基于实体信息增强及多粒度融合的多文档摘要

唐嘉蕊[1]，刘美玲[1,*]，赵铁军[2]，周继云[3]
1.东北林业大学信息与计算机工程学院，哈尔滨，150006
2.哈尔滨工业大学计算机科学系，哈尔滨，150001
3.约翰斯·霍普金斯大学利伯研究所，巴尔的摩，MD 21218，USA
{tjr,mlliu}@nefu.edu.cn,tjzhao@hit.edu.cn,zhoujiyun2010@gmail.com

摘要

神经网络模型的快速发展使得多文档摘要可以获得人类可读的流畅的摘要，对大规模的数据进行预训练可以更好的从自然语言文本中捕捉更丰富的语义信息，并更好的作用于下游任务。目前很多的多文档摘要的工作也应用了预训练模型(如BERT)并取得了一定的效果，但是这些预训练模型不能更好的从文本中捕获事实性知识，没有考虑到多文档文本的结构化的实体-关系信息，本文提出了基于实体信息增强和多粒度融合的多文档摘要模型MGNIE，将实体关系信息融入预训练模型ERNIE中，增强知识事实以获得多层语义信息，解决摘要生成的事实一致性问题。进而从多种粒度进行多文档层次结构的融合建模，以词信息、实体信息以及句子信息捕捉长文本信息摘要生成所需的关键信息点。本文设计的模型，在国际标准评测数据集MultiNews上对比强基线模型效果和竞争力获得较大提升。

关键词： 实体信息增强；预训练语言模型；多粒度融合；多文档摘要

Multi-Document Summarization Based on Entity Information Enhancement and Multi-Granularity Fusion

Jiarui Tang[1],Meiling Liu[1,*],Tiejun Zhao[2],and Jiyun Zhou[3]
1.School of Information and Computer Engineering, Northeast Forestry University, Harbin 150006, China
2.Department of Computer Science, Harbin Institute of Technology, Harbin 150001,China
3.Lieber Institute, Johns Hopkins University, Baltimore, MD 21218, USA
{tjr,mlliu}@nefu.edu.cn,tjzhao@hit.edu.cn,zhoujiyun2010@gmail.com

Abstract

The rapid development of neural network models enables multi-document summarization to obtain human-readable and fluent summaries, and pre-trained on large-scale data can better capture richer semantic information from natural language texts and better serve downstream tasks.Many current works on multi-document summarization also apply pre-trained models (such as BERT) with certain results, but these pre-trained models cannot better capture factual knowledge from texts and do not consider the structure of multi-document texts.This paper proposes a multi-document summarisation model MGNIE based on entity information enhancement and multi-granularity fusion, incorporating entity relationship information into the pre-trained model ERNIE, enhancing knowledge facts to obtain multi-layer semantic information and solving the factual consistency problem of summary generation. In turn, the multi-document hierarchy is fused and modelled at multiple granularities to capture the key information points required for summary generation of long text information in terms of word information, entity information and sentence information. The model designed in

第二十一届中国计算语言学大会论文集，第155页-第165页，南昌，中国，2022年10月14日至16日。

this paper achieves a significant improvement in effectiveness and competitiveness over the strong baseline model on the international standard evaluation dataset MultiNews.

Keywords: Entity Information Augmentation , Pre-trained Language Models , Multi-Granularity Fusion , Multi-document Summarization

1 引言

多文档摘要指的是在保留关键信息的情况下从同一主题相关的多个文档集合中生成简洁的摘要，其各个文档包含的信息虽属于同一个主题却并不相同。近年来，互联网科技迅速发展，使得我们在各种社交媒体上获得大量的数据信息，而随着新闻的快速传播，从同一主题的新闻中获取关键信息显得至关重要。随着深度学习技术在多文档摘要方面的广泛应用以及大规模数据集的发布，如WikiSum(Liu et al.,2018) ,MultiNews(Fabbri et al.,2019),生成式的多文档摘要取得了突破性进展。

