

Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心



基于图结构处理的文本生成

基于图结构处理的文本生成

硕士研究生 张凌浩

2022年2月27日

- 背景简介
- 发展历史
- 基本概念
- 算法原理
 - Graph Writer
 - Graph2seq
- 应用总结
- 参考文献

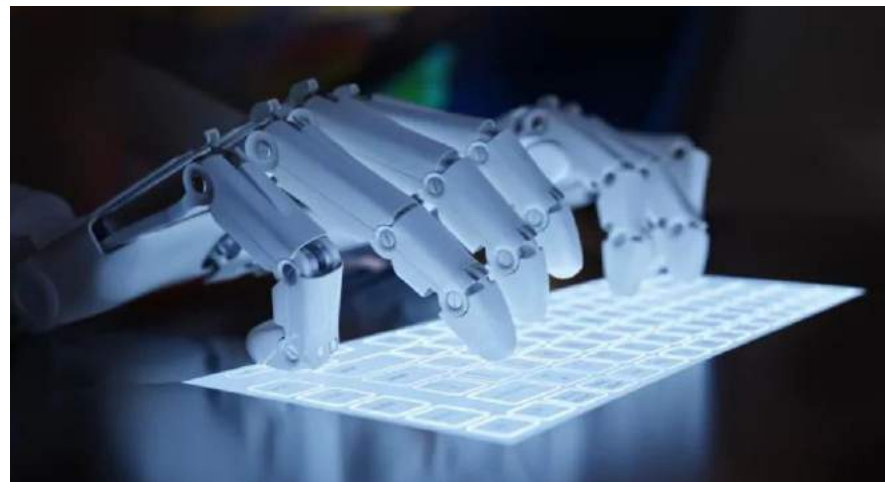
• 预期收获

- 1.熟悉**文本生成**的基本概念
- 2.简单了解**传统的sequence-to-sequence模型和Attention机制**
- 3.理解两种**Graph-to-sequence**的代表算法
- 4.开拓对**图结构数据**的构造与处理思路

- **文本生成**是自然语言处理中一个重要的研究领域，具有广阔的应用前景。

- 机器翻译 (Machine Translation)
- 文本摘要 (Summarization)
- 智能问答 (Dialogue)
- 创意写作 (Creative writing)
- 自动评论 (Comment)

...



- 国内外已经有诸如Automated Insights、Narrative Science以及“小冰”机器人等文本生成系统投入使用。这些系统根据**格式化数据**或**自然语言文本**生成新闻、财报或者其他解释性文本

基于图结构处理的文本生成



历史发展

- 文本生成

- 基于规则

- 早期的文本生成系统多是**基于规则和模板**
 - 通过分析目标文本和**手工制定的规则**来定义内容规划
 - **泛用性低，只适用于特定的领域**

- 基于度量

- 利用文章统计信息，如词频和位置分布等信息，从候选词列表中选出生成词汇链
 - 以统计学为基础
 - 只是基于句子和单词本身的**表层联系**，没有充分利用**词义，语义**等信息

- **基于语义**

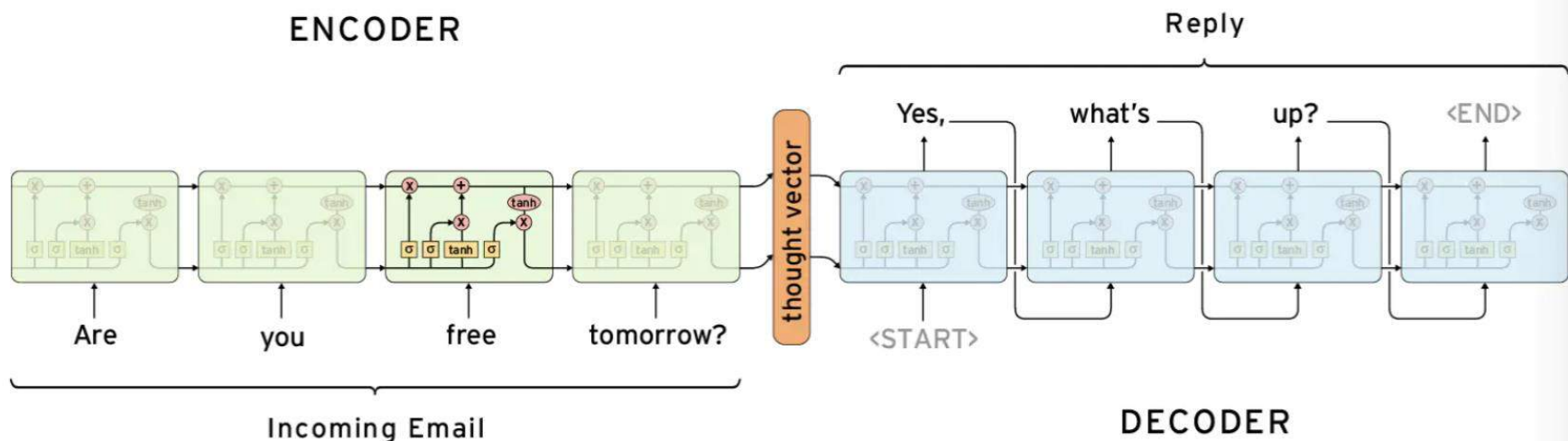
- 对原始语料深层语义信息的挖掘
 - Sequence-to-Sequence, **Graph-to-Sequence**....

