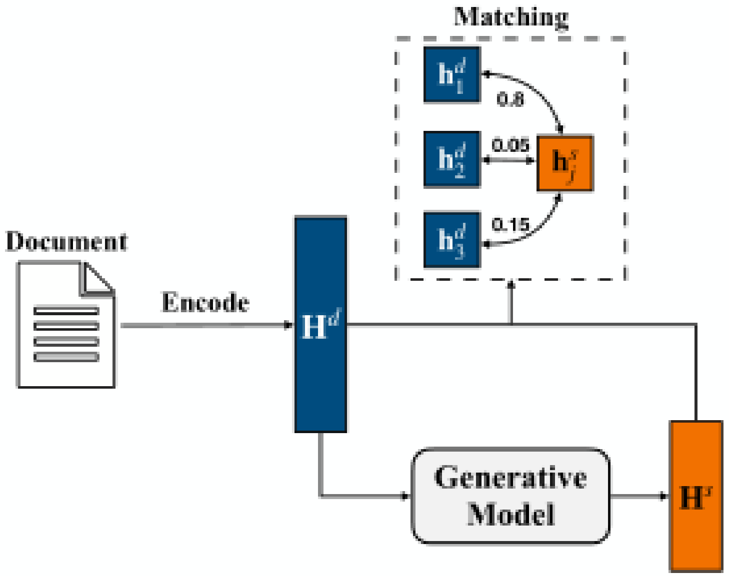


Figure 1: The proposed generation-enhanced extractive summarization framework.

Most existing work formulates it as sequence labeling and give each sentence a {0, 1} label, where label 1 indicates that the sentence will be included in summary.

.