最近，如BERT(Lee et al.,2018)等预训练语言模型的提出，将大规模语料库的训练好的语言模型应用于下游nlp任务，对BERT模型进行微调，使其能够更好的编码文本的上下文信息，捕捉到更深层的语义信息。最近在文本摘要方面，很多工作加入了预训练语言模型，(Liu et al.,2019)首先提出将BERT模型作为预训练模型应用于文本摘要任务，作者通过对BERT模型进行微调，通过将文档中的句子用[CLS]符号分割来学习句子表征，并且更改了区间分割嵌入来区分不同句子，作者还提出通过对编码器和解码器选取不同的优化器来解决预训练模型编码器和解码器不匹配的问题。目前在多文档摘要方面，虽然有加入预训练模型来提高模型性能的工作，但是并没有考虑带有事实信息的预训练模型来提升模型生成的事实一致性的工作。

对于的生成式多文档摘要，获取文本中丰富的语义信息对于生成连贯的摘要是非常重要的，以往的工作中，大部分生成式模型采用单词级语言生成，也有采用词级与句子级进行信息融合的摘要模型，以及应用段落级和篇章级的生成模型，能够充分获得丰富的文本信息。但在目前的工作中缺乏实体级的语义信息与其他语义单元的信息融合的生成式模型，从而丰富层次化的具有结构信息的自然语言文本。

在本文中，我们针对具有结构化的实体信息可以增强生成式摘要的事实一致性，并且融合了实体-关系信息的预训练语言模型能够使文本获得更高层级的语义表征。本文提出了基于实体信息增强以及多粒度融合的多文档摘要模型，它采用了融合了实体-关系结构化信息的ERNIE预训练模型(Zhang et al.,2019)来训练文本，来实现信息增强，将结构化的带有实体-关系的知识图通过tranE算法(Bordes et al.,2013)嵌入到预训练模型中，并实现了实体对齐，获得摘要所需的实体信息。同时我们还采用了多粒度信息融合，将词信息、实体信息和句子信息进行交互融合，从而获得多文档中更具层次化的文本语义信息。针对上文中提出的现有研究的问题，本文的贡献如下:

1）本文提出了一个基于实体信息增强的多文档摘要模型，通过采用具有结构化的实体-关系的知识图通过transE算法将结构化的图信息嵌入到ERNIE预训练模型中，使用实体链接工具TAGME来对文本中提及的实体进行提取，并进行训练从而在丰富上下文信息的基础上进一步加入实体信息实现信息增强。

2）文本提出多个粒度的信息来对原文本进行丰富的语义信息提取，我们将实体信息与词信息进行实体对齐，并通过句子信息和实体信息的融合对词token信息进行更新从而指导解码生成。

3）本文提出的模型在大规模数据集MultiNews上进行实验并取得了先进性结果表明了模型的有效性和可行性，并进行了对多粒度信息，以及是否加入融合实体信息的预训练模型进行了消融实验对比，来说明实体信息增强的有效性。

*通讯作者
基金项目：国家自然科学基金(61702091)

2 相关工作

2.1 基于信息增强的多文档摘要

以往的多文档摘要是基于特征工程和主题模型的(Erkan et al., 2004;Christensen et al., 2013; Yasunaga et al., 2017)，通过特征增强和语义增强来提升模型性能。(Zheng et al.,2019)提出从文档视图和子主题视图中共同生成基础主题表示，通过考虑上下文信息，作者考虑了上下文信息、子主题显著性和相对句子显著性，并且以分级的方式来估计句子的显著性，从而抽取排名最高的句子作为摘要。(Alambo et al.,2020)提出基于中心性聚类的方法，使用相关参考分解从原文档中提取句子集合并且保持他们相互依赖，并且采用增强的多句压缩算法生成主题信息和摘要。其中还有依据数据增强的多文档摘要，(Pasunuru et al.,2021)提出了构建两个新的针对以查询为中心的多文档摘要数据集来实现数据增强，这两个数据集是互补的，并提出采用分层编码的方式来进行编码，同时对局部信息以及全局信息进行了编码，还加入了排序组件和查询组件。

BERT等预训练语言模型的提出，促进了多文档摘要任务的发展，(Li et al.,2020)提出利用图对文档进行编码，能更好的捕捉跨文档的关系，基于图来指导摘要生成，还提出了将BERT模型与作者提出的基于图指导的摘要模型结合起来，以更有效的处理长输入文本。针对事实一致性问题，提出的大多数方法是在评估指标方面对生成摘要的事实一致性进行评估，(Zhang et al.,2019)采用了一种弱监督的方法构造训练集，通过构造的句子文档对来判断是否具有事实一致性。近年来提出了通过外部知识库来生成文本的忠实性，(Dong et al.,2022)把不在原文本中但在与原文本链接的外部知识库中的实体视为对世界知识的忠实，原文本具有提取性，世界知识具有生成性，与之前通过过滤训练实例的只包含提取性的实体来提高事实一致性的工作相反，作者通过提供与来源相关的额外事实，以生成式的角度来提高生成实体的忠实性。

