



# Hierarchical Attention Networks for Sentence Ordering

Tianming Wang & Xiaojun Wan

*{wangtm, wanxiaojun}@pku.edu.cn*

Institute of Computer Science and Technology, Peking University  
Beijing , China

# Outline



- Introduction
- Prior methods
- Our approach
- Experiment & Result

# Outline



- Introduction
- Prior methods
- Our approach
- Experiment & Result



# Introduction

- What's Sentence Ordering?
  - organize a given set of sentences into a coherent text in a clear and consistent manner
  - learn which ordering of sentences is likely to enhance understanding and avoid confusion
- Why do this?
  - modeling discourse coherence
  - help downstream tasks
    - multi-doc summarization
    - question answering
    - text planning

# Outline

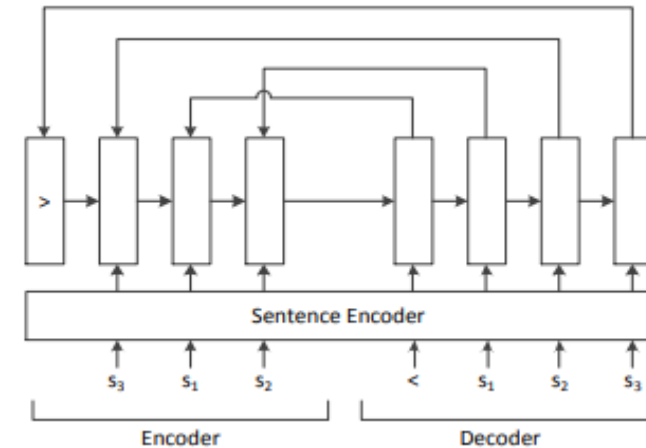
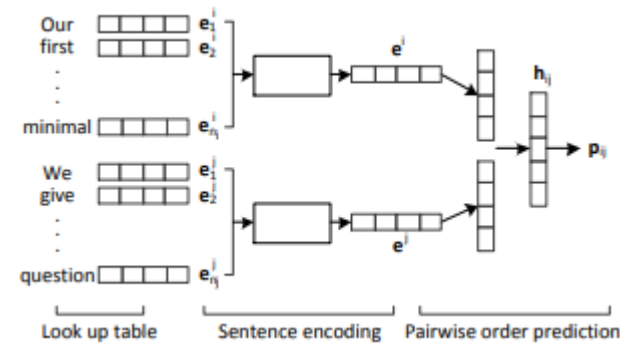


- Introduction
- Prior methods
- Our approach
- Experiment & Result

# Prior methods



- Pair-wise
  - Classification
    - only consider local coherence
    - no contextual information
- End-to-end
  - RNN based pointer network
    - treats the out-of-order set of sentences as a sequential input