基于图结构处理的文本生成

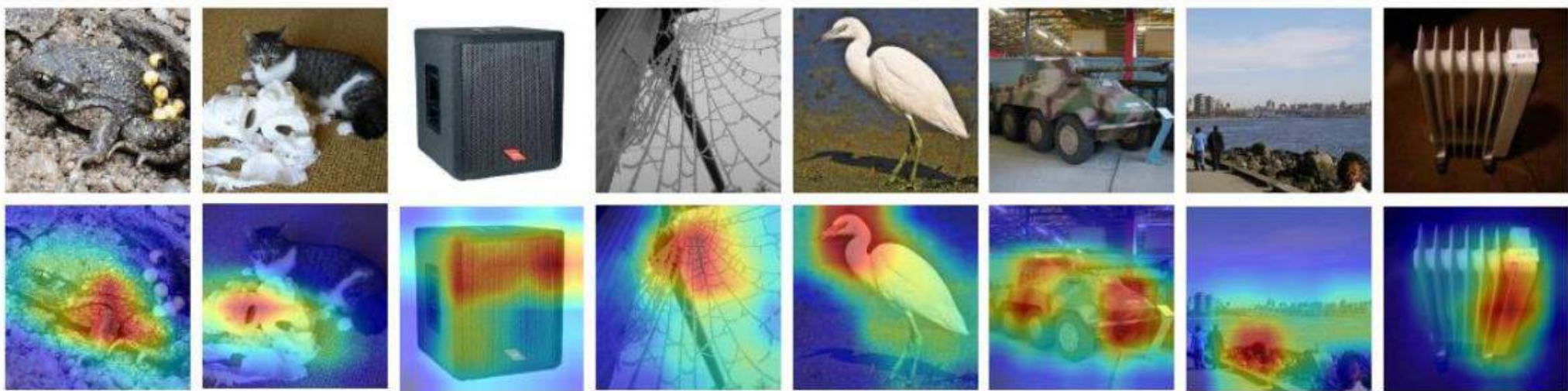


基本概念

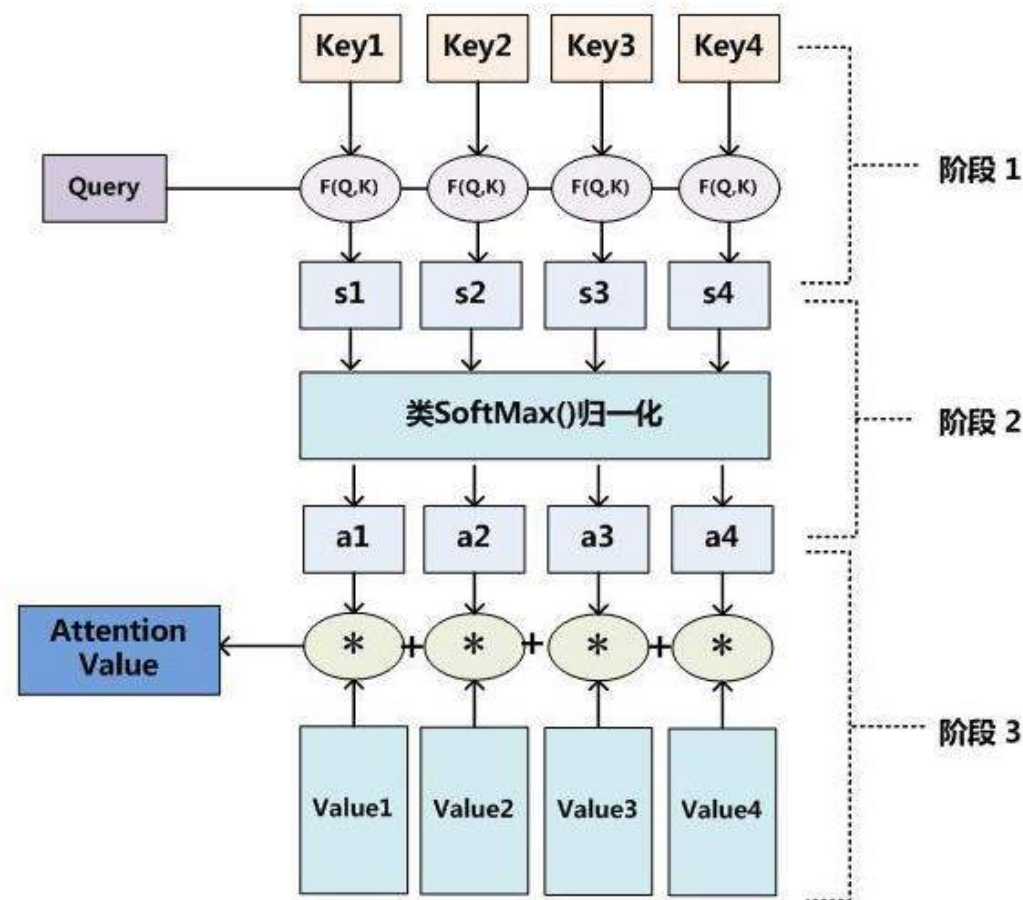
- Sequence-to-Sequence (seq2seq)
 - 大多数基于语义的文本生成都使用的是seq2seq结构
 - seq2seq是Encoder-Decoder 的经典结构，输入是一个序列，输出也是一个序列
 - Encoder 中将一个可变长度的信号序列变为固定长度的语义向量
 - Decoder 将这个固定长度的向量变成可变长度的目标的信号序列
 - 常见的有CNN / RNN / BiRNN / GRU / LSTM



- 传统seq2seq方法存在的主要问题
 - seq2seq框架先用Encoder将输入序列编码成一个统一的**固定大小的语义向量c**，这个过程是对信息压缩的过程，不可避免地会损失很多信息，而且Decoder在解码时无法关注到输入序列的更多细节。导致原始语料中丰富的语义信息得不到利用
- Attention机制
 - 核心目标定义为众多信息中选择出**对当前任务目标更关键的信息**，从而使机器能够更加高效地利用**有限的计算资源来捕捉特征信息**

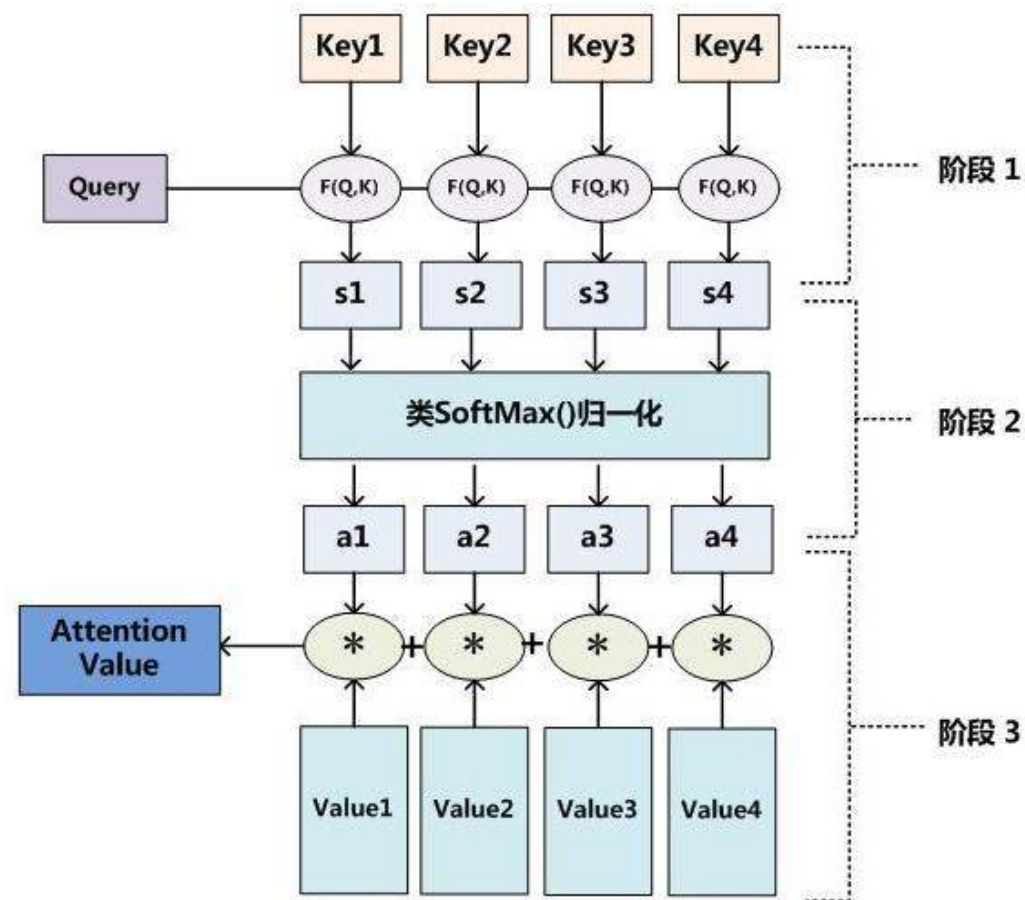


- Attention机制中的Query, Key, Value
 - Target: 我是中国人
 - Source: I am Chinese
 - 翻译目标单词为 I 的时候, Q为I
 - 对应的 “我” “是” “中国人” 都是K,
 - 那么Q就要与每一个K进行相似度计算
 - 相似度的值进行归一化后会生成对齐概率值, 称为注意力权重a
 - 然后相应的V与a进行加权求和, 就得到了当前Q的上下文向量