Approach

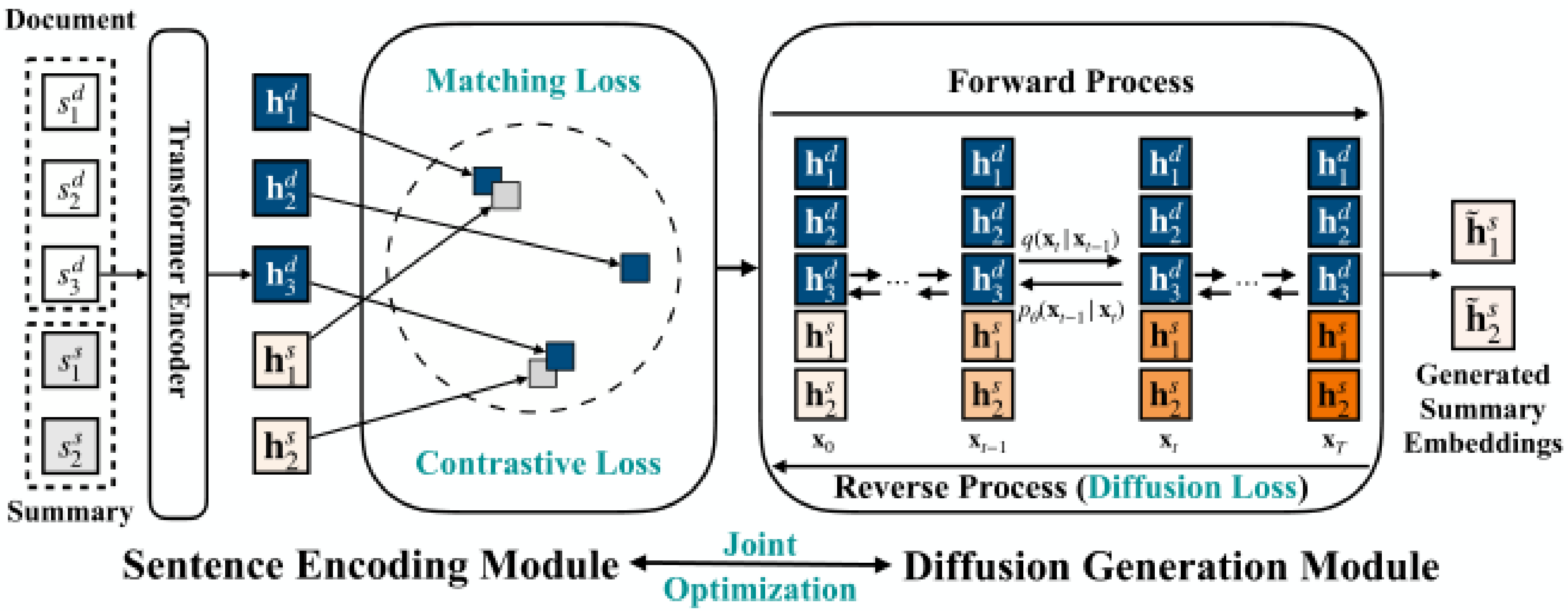
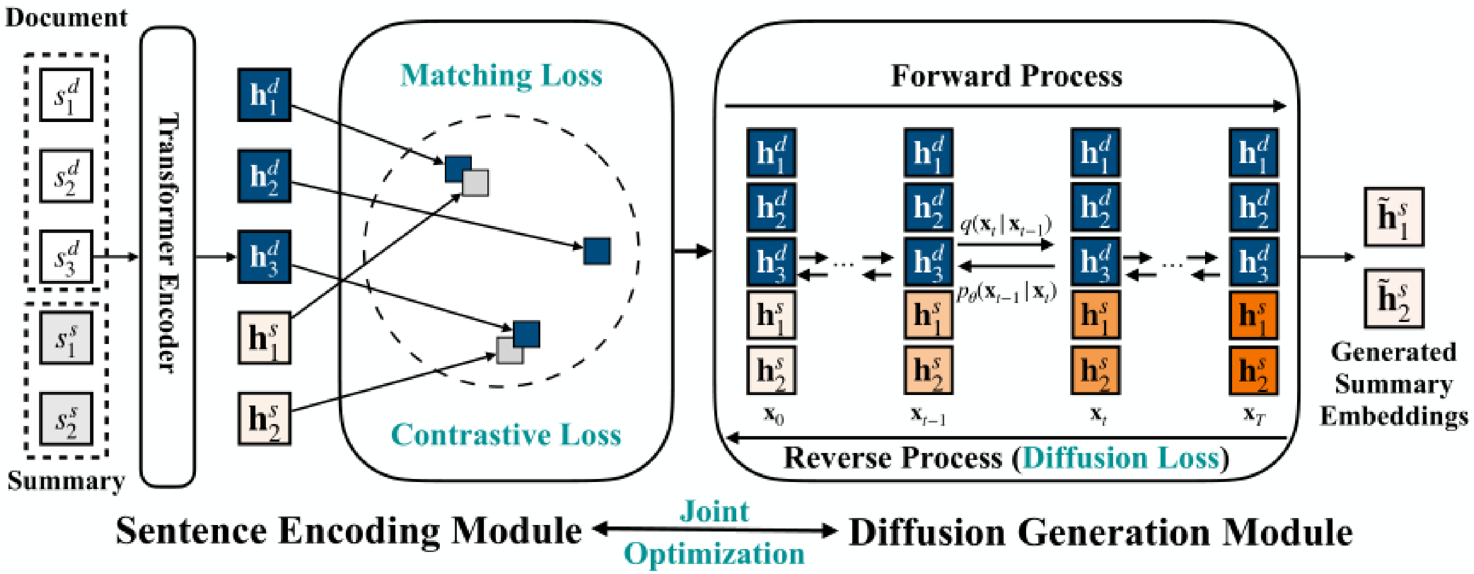


Figure 2: The overall architecture of DiffuSum. The input document is passed to the sentence encoding module and the diffusion generation module. DiffuSum will generate the desired summary sentence representations for inference.

Approach

$$\mathbf{x}_0 \sim q(x) \qquad \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \qquad \beta_t \in (0,1)$$

$$D = \{s_1^d, s_2^d, \ldots, s_n^d\} \qquad S = \{s_1^s, s_2^s, \ldots, s_m^s\}$$



$$q(\mathbf{x}_{t+1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t+1}; \sqrt{1-\beta_t}\mathbf{x}_t, \beta_t \mathbf{I}\right), \quad (1)$$