# Outline

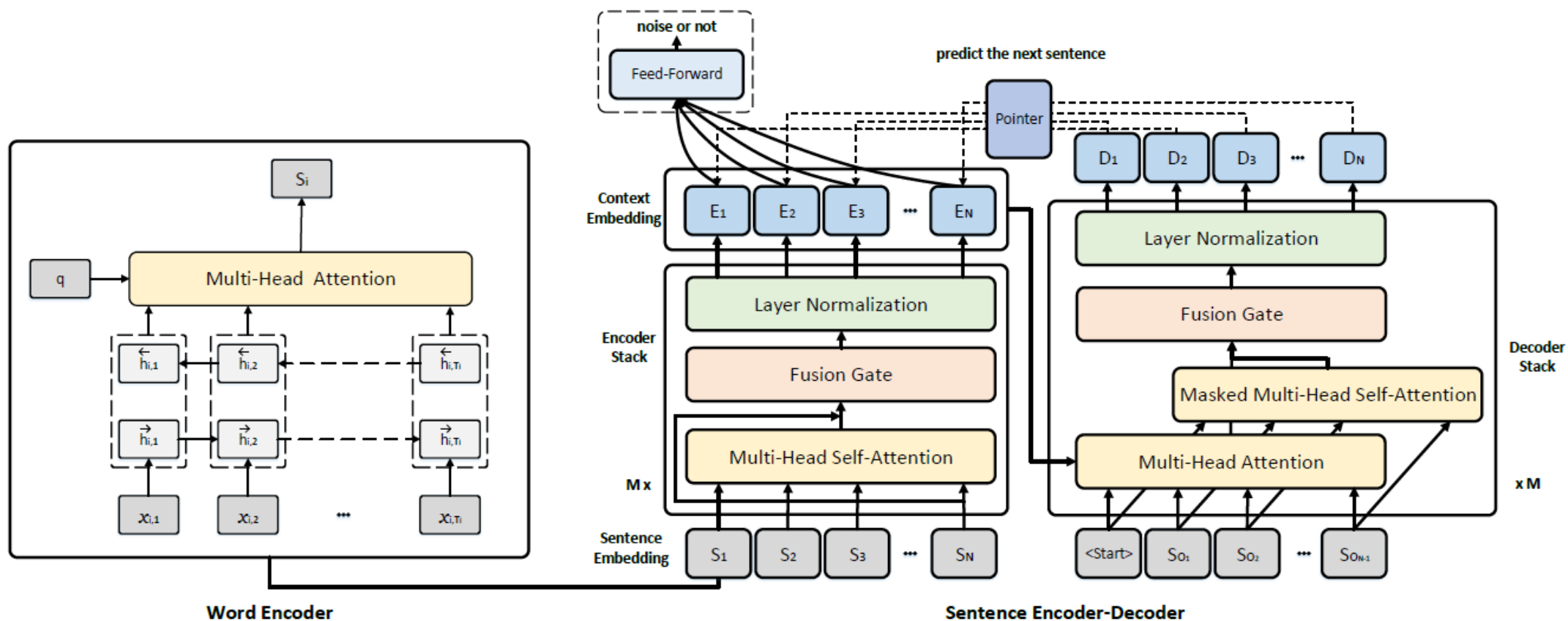


- Introduction
- Prior methods
- Our approach
- Experiment & Result

# Our approach



- Hierarchical attention networks







# Word attention

- Word clues
  - keywords like “first” and “then” provide clues for ordering

- Multi-head attention(Transformer)

$$\hat{Q}^j = \text{ReLU}(QW_Q^j)$$

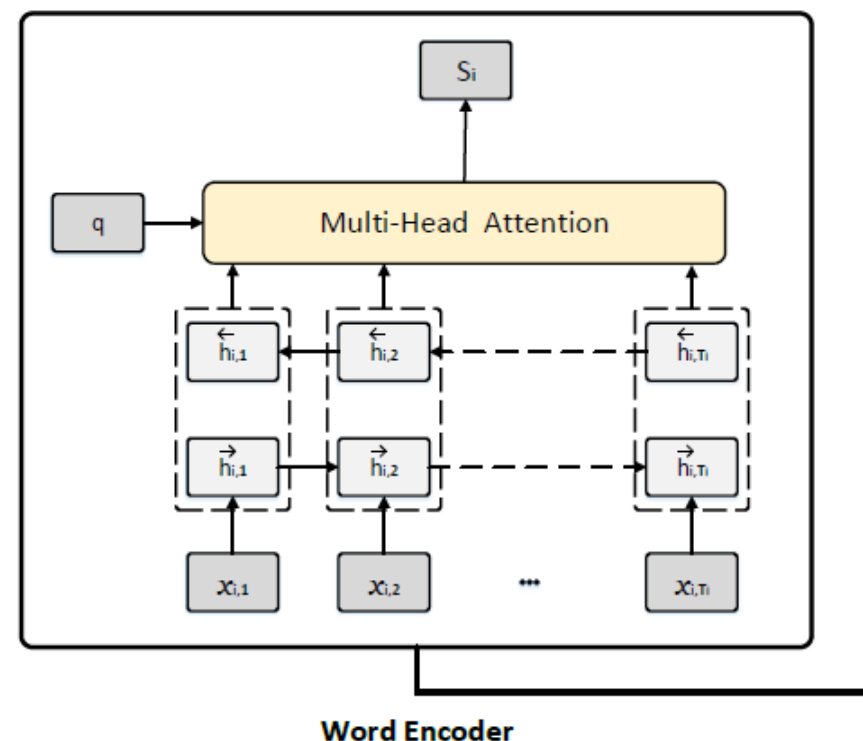
$$\hat{K}^j = \text{ReLU}(KW_K^j)$$

$$\hat{V}^j = \text{ReLU}(VW_V^j)$$

$$\hat{C}^j = \text{softmax}\left(\frac{\hat{Q}^j \hat{K}^{j\top}}{\sqrt{d/H}}\right) \hat{V}^j$$