• 注意力权重计算函数

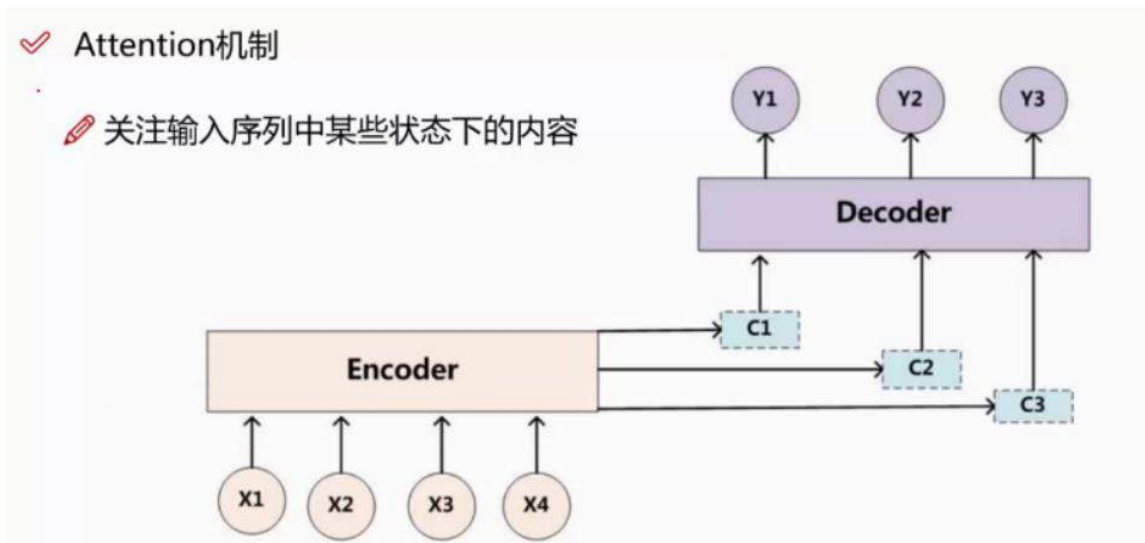
- 点乘: $f(Q, K_i) = Q^T K_i$
- 权重: $f(Q, K_i) = Q^T W K_i$
- 拼接权重: $f(Q, K_i) = W[Q; K_i]$
- 感知机: $f(Q, K_i) = V^T \tanh(WQ + UK_i)$



- Attention机制

- 优点:

- 1. **参数少**: 相比于 CNN、RNN，其复杂度更小，参数也更少
 - 2. **速度快**: Attention 解决了 RNN及其变体模型 不能并行计算的问题。Attention机制每一步计算不依赖于上一步的计算结果，因此可以和CNN一样并行处理
 - 3. **效果好**: Attention 机制引入加强了对长距离的信息的捕捉

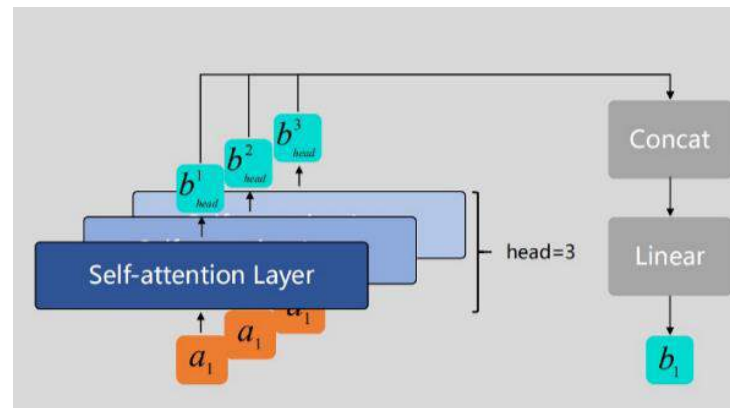
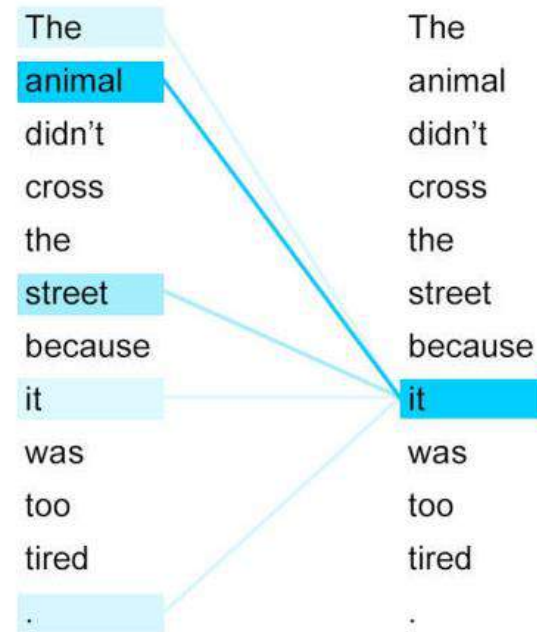


- Self Attention机制

- Self-Attention 顾名思义，指的是不是 Target 和 Source 之间的 Attention 机制，而是 Source 内部元素之间或者 Target 内部元素之间发生的 Attention 机制，也可以理解为 Target = Source 的特殊情况下的 Attention 机制

- 多头Attention机制

- 允许模型在**不同的表示子空间里学习**到相关的信息
- **Head**数通常取3
- 多头注意力的本质是多个独立的**Attention**计算，作为一个集成的作用，防止过拟合



- 信息抽取IE (Information Extraction)
 - 是从自然语言文本中抽取特定的**事件或事实信息**，**自动分类、提取和重构**。这些信息通常包括实体 (entity)、关系 (relation)、事件 (event)。例如从新闻中抽取时间、地点、关键人物
- 开放信息抽取 (OpenIE)
 - 抽取实体关系的**模式不需要事先指定**；通常，关系名称只是链接两个参数的文本。例如，Barack Obama was born in Hawaii 将会创建一个**三元组** (Barack Obama; was born in; Hawaii), 对应于开放域的关系为 was-born-in(Barack-Obama, Hawaii)
- 关键词抽取 (TextRank)
 - 基于PageRank，用于为文本生成关键词
 - 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序
 - 仅利用**单篇文档**信息即可，无需使用多篇文章进行学习

基于图结构处理的文本生成



算法原理

算法流程

T	图转换网络文本生成
I	原始文章与标题
P	1.将原始文章抽取的实体关系三元组构造为知识图谱 2.利用Graph Transformer编码器处理知识图谱
O	文本摘要