2.2 多粒度信息融合

对于的生成式多文档摘要，获取文本中丰富的语义信息对于生成连贯的摘要是非常重要的，以往的工作中，有采用单词级语言生成的，采用词级与句子级进行信息融合的摘要模型，以及应用段落级和篇章级的生成模型，从而获得丰富的文本信息。而在目前的工作中缺乏实体级的语义信息与其他语义单元的信息融合从而丰富层次化的具有结构信息的自然语言文本。

在近些年的工作中，随着深度学习的快速发展，对于多文档摘要的研究从多个粒度方面进行，大多数工作是采用单词级的文本嵌入表征来获得上下文信息，也将其他粒度的信息如段落、文档进行融合来输入表征。(Li et al.,2020) 提出了一种神经生成式多文档摘要(MDS)模型，该模型利用段落级和词级的图表示结构，如相似图和篇章图，来有效地处理多个输入文档并产生生成式摘要。transformer(Vaswani et al.,2017)的提出使得生成式文本摘要取得了突破性进展。(Zhao et al.,2020)提出SummPip模型是第一种结合语义知识和深度神经表示构造句子图的无监督摘要方法，(Jin et al.,2020)提出采用文档、句子、词多粒度信息交互网络，在不同语义粒度信息表征进行交互。(Yasunaga et al.,2017)提出在关系图上使用图卷积网络(GCN)，并将从递归神经网络获得的句子嵌入作为输入节点特征。通过多层分层传播，GCN生成高级隐藏句特征以进行显着性估计。

以上这些已提出的方法虽然在一定程度上解决了多文档摘要生成的事实一致性以及丰富的文本语义信息特征提取问题，但是针对通过预训练模型嵌入结构化的实体信息来进行信息增强的多文档摘要模型还很少，通过实验我们发现将实体信息融入文本单元中进行特征融合对于生成式文本摘要性能的提升具有有效性。

3 基于实体信息增强及多粒度融合的多文档摘要模型MGNIE

在这一节中，我们详细描述了我们提出的模型。模型的结构如图1所示。在本文中，首先使用transE算法将结构化的实体信息嵌入ERNIE预训练模型中，我们使用TAGME实体链接工具来提取文本中提及的实体，来对原文本进行实体信息融合的预训练，从而得到预训练后的词嵌入信息和实体嵌入信息，同时通过对句子进行编码获得句子嵌入信息，输入到Transformer编码层进行融合，最后通过句子信息和实体信息的融合对词token信息进行更新从而指导解码生成。