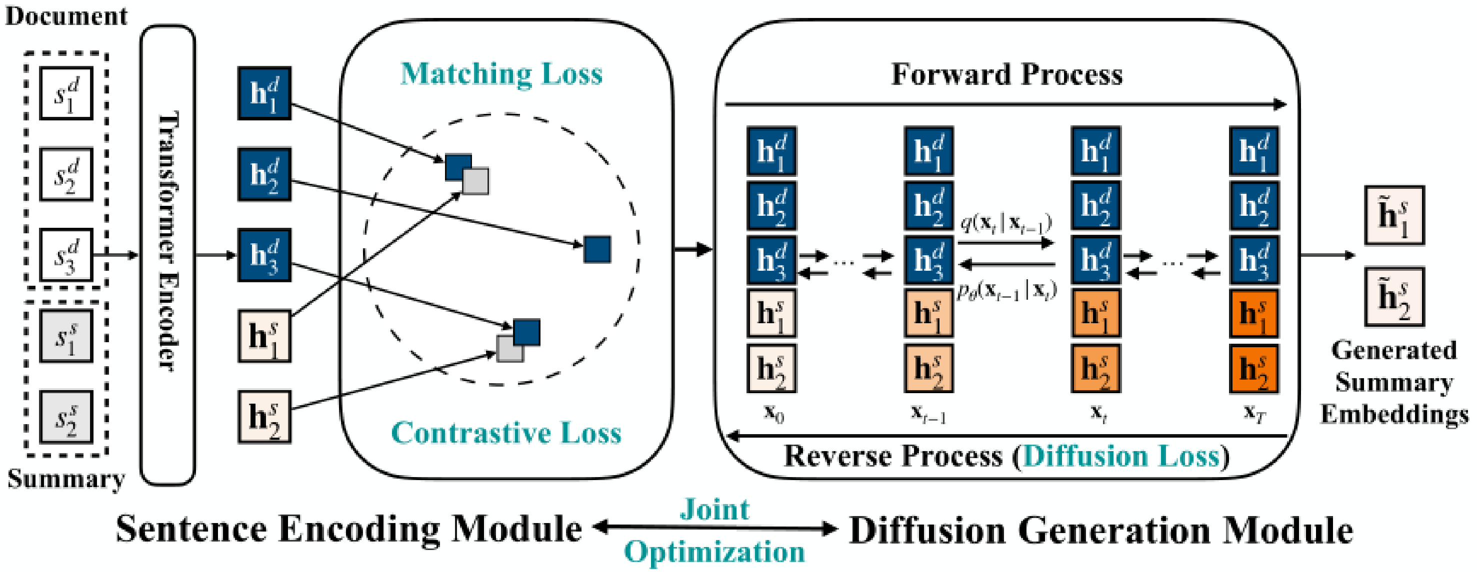
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta^2(t)\mathbf{I}\right), \quad (2)$$

$$\mathcal{L}_{\text{simple}} = \sum_{t=1}^{T} \left\| \mathbf{x}_0 - \tilde{f}_\theta(\mathbf{x}_t, t) \right\|^2, \quad (3)$$

$\tilde{f}_\theta(\mathbf{x}_t, t)$ is the reconstructed $\mathbf{x}_0$ at step t

$$\tilde{\mathbf{y}}_j = \text{softmax}(\tilde{\mathbf{h}}_j^s \cdot \mathbf{H}^{d^T}). \quad (4)$$

Approach



$$D = \{s_1^d, s_2^d, ..., s_n^d\}$$

$$\mathbf{E}^d = [\mathbf{e}_1^d, \mathbf{e}_2^d, ..., \mathbf{e}_n^d]$$

$$\mathbf{h}_i^d = \mathrm{MLP}(\mathrm{Transformer}(\mathbf{e}_i^d)). \quad (5)$$

$$\mathbf{H}^s = [\mathbf{h}_1^s, \mathbf{h}_2^s, ..., \mathbf{h}_m^s] \in \mathbb{R}^{m \times h}$$