P	原始语料中文章的全局特征未被有效利用
C	科学领域文本
D	图具有异构性、长距离依赖易崩溃等特性
L	2019 NNACL

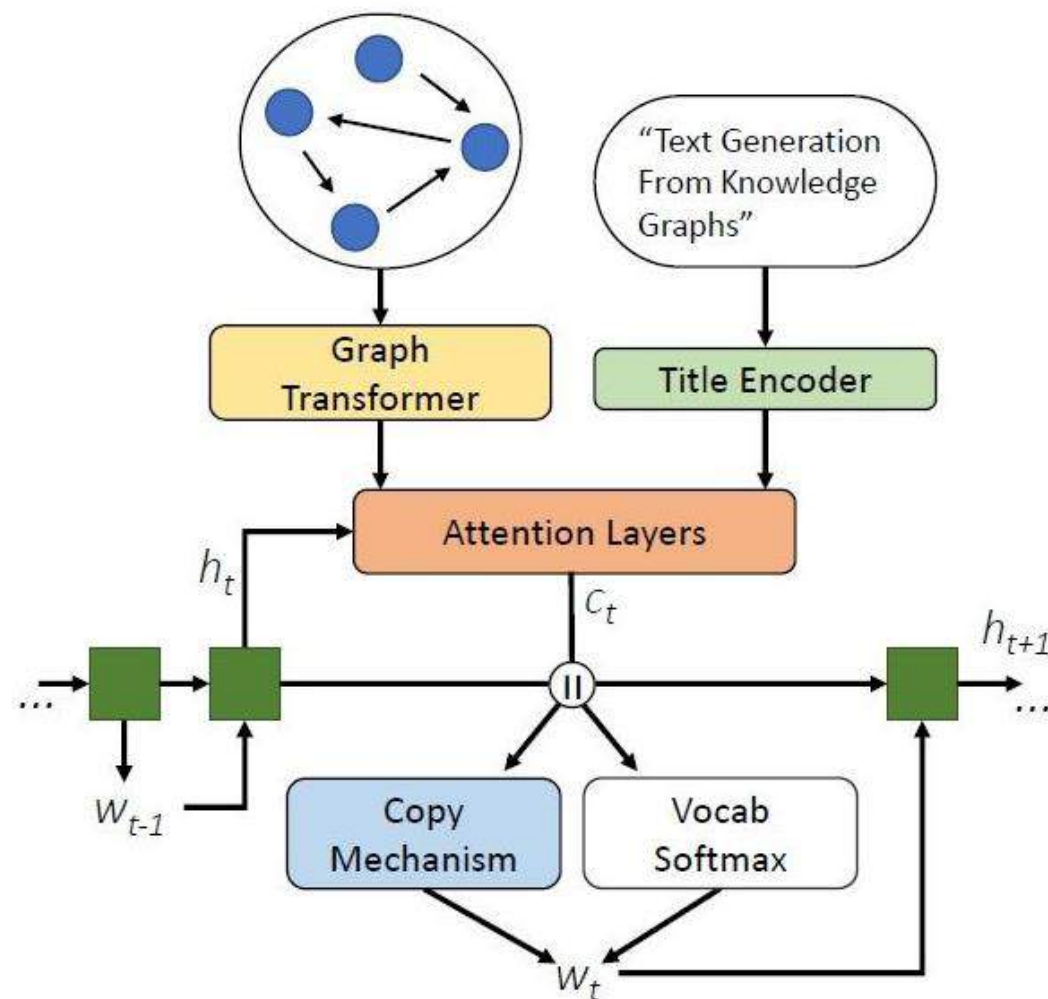
- 算法流程

- 知识图谱构建

- 图转换网络 Graph Transformers编码

- 标题编码

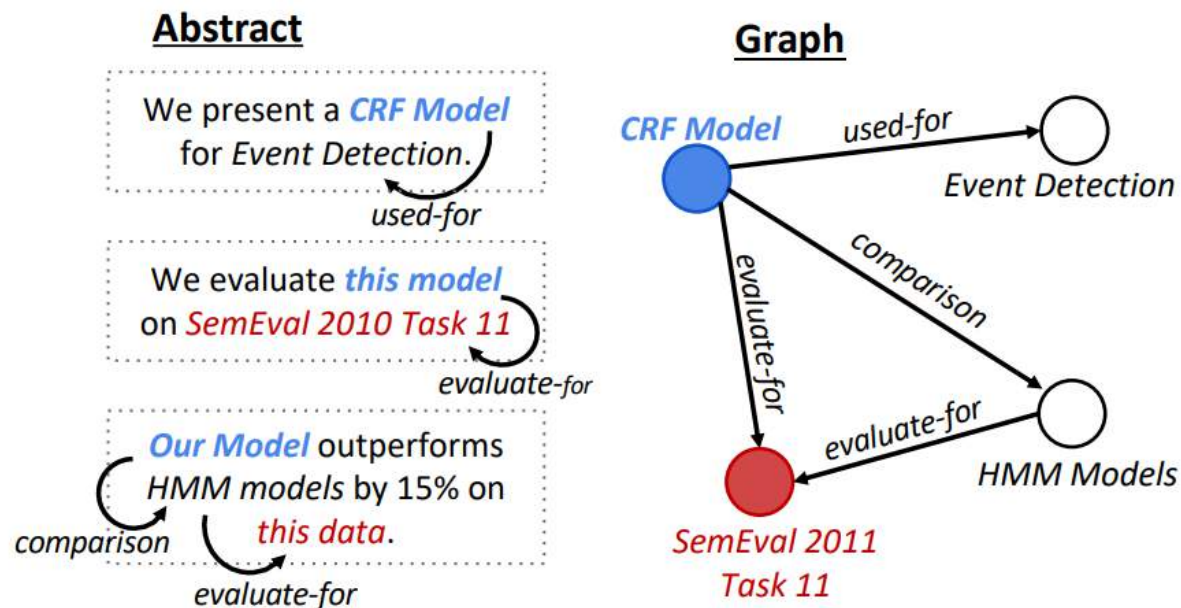
- 输出解码



- 知识图谱的构建

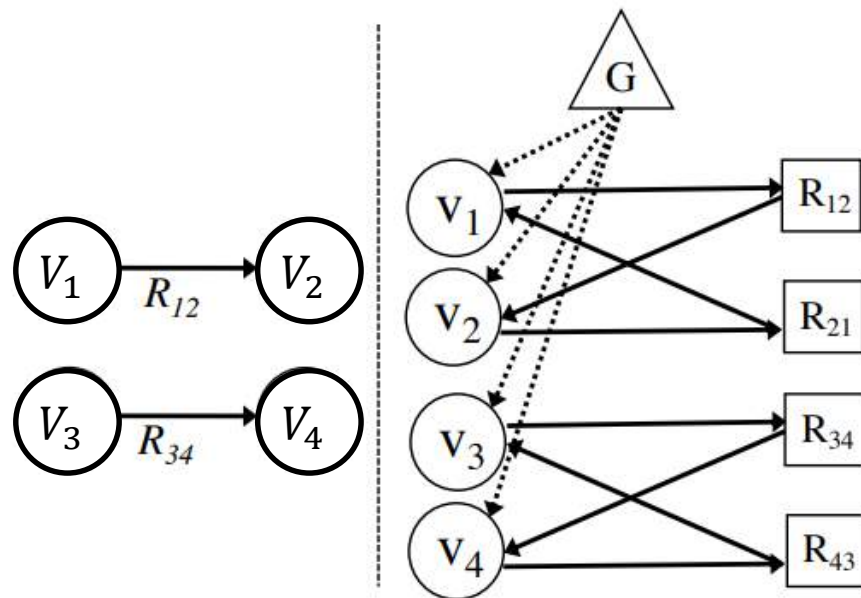
- 利用the SciIE system（科学领域的IE的sota）抽取原始文章中知识并建模成实体关系三元组。实体类型包括任务、方法、度量、材料或其他科学术语。该实体之间的关系被建模成以下7种：

- Compare
- Used-for
- Feature-of
- Hyponym-of
- Evaluate-for
- Conjunction



- 知识图谱的构建

- 将作为边的“关系”也构造为新的节点参与连接
- 构造一个全局节点G连接所有实体以促进断开部分之间的信息流动。同时这个全局节点将用于初始化解码器
- 最终形成 $G=(V, E)$ ，其中V是实体、关系和全局节点的列表，E是描述有向边的邻接矩阵