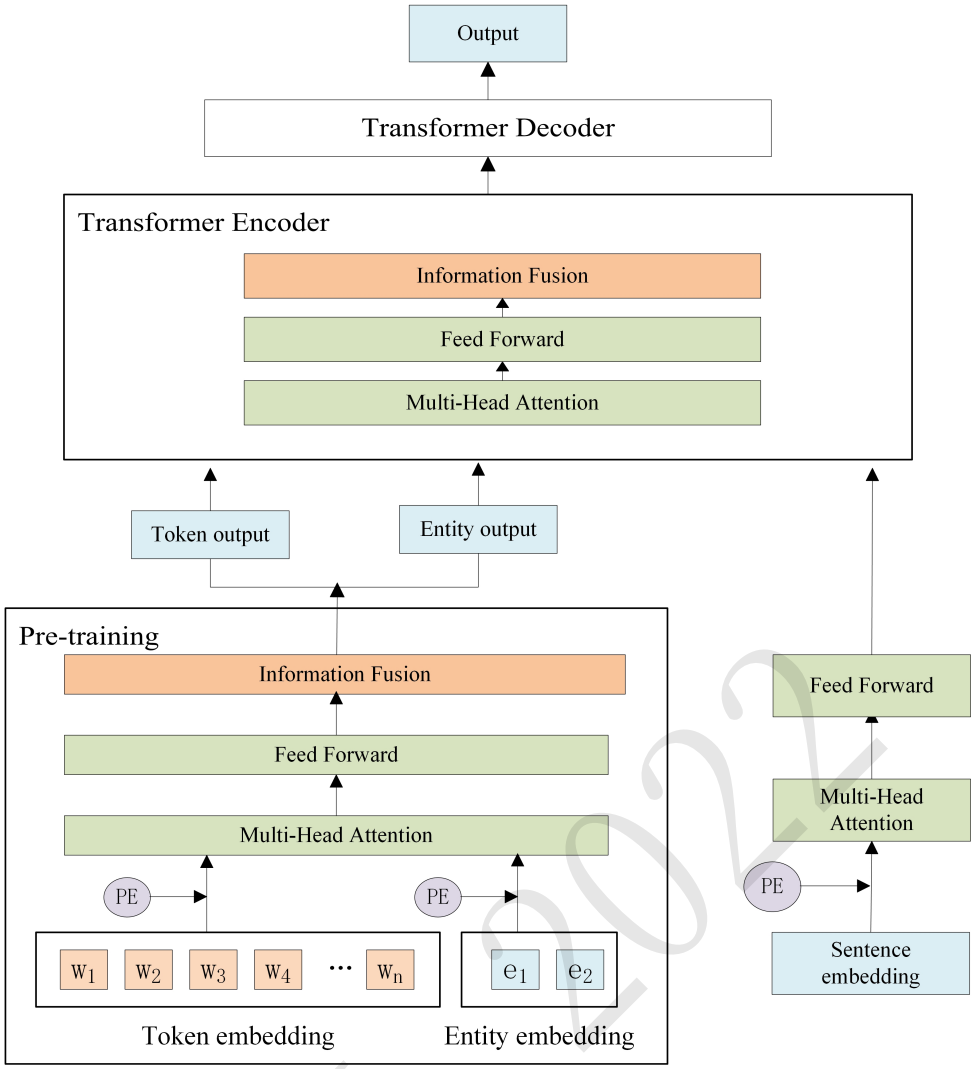

Figure 1: 基于实体信息增强及多粒度融合的多文档摘要模型MGNIE

3.1 嵌入知识图的预训练模型ERNIE

本文通过加入知识图结构来增强预训练模型的事实一致性，采用transE算法将Wikidata知识图的实体-关系信息输入到ERNIE模型进行训练，然后将带有实体信息的原文本输入到预训练模型中进行预训练从而得到词嵌入和实体嵌入。

我们将词序列集合定义为$W = \{w_1, w_2, \ldots, w_n\}$，其中n表示词序列的长度，将实体序列定义为$E = \{e_1, e_2, \ldots, e_m\}$,其中m为实体序列的长度，将句子序列表示为$S = \{s_1, s_2, \ldots, s_o\}$,其中o为句子序列的长度。知识图$KG_s$中的所有实体表示为E，我们将源文本中的实体与$KG_s$中的实体对齐。

在预训练模型中，我们将词信息与KG_s中的实体信息分别进行编码，然后输入到前馈神经网络层，进行异质信息融合。首先对词序列进行编码，将词嵌入e_w和段嵌入s_w以及位置嵌入p_w相加获得最终词嵌入:

$$h_w^0 = e_w + s_w + p_w \tag{1}$$

同样的，实体嵌入可计算为:

$$h_e^0 = e_e + s_e + p_e \tag{2}$$

则获得最终的词嵌入$h_w^{l-1} = \{h_{w_1}^{l-1}, h_{w_2}^{l-1}, \ldots, h_{w_n}^{l-1}\}$和实体嵌入$h_e^{l-1} = \{h_{e_1}^{l-1}, h_{e_2}^{l-1}, \ldots, h_{e_m}^{l-1}\}$,将它们作为输入，送入多头注意力中，