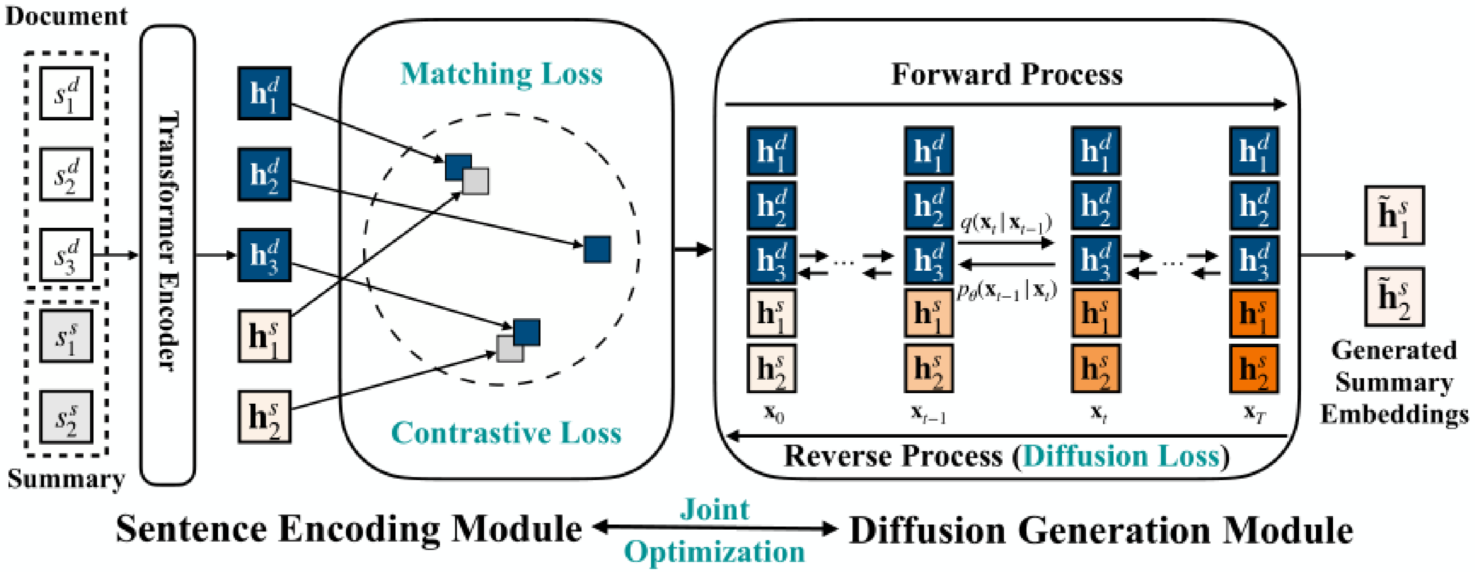
$$\hat{\mathbf{y}}_j = \mathrm{softmax}(\mathbf{h}_j^s \cdot \mathbf{H}^{d^{\mathrm{T}}}). \quad (6)$$

$$\mathcal{L}_{\mathrm{match}} = \sum_{j=1}^{m} \mathrm{CrossEntropy}\left(\mathbf{y}_j, \hat{\mathbf{y}}_j\right). \quad (7)$$

$$\mathbf{H}^{in} = \mathbf{H}^d \| \mathbf{H}^s \in \mathbb{R}^{(n+m) \times h}$$

$$\mathbf{H}^{in} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{n+m}] \in \mathbb{R}^{(n+m) \times h}$$

Approach



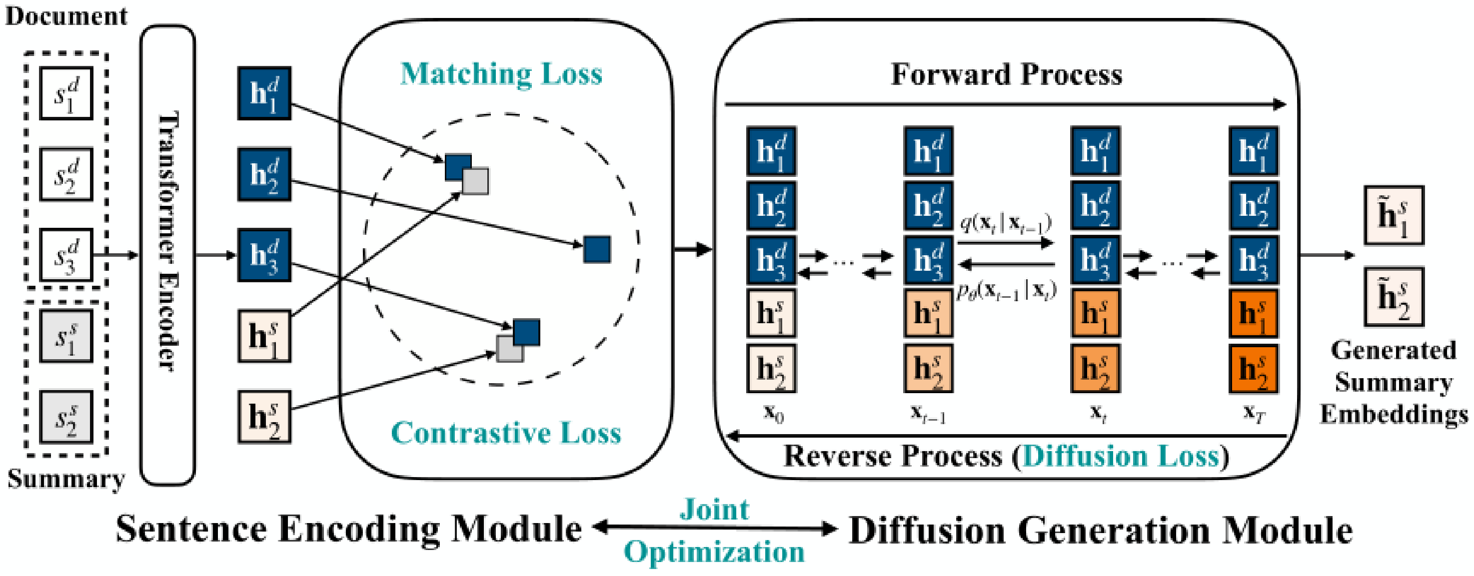
$$y_p^c = \begin{cases} q, & \text{if } p \le n \text{ and } s_p^d = s_q^s \\ q, & \text{if } p = n+q \\ 0, & \text{otherwise} \end{cases}, \quad (8)$$