$$C = [\hat{C}^1; \hat{C}^2; \dots; \hat{C}^H]$$

where  $Q, K, V$  are the packages of a set of queries, keys and values,  $W_Q^j, W_K^j, W_V^j \in \mathbb{R}^{d \times d/H}$  are parameter matrices,





# Sentence encoder

- Recode sentence
  - capture global dependencies
  - adjust representations of sentences in context
- Encoder stack

$$E_{in}^j = E_{out}^{j-1}$$

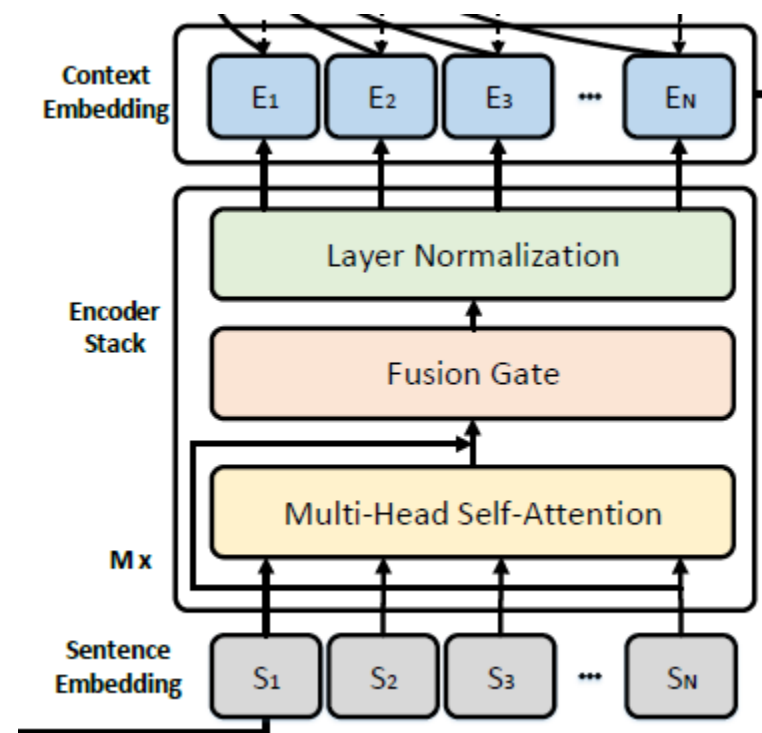
$$C = MultiHead(E_{in}^j, E_{in}^j, E_{in}^j)$$