第二十一届中国计算语言学大会论文集，第155页-第165页，南昌，中国，2022年10月14日至16日。

$$h_w^l = MHAtt(h_{w_i}^{l-1}, h_{w_n}^{l-1}) \tag{3}$$

$$h_e^l = MHAtt(h_{e_j}^{l-1}, h_{e_m}^{l-1}) \tag{4}$$

原文本中的词token嵌入包含实体信息，与KG_s中的实体对齐，并将实体信息融入到词序列中进行异质信息融合，实现了外部实体嵌入的信息增强。源文本词序列token包含实体的表示为h_w，KG_s中的实体在源文本序列中有对应的表示为h_e，则可以表示为$h_e = f(h_w)$，融合后的表征为:

$$h_1 = \sigma(h_w^l, h_e^l) = (W_t w_i + W_e e_j + b) \tag{5}$$

$$h_w = \sigma(h_1 W_t + b_t) \tag{6}$$

$$h_e = \sigma(h_1 W_e + b_e) \tag{7}$$

其中W_t、W_e、b表示可训练的权重参数，h_1表示整合了token和实体信息的内部隐藏状态，本文使用非线性激活函数$GELU$对实体嵌入和词嵌入进行融合。对于没有相应实体的词嵌入，不进行信息融合而直接输出。为了简化，我们将经过预训练模型的词嵌入向量h_w仍表示为$h_w^l = \{h_{w_1} l, h_{w_2}^l, \ldots, h_{w_n}^l\}$，实体嵌入向量$h_e$表示为$h_e^l = \{h_{e_1}^l, h_{e_2}^l, \ldots, h_{e_m}^l\}$

本文将外部KG_s中的实体信息融入到源文本中，通过mask表示实体的词token，通过上下文对进行实体预测来预训练模型，使模型获得更丰富以及更高语义的信息，从而能够生成更好的表征。预训练模型mask实体自动编码过程的损失函数可以用下述公式来计算:

$$p(e_x|w_i) = \frac{exp(linear(w_i^0)e_x)}{\sum_{j=1}^{m} exp(linear(w_i^0)e_j)} \tag{8}$$

其中$linear()$表示一个线性层。

3.2 多粒度信息融合

本文采用多粒度信息，包括词嵌入、实体嵌入和句子嵌入，在预训练模型阶段，我们将外部的实体信息与原文本包含的实体信息进行融合从而获得实体级的信息增强，在摘要模型输入阶段，我们分别将源文本划分为词序列、实体序列和句子序列，并对不同粒度的信息进行信息融合，从而获得包含更加丰富语义的语言模型表征。

我们将经过预训练模型的源文本中的词向量嵌入表示为$h_w^{l-1} = \{h_{w_1}^{l-1}, h_{w_2}^{l-1}, \ldots, h_{w_n}^{l-1}\}$，将源文本中提取的实体向量嵌入表示为$h_e^{l-1} = \{h_{e_1}^{l-1}, h_{e_2}^{l-1}, \ldots, h_{e_m}^{l-1}\}$，采用同样的方法，我们可以获得句子序列表示$S = \{s_1, s_2, \ldots, s_o\}$进行编码获得句子嵌入:

$$h_s^0 = e_s + s_s + p_s \tag{9}$$

则句子嵌入表示为$h_s^{l-1} = \{h_{s_1}^{l-1}, h_{s_2}^{l-1}, \ldots, h_{s_o}^{l-1}\}$，送入多头注意力，可以得到句子的上下文信息:

$$h_s^l = MHAtt(h_{s_z}^{l-1}, h_{s_o}^{l-1}) \tag{10}$$

对多粒度信息进行融合以获得对源文本更加丰富的特征，实体信息的融合体现了生成摘要过程中对事实的准确性。我们加入的实体特征是在源文本中出现的，且能够链接到外部知识KG_S的结构化的实体关系。我们使用融合函数进行融合，首先获得词$token$与实体的融合信息h_w。

融合后的实体信息与词信息表示为部分词$token$融入了实体的嵌入信息，再将融合后的词序列信息与句子序列信息进行融合，得到融合后的词向量h_w':

$$h_2 = \sigma(h_1, h_s^l) = \sigma(\sigma(h_w^l, h_e^l), h_s^l) \tag{11}$$

$$h_w^{'} = \sigma(h_2 W_t + b_t) \tag{12}$$

将获得的融合的词表征进一步送入前馈神经网络用来进一步的转化丰富的语义信息：

$$h = LayerNorm(h^{l-1} + h_w^{'}) \tag{13}$$

$$h^l = LayerNorm(h + FFN(h)) \tag{14}$$

FFN是两层前馈网络，采用ReLU隐藏激活函数，其中layerNorm是层规范化。h^l表示编码器的输出。将经过编码器融合的输入向量h^l以及隐藏状态输入到transformer解码器中进行逐词解码，编码器输出作为key和value，将输入嵌入和词位置编码输入到解码器中经过多头注意力机制以及前馈神经网络层，得到上下文表征作为query，输入多头注意力机制中，得到输出g^l，最后送入softmax，来计算目标词汇生成分布：