$$\mathcal{L}_{\text{contra}} = \frac{-1}{2N_{y_p^c} - 1} \sum_{p=1}^{n+m} \mathcal{L}_{\text{contra}}^p,$$

$$\mathcal{L}_{\text{contra}}^p = \sum_{\substack{q=1;q\neq p;\\ y_q^c = y_p^c}}^{n+m} \log \frac{\exp\left(\mathbf{h}_p \cdot \mathbf{h}_q^T / \tau\right)}{\sum_{k=1;p\neq k}^{n+m} \exp\left(\mathbf{h}_p \cdot \mathbf{h}_k^T / \tau\right)}, \quad (9)$$

$$\mathcal{L}_{\text{se}} = \mathcal{L}_{\text{match}} + \gamma \mathcal{L}_{\text{contra}}, \quad (10)$$

Approach



$$q(\mathbf{x}_0|\mathbf{H}^{in}) = \mathcal{N}\left(\mathbf{H}^{in}, \beta_0\mathbf{I}\right)$$

$$\mathbf{x}_t = \mathbf{x}_0^d || \mathcal{N}\left(\mathbf{x}_t^s; \sqrt{1-\beta_t}\mathbf{x}_{t-1}^s, \beta_t\mathbf{I}\right), \quad (11)$$

the noised representations is $\mathbf{x}_t$

$$p_\theta\left(\mathbf{x}_{t-1}^s|\mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}^s; \mu_\theta\left(\mathbf{x}_t, t\right), \sigma_\theta^2(t)\mathbf{I}\right), \quad (12)$$

$$\mathcal{L}_{\text{diffusion}} = \sum_{t=2}^{T} \left\|\mathbf{x}_0 - \tilde{f}_\theta\left(\mathbf{x}_t, t\right)\right\|^2 + \left\|\mathbf{H}^{in} - \tilde{f}_\theta\left(\mathbf{x}_1, 1\right)\right\|^2 + \mathcal{R}\left(\mathbf{x}_0\right), \quad (13)$$

$\tilde{f}_\theta\left(\mathbf{x}_t, t\right)$ is the reconstructed $\mathbf{x}_0$ at step t

$$\mathcal{L} = \mathcal{L}_{\text{se}} + \eta\mathcal{L}_{\text{diffusion}} \quad (14)$$

Experiments

Dataset	Domain	Doc #words	Sum #words	#Ext
CNN/DM	News	766.1	58.2	3
XSum	News	430.2	23.3	2
PubMed	Paper	444	209.5	6

Table 1: Statistics of the experimental datasets. Doc # words and Sum # words refer to the average word number in the source document and summary. # Ext refers to the number of sentences to extract.

Experiments

Model	R-1	R-2	R-L
LEAD	40.43	17.62	36.67
ORACLE	52.59	31.23	48.87
One-stage Systems			
Transformer (2017)	40.90	18.02	37.17
HIBERT* (2019)	42.37	19.95	38.83
PNBERT* (2019)	42.69	19.60	38.85
BERTEXT* (2019)	42.76	19.87	39.11
BERTSum* (2019)	43.85	20.34	39.90
$\text{COLO}^{*}_{\text{Ext}}$ (2023)	44.58	21.25	40.65
DiffuSum (ours)	**44.62**	**22.51**	**40.34**
Two-stage Systems			
MATCHSUM*(BERT) (2020)	44.22	20.62	40.38
MATCHSUM*(Roberta)	44.41	20.86	40.55
DiffuSum (ours)	**44.83**	**22.56**	**40.56**

Table 2: Experimental results on CNN/DailyMail dataset. Models using pre-trained language models are marked with*.