$$G = sigmoid(E_{in}^j W_{in}^j + C W_{out}^j)$$

$$F = G E_{in}^j + (1 - G) C$$

$$E_{out}^j = LayerNorm(F)$$

where  $W_{in}^j, W_{out}^j \in \mathbb{R}^{d \times 1}$  are parameter matrices

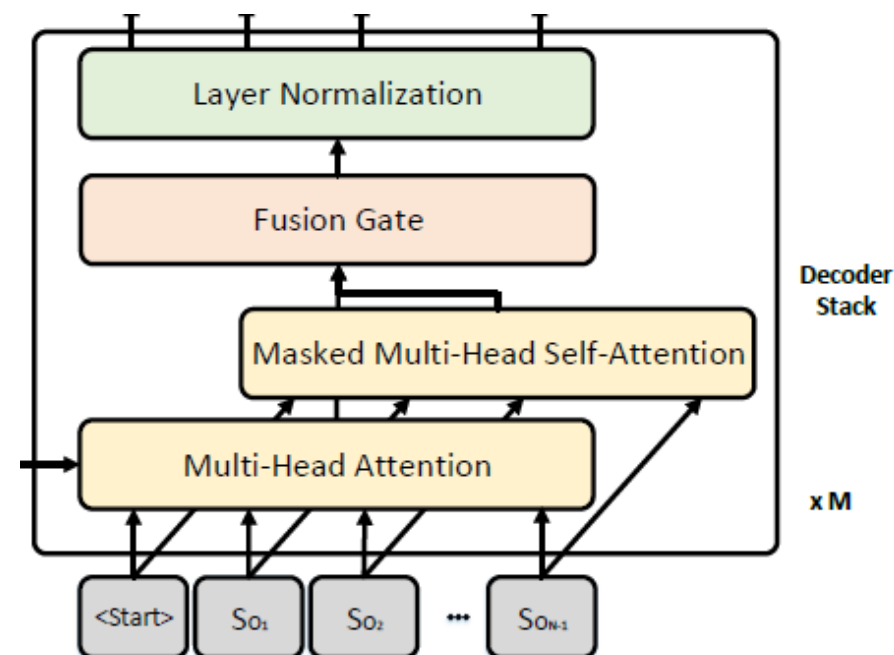




# Sentence Decoder

- Masked multi-head attention
  - prevent earlier decoding steps from accessing information from later steps
  - utilize the information of ordered subsequence and construct a context for predicting the next sentence

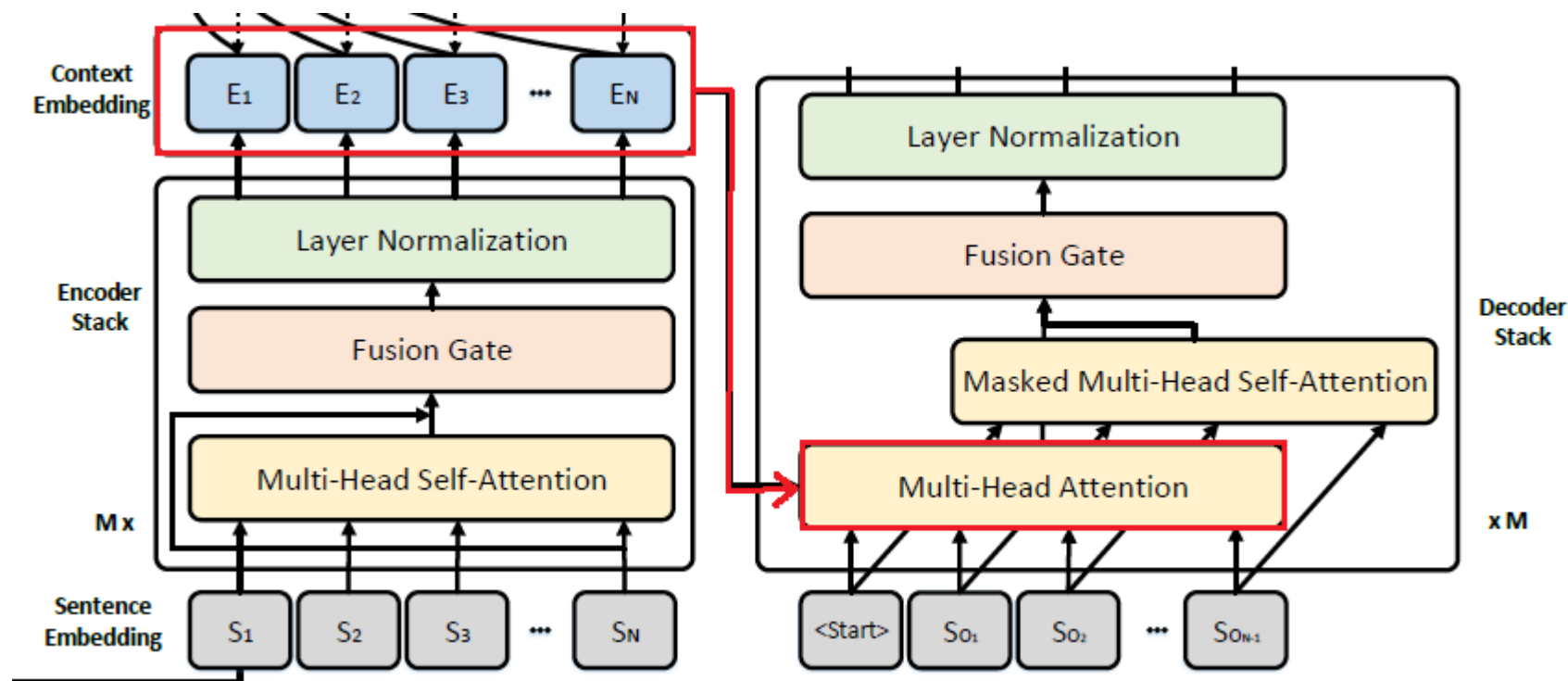
$$Mask_{x,y} = \begin{cases} 0 & x \leq y \\ -\infty & otherwise \end{cases}$$
$$\hat{C}^j = softmax(\frac{\hat{Q}^j \hat{K}^{j\top} + Mask}{\sqrt{d/H}}) \hat{V}^j$$
$$C = [\hat{C}^1; \hat{C}^2; \dots; \hat{C}^H]$$