- 图转换网络Graph Transformer编码

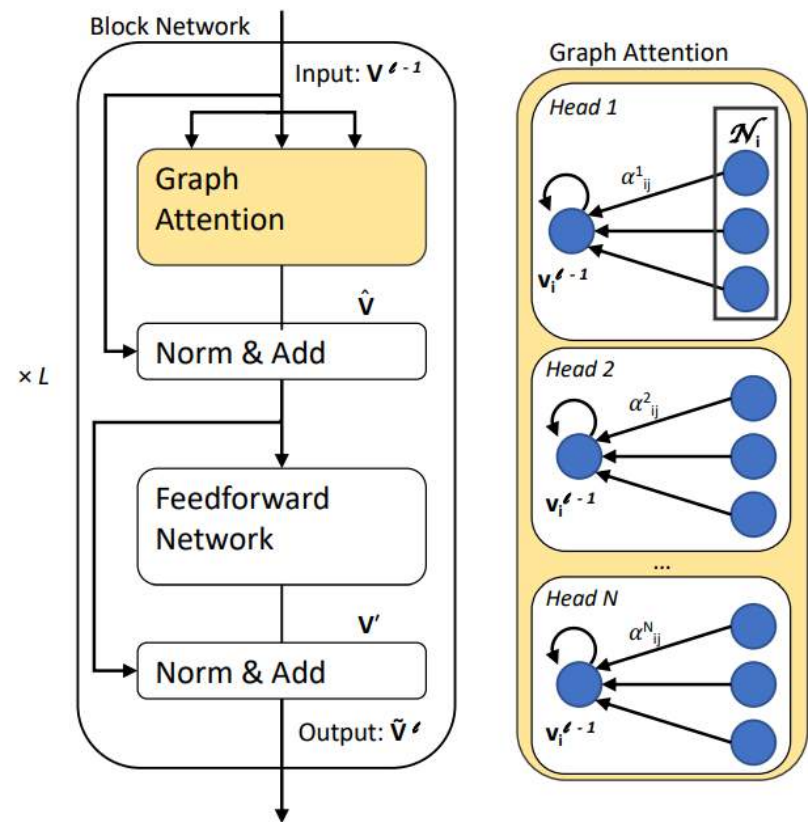
- 图G中所有顶点的集合作为输入

$$V^0 = [v_i], \quad v_i \in \mathbb{R}^d$$

- 每个顶点表示为 v_i ，通过关注 v_i 在G中连接到的其他顶点来进行语境化。用N个抽头的自注意层，N个独立注意层在计算和连接后参与到其他连接中

$$\hat{v}_i = v_i + \prod_{n=1}^N \sum_{j \in N_i} a_{ij}^n W_V^n v_j$$

$$a_{ij}^n = a^n(v_i, v_j)$$

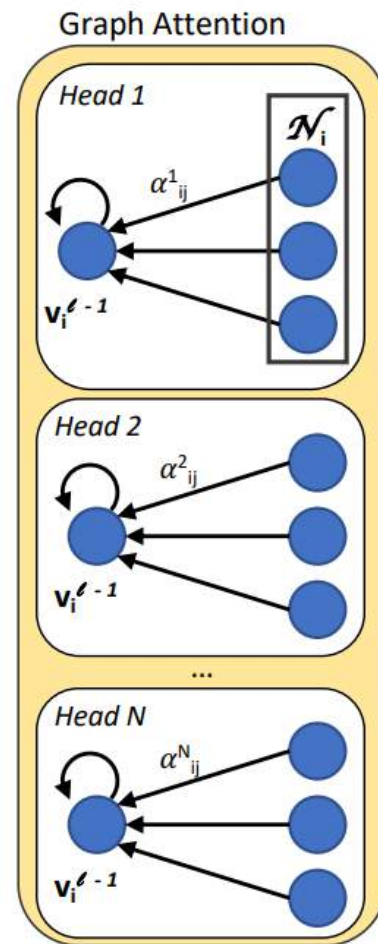


- 图转换网络Graph Transformer编码

- k表示N个注意头的串联， $\mathcal{N}_i$ 表示G中 v_i 的邻域， $w_v^n \in \mathbb{R}^{d \times d}$ ，其中 a^n 是每个注意头的**注意力权重**。使用以下形式的注意力函数

$$a(\mathbf{q}_i, \mathbf{k}_j) = \frac{\exp((\mathbf{W}_K \mathbf{k}_j)^\top \mathbf{W}_Q \mathbf{q}_i)}{\sum_{z \in \mathcal{N}_i} \exp((\mathbf{W}_K \mathbf{k}_z)^\top \mathbf{W}_Q \mathbf{q}_i)}$$

- 每个a学习k和q的独立变换： W_k 与 W_Q 。所得乘积在所有连通边上归一化，为了**削弱这些点积阻碍梯度下降**的趋势，将其梯度标为 $1/\sqrt{d}$



• 图转换网络Graph Transformer编码

- 在整个图转换网络中，为了增强多头注意力层。构建了L个块网络。每个块都应用了以下的变换：

$$\bar{v}_i = \text{LayerNorm}(v_i' + \text{LayerNorm}(v_i))$$

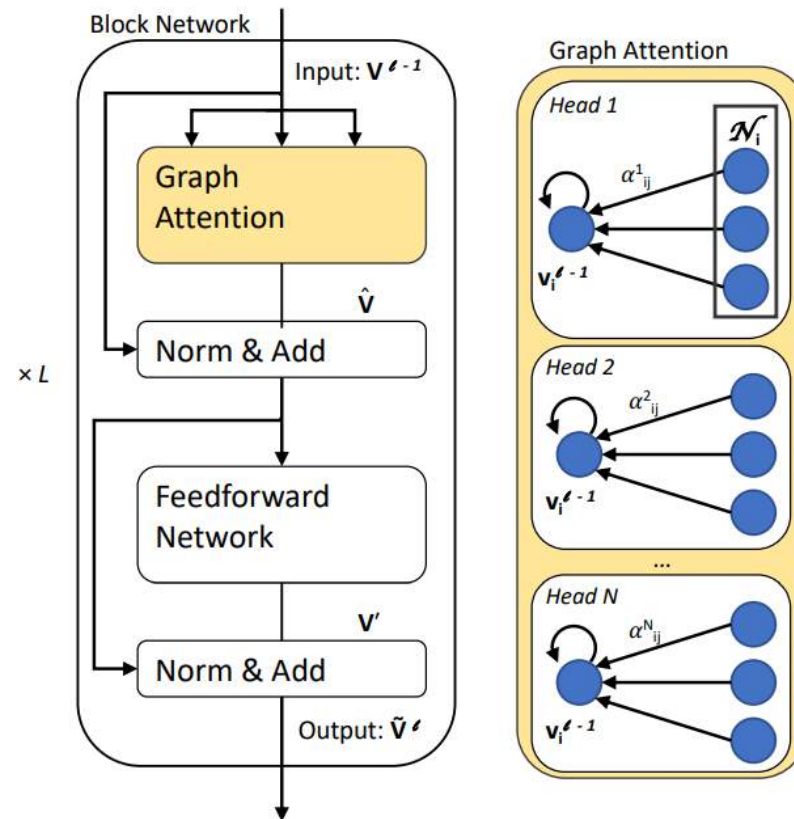
$$v_i' = \text{FFN}(\text{LayerNorm}(v_i))$$