Experiments

Model	PubMed			XSum		
	R-1	R-2	R-L	R-1	R-2	R-L
ORACLE	45.12	20.33	40.19	25.62	7.62	18.72
LEAD	37.58	12.22	33.44	14.40	1.46	10.59
BERTSUM	41.05	14.88	36.57	22.86	4.48	17.16
MatchSUM	41.21	14.91	36.75	**24.86**	4.66	**18.41**
DiffuSum	**41.40**	**15.55**	**37.48**	24.00	**5.44**	18.01

Table 3: Experimental Results on PubMed and XSum datasets.

Experiments

Model	R-1	R-2	R-L
DiffuSum	**44.83**	**22.56**	**40.56**
w/o Sentence-BERT	43.53	21.63	40.23
w/o ORACLE	39.19	17.12	34.38
w/o Contrastive Loss	44.57	22.35	40.34

Table 4: Ablation study results on CNN/DailyMail dataset.

Experiments

Model	R-1	R-2	R-L
DiffuSum(T=500, h=128)	**44.83**	**22.56**	**40.58**
DiffuSum(T=500, h=64)	43.36	21.27	39.89
DiffuSum(T=500, h=256)	44.53	22.49	40.27
DiffuSum(T=50, h=128)	42.60	19.71	38.96
DiffuSum(T=100, h=128)	44.61	22.24	40.32
DiffuSum(T=1000, h=128)	44.65	22.36	40.37
DiffuSum(T=2000, h=128)	44.64	22.37	40.40

Table 5: The performance of DiffuSum with different hyperparameter settings on CNN/DM dataset.

Experiments

Train \ Test	CNN/DM	XSum	PubMed
CNN/DM	44.83/22.56	21.35/3.85	39.83(-1.57)/13.25
XSum	42.85/21.37	24.0/5.44	38.71(-2.69)/12.93

Table 6: ROUGE-1 and ROUGE-2 results for cross-dataset evaluation.

Experiments

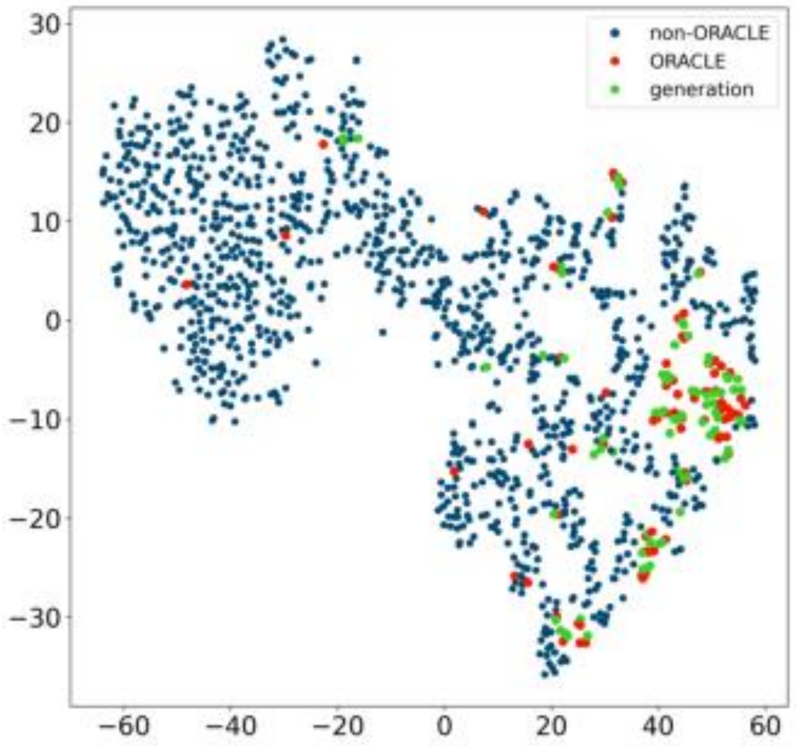


Figure 3: T-SNE visualization of sentence embeddings from 25 CNN/DM dataset documents.

Thank you！