# Sentence Decoder

- global attention
  - utilizes the global information captured by the encoder





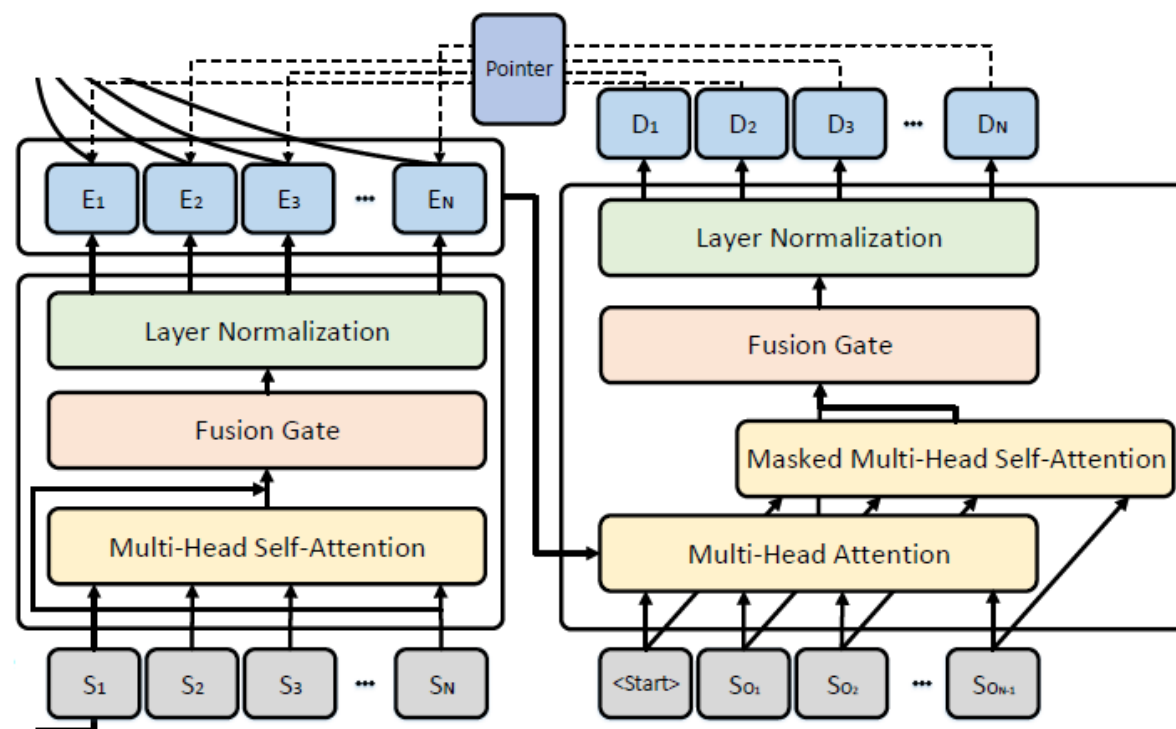
# Sentence Decoder

- Pointer
  - predicting the next sentence
- Inference
  - beam search

$$Q = \text{ReLU}(D_{out}^M W_Q)$$

$$K = \text{ReLU}(E_{out}^M W_K)$$

$$P = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$



# Outline



- Introduction
- Prior methods
- Our approach
- Experiment & Result

# Experiment



- Dataset

- arXiv

- 884912 training abstracts, 110614 validation abstracts and 110615 testing abstracts of papers on arXiv website
    - composed of 2 to 20 sentences