- 第 $l-1$ 层的输出作为第 l 层的输入：

$$V_i^l = \hat{V}_i^{l-1}$$

- 最终得到顶点编码

$$V_L = [V_i^L]$$



- 标题编码器

- 双向RNN编码作为标题词嵌入

$$T = \text{BiRNN}(x'_1 \dots x'_m)$$

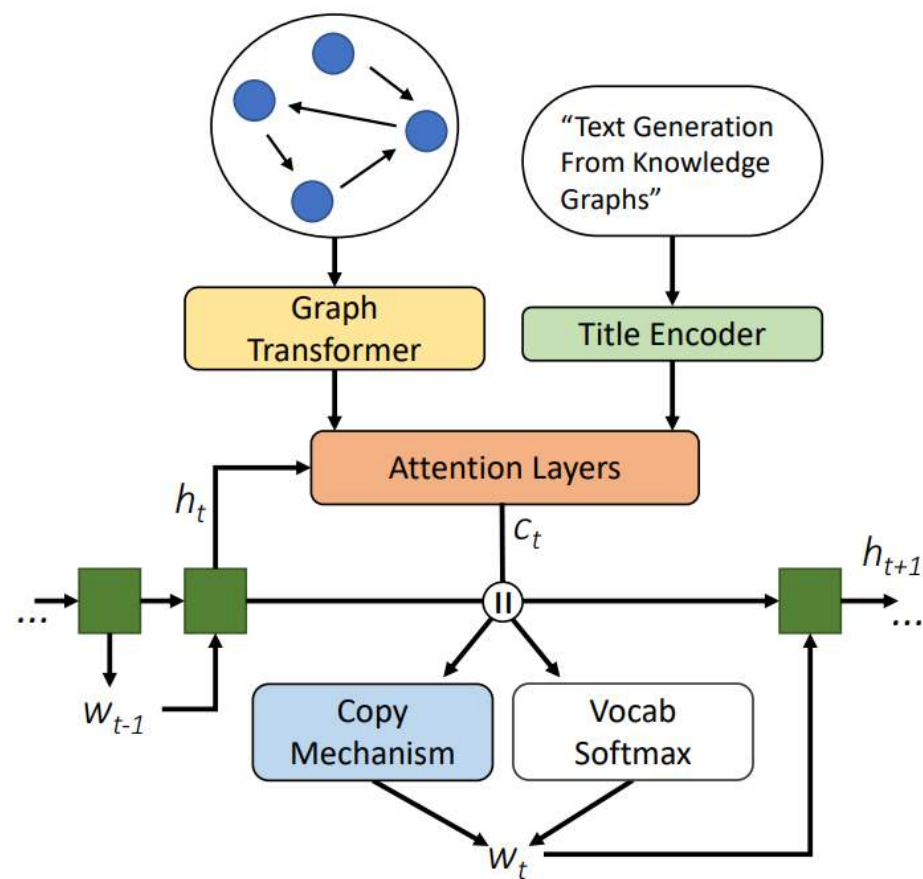
- 输出解码器

- 基于注意力机制的解码器进行解码得到上下文向量

$$c_g = h_t + \prod_{n=1}^N \sum_{j \in V} a_j^n W_G^n v_j^L$$

- 利用 h_t 和 c_t 决定了每个时刻的输出是来自词表还是从图或标题中的实体中复制

$$p * a^{copy} + (1 - p) * a^{vocab}$$



- 实验数据

- AGENDA (the **A**bstract **GE**Neration **DA**taset)数据集——科技论文领域

- 语料库摘自12届人工智能会议论文集
 - 包含顶级人工智能会议的4万篇论文

- 实验方法

- 人工评价

- Best-Worst Scaling的方法

- 自动评价

- BLEU值——Bilingual Evaluation Understudy（双语评估替换）
 - METEOR值——词序纳入评估范畴，设立基于词序变化的罚分机制

- 人工评价- Best-Worst Scaling方法

- 在Rewriter模型、GraphWriter模型与人类写作产生的候选结果中，由专家投票选出Best 和 Worst
- 在Rewriter和GraphWriter两个模型之间，根据grammar, fluency, coherence, informativeness 4个指标，选出胜负

- 实验结果

	Best	Worst
Rewriter (No knowledge)	12%	64%
GraphWriter (Knowledge)	24%	36%
Human Authored	64%	0%

	Win	Lose	Tie
Structure	63%	17%	20%
Informativeness	43%	23%	33%
Grammar	63%	23%	13%
Overall	63%	17%	20%

- 自动评价
 - 计算模型生成结果的指标
- 对比方法
 - 未引入全局节点的模型： GAT (NIPS2018)
 - 未采用图结构表示知识的模型： EntityWriter (ACL2019)
 - 仅采用标题的模型： ReWriter (ACL2018)

- 评价指标

- BLUE值
- METEOR值

	BLEU	METEOR
GraphWriter	14.3 $\pm$ 1.01	18.8 $\pm$ 0.28
GAT	12.2 $\pm$ 0.44	17.2 $\pm$ 0.63
EntityWriter	10.38	16.53
Rewriter	1.05	8.38

- 结论

- 优点:

- 用基于Transformer架构的图编码器对知识图谱进行特征编码，从而实现对文章全局特征的提取，提高了生成文本摘要的质量

- 缺点:

- 由模型生成的所有句子中有约18%的重复。
 - 同时约有40%的实体未出现在生成的文本中。
 - 句式相对匮乏、单一

in this paper , we present a novel approach for PRIMAL-DUAL ALGORITHM based on a CONVEX FORMULATION of the MINIMAL TWO-DIMENSIONAL SURFACE and a GLOBALLY OPTIMAL SOLUTION . the PRIMAL-DUAL ALGORITHM is based on the GLOBALLY OPTIMAL SOLUTION and a CONVEX FORMULATION for the CONVEX MINIMIZATION PROBLEM . the PRIMAL-DUAL ALGORITHM is evaluated on the CORRESPONDENCE PROBLEM

基于图结构处理的文本生成



算法原理

T	基于图到序列模型的中文文章连贯评论生成
I	新闻文档
P	1.主题交互图建模 2.顶点编码与图编码对主题交互图进行处理 3.融合attention机制的解码器进行输出
O	文章评论

P	传统的Seq-Seq模型不适用于从较长的新闻文章中生成短评论
C	评论生成场景（短文本）
D	如何围绕文章主题进行图建模与处理
L	2020 ACL

• 算法背景

- 聚焦评论（短文本）生成的场景
- 实际问题：
 - 经典seq-to-seq模型难以处理长新闻
 - 新闻标题夸张且信息量少
 - 评论的宽角度和多样化

Title

这部影片被称为“十年来最搞笑漫威电影”，你看了吗？

Have you seen the movie intitled as “the most hilarious [Marvel movie](#)”?

Content

点击“IPTV4K超高清”订阅，精彩内容等你共享
《[复仇者联盟3：无限战争](#)》中的巅峰一役，将战火燃遍了整个宇宙...作为接档《[复联3](#)》的漫威电影，《[蚁人2](#)》的故事爆笑中带着温情，无疑成为了现阶段抚平漫威粉心中伤痛的一味良药...看过《[复联3](#)》的漫威粉们，心中都有同一个疑问：在几乎整个[复仇者联盟](#)都参与到无限战争的关键时刻，[蚁人](#)究竟去哪儿了？ ...

Click on the “IPTV4K ultra HD” to subscribe, fantastic contents are waiting for you to share. The battle in “[Avengers: Infinity War](#)” has spread the flames of war throughout the universe ... As the continuation [Marvel movie](#) to “[Avengers 3](#)”, the hilarious and warm “[Ant-Man and the Wasp](#)” is no doubt a good dose to heal the fans of [Marvel](#) at the time. ... Fans of the [Marvel](#) who have watched “[Avengers 3](#)” all have a doubt about where [Ant-Man](#) is when all other [Avengers](#) have been involved in the infinity war.