$$P_t = softmax(g^l W_g + b_g) \tag{15}$$

其中W_g、b_g为可训练的参数。本文使用的交叉熵损失函数为：

$$L = -\frac{1}{N}\sum_{n=1}^{N} logP(y_w^{(n)}) \tag{16}$$

其中$y_w^{(n)}$表示生成的真实摘要，N表示语料库的样本数。

3.3 Wikidata知识图实体嵌入

我们采用外部知识图来对文本的实体信息进行增强，采用transE算法将Wikidata知识图的实体-关系信息输入到ERNIE模型进行训练。Wikidata知识图是一个开放的多关系知识图谱，它包含了包括维基百科的结构化的数据，我们从Wikidata知识图抽取实体-关系三元组，并且通过transE算法学习实体嵌入。该算法将关系数据中的实体和关系嵌入低维向量空间。给定一个实体-关系三元组(h,l,t)，他们由h头实体，t尾实体，l关系组成，通过模型学习实体和关系的嵌入向量，算法的原理就是通过边所对应的关系对应于嵌入的转换，当(h,l,t)成立时，使得头实体向量和关系向量尽可能的靠近尾实体向量，并计算(h,l)和t之间的距离。

transE训练模型原理是从实体矩阵和关系矩阵中各自抽取一个向量，进行运算得到的结果近似等于实体矩阵中另一个实体的向量，从而达到通过词向量表示知识图中已存在的三元组。transE的损失函数为：

$$\mathcal{L} = \sum_{(h,l,t)\in S} \sum_{(h',l,t')\in S'_{(h,l,t)}} [\gamma + d(h+l,t) - d(h^{'}+l,t^{'})]_+ \tag{17}$$

$$S'_{(h,l,t)} = \{(h^{'},l,t) \in E\} \cup \{(h,l,t^{'}) | t^{'} \in E\} \tag{18}$$

其中公式中$S^{'}$表示头实体或尾实体被替换的负采样三元组。

4 实验

4.1 数据集与评价指标

MultiNews数据集由(Fabbri et al.,2019)提出，MultiNews数据集由新闻文章和人工撰写的摘要组成。该数据集来自不同的新闻来源(超过1500个网站)。MultiNews更类似于传统的多文档只摘要数据集，如DUC，但规模更大。正如Fabbri等人所述，数据集分为44，972个用于训练的实例，5622个用于验证的实例和5622个用于测试的实例。源文档和输出摘要的平均长度分别为2103.5个标记和263.7个标记。我们将源文档截断为句子S，并按照原始顺序将句子序列连成一个序列。我们使用Stanford coreNLP工具对数据集进行预处理，并采用TAGME实体链接工具提取源文档中的实体token。本文使用F1 ROUGE对生成摘要与标准摘要进行评估。

4.2 基线和实施细节

在对比实验中，本文将提出的模型与现有几种先进的方法进行了比较：Lead是连接标题和排序的段落，并提取前k个标记；LexRank (Erkan et al., 2004)是一个广泛使用的基于图形的抽取式摘要，类似PageRank的算法来排列和选择段落；MMR (Carbonell et al.,1998)提取具有排序列表的句子基于相关性的候选句子和冗余；HIBERT (Zhang et al., 2019) 提出用BERT模型对句子进行预训练，然后对整个文档进行编码；PGN(See et al.,2017)是一个基于RNN的模型，具有注意力机制，允许系统通过指向从源文本复制单词进行抽象概括；Hierarchical Transformer (HT) (Liu et al.,2019)该模型将标题和段落作为输入来产生目标摘要；Hi-MAP(Fabbri et al.,2019)将指针生成器网络模型扩展为分层网络，并集成MMR模块来计算句子级得分；Flat Transformer (FT)是将基于转换器的编码器-解码器模型应用于平面令牌序列的基线；MGSum(Jin et al.,2020)是一个用于抽取式和生成式多文档摘要网络，联合学习了单词、句子和文档的语义表示。CTF+DPP(Perez-Beltrachini et al.,2021)提出基于DPP的注意力模型，将注意力权重用行列式点过程(DPP)给出的概率来计算，并将注意力机制与已有的模型相结合。