- VIST

- 40155 training stories, 4990 validation stories and 5055 testing stories
    - composed of 5 sentences

- ROCStory

- 78529 training stories, 9816 validation stories and 9817 testing stories
    - composed of 5 sentences



# Experiment

- Metrics

- Kendall's tau

$$\tau = 1 - \frac{2(InvertPairs)}{N(N-1)/2}$$

where  $N$  is the number of sentences being ordered and  $(InvertPairs)$  is the number of interchanges of consecutive elements necessary to arrange them in their natural order.

- Perfect Match Ratio

- the ratio of cases of exact match of the whole sequence



# Result



Table 1: Comparison of results on three datasets

Methods	arXiv		VIST		ROCStory	
	$\tau$	PMR	$\tau$	PMR	$\tau$	PMR
random	0	0.0827	0	0.0083	0	0.0083
LSTM+Pairwise (Chen, Qiu, and Huang 2016)	0.6594	0.3343	-	-	-	-
SkipThought+Pairwise (Agrawal et al. 2016)	-	-	0.4640	-	-	-
LSTM+PtrNet (Gong et al. 2016)	0.7158	0.4044	0.4842	0.1234	-	-
Seq2Seq+Pairwise (Li and Jurafsky 2017)	0.0593	0.1370	0.1892	0.1250	0.3419	0.1793
LSTM+Set2Seq (Logeswaran, Lee, and Radev 2018)	0.7281	0.4157	0.4919	0.1380	0.7112	0.3581
WordAtt+PtrNet	0.7367	0.4210	0.4925	0.1346	0.7024	0.3285
Our	<b>0.7536</b>	<b>0.4455</b>	<b>0.5021</b>	<b>0.1501</b>	<b>0.7322</b>	<b>0.3962</b>

- Parameter study on the arXiv dataset

	M	H	$P_{drop}$	WordAtt	$\tau$	PMR
base	3	4	0.05	yes	<b>0.7536</b>	<b>0.4455</b>
(A)	2				0.7484	0.4419
	4				0.7515	0.4409
(B)		1			0.7437	0.4368
		2			0.7496	0.4420
		8			0.7481	0.4407
(C)			0		0.7475	0.4413
			0.1		0.7526	0.4442
(D)				no	0.7399	0.4301



# Experiment with noise

- Noise
  - Randomly chosen from another abstract or story
  - Why?
    - to test the robustness and effectiveness
    - For example, if the model arranging sentence only according to words clues like “First” and “Then”, we can add a noisy sentence like “First, we...”, which is irrelevant to the abstract to cheat the model
- Strategy
  - add 1 noisy sentence (1 noise)
  - add 1 noisy sentence with a probability of 50% (0/1 noise)



# Experiment with noise

- Discrimination

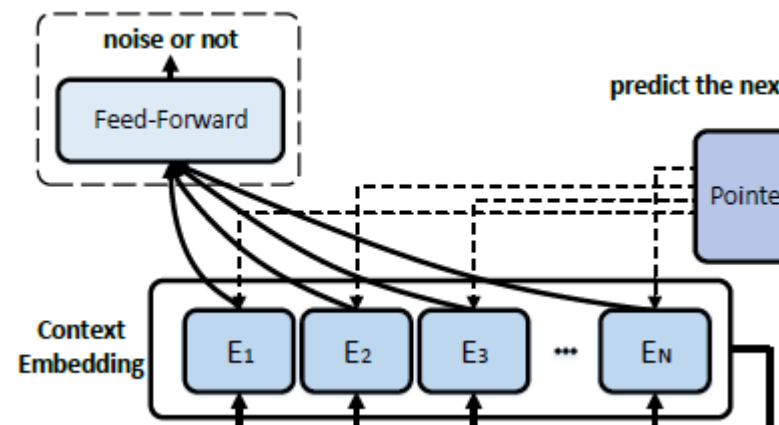
Table 3: Performance of noise discrimination.