Comment

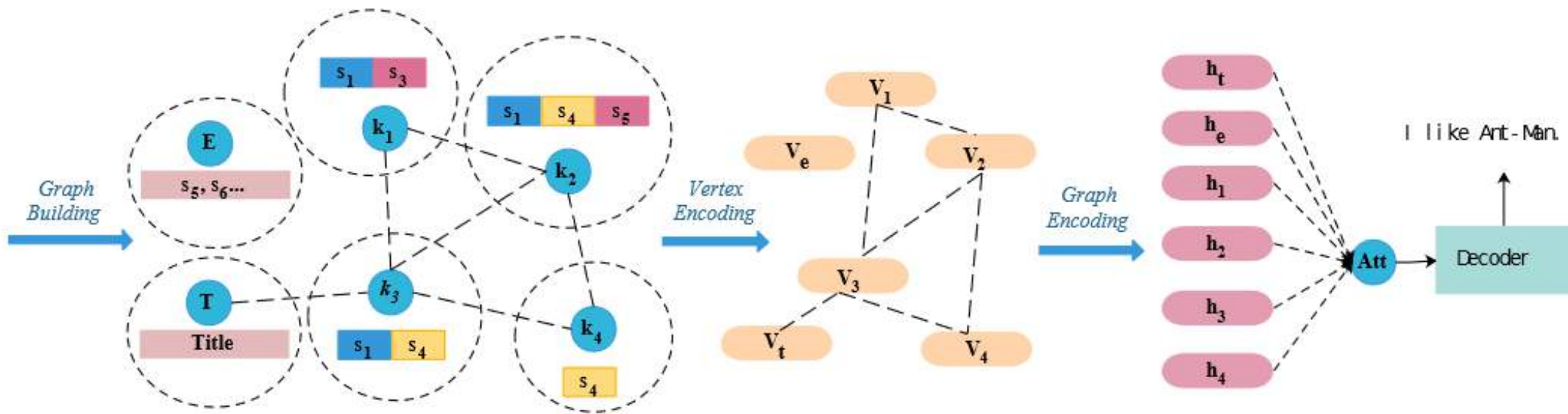
只有我觉得那个头盔像[蚁人](#)的头盔吗？

Am I the only one that thinks the helmet similar to the helmet of [Ant-Man](#)?

- 算法流程

- 图的构建→顶点编码→图编码→解码器→文本输出

Title: Have you seen the movie intituled as "the most hilarious Marvel movie?"
S1: Click on the "IPTV4K ultra HD" to subscribe.
S2: fantastic contents are waiting for you to share .
S3
S4
S5
S6
...



• 图的构建

- 使用了**CoreNLP**进行分词和NER构建关键词集合作为顶点
- 通过**TextRank**扩充关键词集合
- 边的构建
 - 只要**关键词 κ** 出现在**句子 S_i** 中
 - 就把 S_i 作为**关键词 κ** 作为属性
 - 显然一个句子可以出现多个 κ
 - 满足这个条件中的**不同的 κ** 就连接一边
 - **边的权重**被定义为共现句子的数量

Algorithm 1 Graph Construction

Require: The title $title$ and article text D , weight calculation function λ

```
1: Segment  $title$  and  $D$  into words
2: Do named entity recognition and keyword detection and get the keywords  $\kappa$ 
3: for sentence  $s$  do
4:   if  $s$  contains  $k \in \kappa$  then
5:     Assign  $s$  to vertex  $v_k$ 
6:   else
7:     Assign  $s$  to vertex  $v_{empty}$ 
8:   end if
9: end for
10: for vertex  $v_i$  and  $v_j$  do
11:   Calculate edge weight:  $w_{ij} = \lambda(v_i, v_j)$ 
12: end for
```

- 图的构建

Title: Have you seen the movie intitled as ``the most hilarious Marvel movie?

S1: Click on the ``IPTV4K ultra HD" to subscribe.

S2: fantastic contents are waiting for you to share .

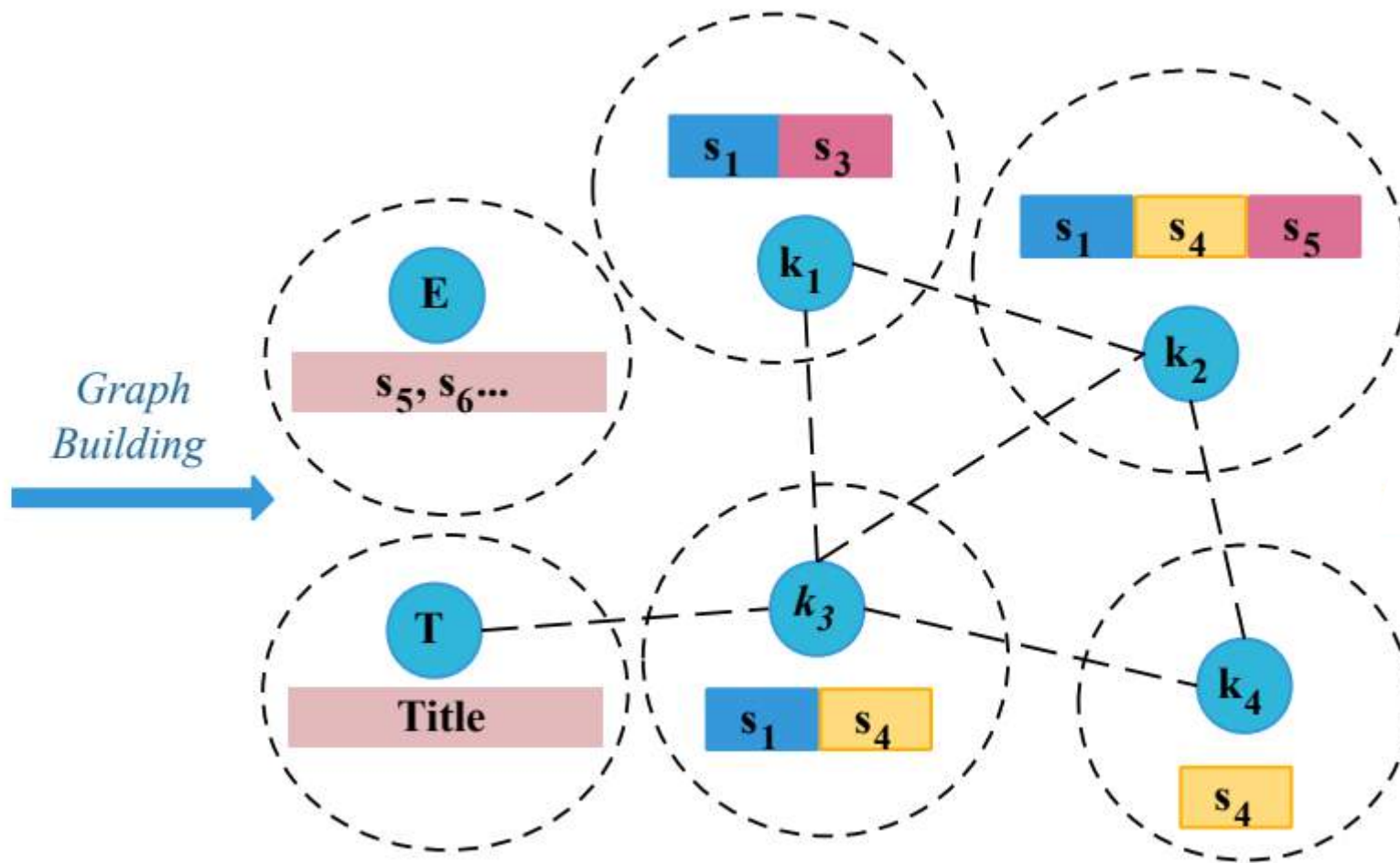
S3

S4

S5

S6

...



- 顶点编码 Vertex Encoding

- 每个顶点下都有一段相对应的文本，先用词嵌入提取每个文本的每个词向量的特征同时再加上位置信息嵌入表示，节点所代表的keyword放在第一位
- 现在得到了顶点的词嵌入初始表示后，作者再对每个点的文本序列中每个词做多头自注意力机制

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{softmax}(QK^T)V \\ \text{MultiHead}(Q, K, V) &= [\text{head}_1; \dots; \text{head}_h]W^O \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned}$$

- Q,K,V都是一样的，最后再提取keyword的隐藏层状态作为该点的特征向量

- 图编码Graph Encoding

- 使用标准的GCN模型。这种方法的优点在于既可以**对顶点的内容进行建模**，也可以利用**图的结构信息**。由于新闻标题是重要信息，使用图编码器的标题顶点的隐藏输出作为解码器的初始状态。表示如下：

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^l W^l)$$
$$\tilde{A} = A + I_N$$

- 为了过渡平滑问题，在层与层之间加上了**残差连接**。通过tanh得到输出

$$g^{l+1} = H^{l+1} + H^l$$
$$g^{out} = \tanh(W_o g^K)$$

- 解码输出 Decoder

- 从图编码器中，得到了GCN的编码结果：

$$\langle g_0, g_1, \dots, g_n \rangle$$

- 在每一个时间步上， c_i 由GCN的输出做Attention得到

$$\alpha_j = \frac{\exp(\delta(t_i, g_j))}{\sum \exp(\delta(t_i, g_k))} \quad c_i = \sum \alpha_j \times g_j$$

- 考虑到抽取出的关键词本身的重要性，应用copy机制，来选择从实体复制还是词表生成

$$y_i = \text{softmax}(W_o (\tanh(W ([t_i; c_i]) + b)))$$

$$p_{\text{copy}} = \sigma(W_{\text{copy}} (t_i; c_i))$$

$$p = (1 - p_{\text{copy}}) \times y + p_{\text{copy}} \times \alpha$$

- 实验数据

- 语料库收集自在线新闻平台腾讯快报
- 主题选取：娱乐模块和体育模块

	average word #		average character #	
	Ent	Sport	Ent	Sport
content	456.1	506.6	754.0	858.7
title	16.4	15.7	28.1	27.4
comment	16.3	19.4	26.2	31.2
keyword	8.4	9.0	-	-

- 评价方法：由三位人工专家打分，分别对三个指标进行评分（0-10）
 - Coherence 相关度
 - Informativeness 信息丰富度
 - Fluency 流畅程度

• 对比方法

- seq2seq模型：分别以标题（T）、内容（C）、标题+内容（TC）作为输入，最大长度设为100.
- self-attention编码模型
- hierarchical attention HA编码模型

Title	被王丽坤美到了，《上新了·故宫》里穿古装温婉又娴静，气质惊艳 <i>In "updates of the Palace Museum" Likun Wang appears so gentle, refined and astonishingly elegant wearing ancient costume that audiences are touched by her beauty.</i>
S2S-T	我觉得还是喜欢看的古装，古装扮相，古装扮相很好看 <i>I still think I like ancient costume, appearance in ancient costume, appearance in ancient costume is pretty.</i>
S2S-C	我觉得还是喜欢看的 <i>I still think I like to watch</i>
S2S-TC	我觉得还是喜欢看的 <i>I still think I like to watch</i>
SA-B	我觉得赵丽颖的演技真的很好 <i>I think the acting skill of Liying Zhao is very good</i>
SA-K	我觉得还是喜欢李沁 <i>I still think I like Qin Li</i>
HA	我觉得还是喜欢看她的剧 <i>I still think I like her plays</i>
graph2seq	王丽坤的演技真的好 <i>The acting skill of Likun Wang is really good</i>

• 实验结果

娱乐模块

Models	Coherence	Informativeness	Fluency	Total
seq2seq-T (Qin et al., 2018)	5.38	3.70	8.22	5.77
seq2seq-C (Qin et al., 2018)	4.87	3.72	8.53	5.71
seq2seq-TC (Qin et al., 2018)	3.28	4.02	8.68	5.33
self-attention-B (Chen et al., 2018)	6.72	5.05	8.27	6.68
self-attention-K (Chen et al., 2018)	6.62	4.73	8.28	6.54
hierarchical-attention (Yang et al., 2016)	1.38	2.97	8.65	4.33
graph2seq (proposed)	8.23	5.27	8.08	7.19

体育模块

Models	Coherence	Informativeness	Fluency	Total
seq2seq-T (Qin et al., 2018)	4.30	4.38	6.27	4.98
seq2seq-C (Qin et al., 2018)	3.88	3.85	6.02	4.58
seq2seq-TC (Qin et al., 2018)	4.70	5.08	6.37	5.38
self-attention-B (Chen et al., 2018)	5.15	5.62	6.28	5.68
self-attention-K (Chen et al., 2018)	6.68	5.83	7.00	6.50
hierarchical-attention (Yang et al., 2016)	4.43	5.05	6.02	5.17
graph2seq (proposed)	7.97	6.18	6.37	6.84

- 结论

- 优点:

- Coherence 相关度

- 在一致性方面获得了更高的分数
 - 引入图结构使模型可以**更好地获得文章的要点**，而不是参考那些与文章仅略有关联甚至无关的高频术语

- Informativeness 信息丰富度

- 信息丰富度与相关度存在联系。相关度得分较高的模型往往信息量更大。引入的图结构使得文本生成的**语义空间**得到了扩充，生成了**信息丰富度更高**的文本

- 缺点

- Fluency 流畅程度

- 流畅程度大体与其他模型相似，原因在于**解码器**的结构基本相同。存在部分重复和知识矛盾的问题，有待进一步解决

» “**鹿晗是谁？我只认识鹿晗**” “**大餐是个好演员**”

- 文本摘要

- 能够从原始信息中提炼出简短的摘要，可以有效降低用户信息负担、提高用户信息获取速度
- 在信息检索、舆情分析、内容审查等领域具有重要作用

- 评论生成

- 自动评论有助于鼓励在线平台上的用户参与和互动
- 对智能对话、创意写作有着启发性的借鉴意义

- 未来展望

- 引入修辞句法学习，丰富句式结构
- 融合更多外部知识，优化生成句子的逻辑
- 图结构处理思维的扩展：漏洞发现，自动编程等

- [1] Koncel-Kedziorski R , Bekal D , Luan Y , et al. Text Generation from Knowledge Graphs with Graph Transformers[J]. 2019.**
- [2] Li W , J Xu, Y He, et al. Coherent Comments Generation for Chinese Articles with a Graph-to-Sequence Model[J]. 2019.**
- [3] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.**

道可道，非常道。名可名，非常名。无名天地之始。有名万物之母。故常无欲以观其妙。常有欲以观其徼。此两者同出而异名，同谓之玄。玄之又玄，众妙之门。

谢谢！