Strategy	Methods	arXiv	VIST	ROCStory
		acc	acc	acc
1 noise	random	0.1819	0.1667	0.1667
	Our	0.9664	0.8462	0.9382
0/1 noise	random	0.2955	0.5833	0.5833
	Our	0.9330	0.9151	0.9698

- Sentence ordering

Table 4: Performance of sentence ordering on three datasets with noise.

Strategy	Methods	arXiv		VIST		ROCStory	
		$PM_F$	PMR	$PM_F$	PMR	$PM_F$	PMR
0 noise	random	0.5000	0.0827	0.5000	0.0083	0.5000	0.0083
	LSTM+PtrNet (Gong et al. 2016)	0.8579	0.4044	-	-	-	-
	WordAtt+PtrNet	0.8683	0.4210	0.7463	0.1346	0.8512	0.3285
	Our	<b>0.8768</b>	<b>0.4455</b>	<b>0.7510</b>	<b>0.1520</b>	<b>0.8661</b>	<b>0.3962</b>
1 noise	random	0.3178	0.0238	0.3357	0.0011	0.3326	0.0010
	LSTM+PtrNet (Gong et al. 2016)	0.8228	0.3733	-	-	-	-
	WordAtt+PtrNet	0.8271	0.3805	0.6432	0.0980	0.7852	0.2930
	Our	<b>0.8586</b>	<b>0.4325</b>	<b>0.6992</b>	<b>0.1283</b>	<b>0.8376</b>	<b>0.3883</b>
0/1 noise	random	0.3830	0.0259	0.4096	0.0049	0.4227	0.0069
	LSTM+PtrNet (Gong et al. 2016)	0.8344	0.3675	-	-	-	-
	WordAtt+PtrNet	0.8407	0.3740	0.6706	0.1064	0.7967	0.3055
	Our	<b>0.8516</b>	<b>0.4094</b>	<b>0.6974</b>	<b>0.1300</b>	<b>0.8293</b>	<b>0.3879</b>





# Visualization

- Word clues

- Darker shades correspond to higher attention weights

$$\alpha = \frac{1}{H} \sum_{j=1}^H \text{softmax}\left(\frac{\hat{Q}^j \hat{K}^{j\top}}{\sqrt{d/H}}\right)$$

In this paper some important inequalities are revisited .

**First** , as motivation , we give another proof of the Hardy 's inequality applying convenient vector fields as introduced by Mitidieri , see [6] .

**Then** , we investigate a particular case of the Caffarelli-Kohn-Nirenberg 's inequality .

**Finally** , we study the Rellic 's inequality .

Wireless microsensor networks , which have been the topic of intensive research in recent years , **are now** emerging in industrial applications .

An important milestone in **this transition** has **been** the release of the IEEE 802.15.4 standard that specifies interoperable wireless physical and medium access control layers targeted to sensor node radios .

In this paper , we **evaluate** the potential of an 802.15.4 radio for use in an ultra low power sensor node operating in a dense network .

**Starting** from measurements carried out on the off-the-shelf radio , effective radio activation and link adaptation policies **are derived** .

It is **shown that** , in a **typical** sensor network scenario , the average power per node can be reduced down to 211m mm mW .

**Next** , the energy consumption breakdown between the different phases of a packet transmission is **presented** , **indicating** which part of the transceiver architecture can most effectively be optimized in order to further reduce the radio power , **enabling** self-powered wireless microsensor networks .

Jimmy **needed** to break up with his girlfriend . He **drove** to her house and **knocked** on her door . She **answered** a minute **later** and they began to talk . Jimmy **told** her the bad news and she began to cry . Jimmy **left** the scene and felt very bad about himself .

Thanks