本文使用Pytorch来实现提出的模型，使用Stanford coreNLP工具对数据集进行预处理。优化器是Adam (Kingma et al.,2014)，学习率为$2e^{-5}$，$\beta_1 = 0.9$，$\beta_2 = 0.998$。所有模型都在1个GPU上进行500，000步的训练。模型中的所有线性层之前应用概率为0.1的下降。最大序列长度设置为256，batch size大小为32，模型中的隐藏单元的数量被设置为256，前馈隐藏大小为1024，头的数量为8。

4.3 MGNIE整体性能分析：

我们在MultiNews数据集上与先前的几种模型进行了对比，实验结果如下表，MGNIE表示本文提出的模型，它包含嵌入了实体信息的预训练模型，以及多粒度融合信息的摘要模型。MGNIE-BERT将预训练模型换成BERT模型，表明预训练模型中不包含实体关系。MGEIN模型在无论是在ROUGE-1，ROUGE-2还是ROUGE-L，较先前的工作都取得了优秀的性能，并有所提升。

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead	41.24	12.91	18.84
LexRank	38.27	12.70	13.20
MMR	38.77	11.98	12.91
TIBERT	43.86	14.62	18.34
MGSum-ext	44.75	15.75	19.30
MGNIE-BERT	45.75	17.01	20.39
MGNIE	**46.80**	**17.22**	**22.85**

Table 1: 在MultiNews上抽取式模型的对比实验

表1是本文提出的模型与抽取式模型在MultiNews数据集上的对比实验结果，根据结果我们可以发现，本文提出的模型较先前提出的基于抽取式的模型有较大提升。与TIBERT模型相比，我们采用BERT模型做为预训练模型，同时采用多粒度信息融合的方法，较TIBERT模型提升了1.89，表明我们的对多粒度信息进行融合方法对生成式模型的摘要生成有提升作用。与MGSum-ext模型相比，本文提出的模型在多粒度信息融合的基础上，加入了BERT预训练语言模型，较没有预训练模型的摘要模型提升了1.0，同时本文提出的模型在预训练模型中加入了来自知识图的实体信息，进一步取得了模型性能的提升。

Model	ROUGE-1	ROUGE-2	ROUGE-L
PGN	41.85	12.91	16.46
HT	42.36	15.27	22.08
Hi-MAP	43.47	14.89	17.41
FT	44.32	15.11	20.50
MGSum-abs	46.00	16.81	20.09
CTF+DPP	45.84	15.94	21.02
MGNIE-BERT	45.75	17.01	20.39
MGNIE	**46.80**	**17.22**	**22.85**

Table 2: 在MultiNews上生成式模型的对比实验

表2 是本文提出的模型与生成式模型MultiNews数据集上的对比实验结果，根据结果我们可以发现本文提出的模型较先前的具有先进性的生成式模型取得了更好的效果。本文的模型相比于MGSum-abs模型提升了0.8，这表明在多粒度信息融合方面，本文的模型加入了实体信息，并将知识图的实体信息与文本中的实体信息进行融合做预训练模型，更加丰富了文本的上下文表征，从而增强模型理解能力，对于模型性能的提升有更好的影响。可以观察到，在多粒度的基础上将预训练模型换成BERT模型，较于先前的工作有部分提升，而在预训练模型的基础上，不仅对原文本进行模型训练，还加入了外部知识的实体信息，使得预训练获得文本向量表示的效果更加好进一步展开分析，在ROUGE分数评估下，MGNIE以及MGNIE-BERT模型都优于现有的工作模型表现，说明我们采用多粒度的方法进行信息提取获得语义表征，能够挖掘到更多信息特征，从而获得更好的生成摘要效果。

通过观察表1和表2的数据我们可以发现，本文提出的模型对比先前的抽取式模型的提升比生成式模型的效果要好，因为我们使用了融合了丰富知识图信息的预训练模型对原文本进行训练，并且采用的多粒度信息融合来对文本的长距离信息进行了交互，实现了实体的信息增强，对于关键词的解码生成有重要的影响，本文提出的模型生成的摘要文本会相比于抽取式模型生成的摘要更具有连贯性，同时说明了实体信息对于摘要生成的有效性。

4.4 实体信息对摘要性能的影响：

本文通过引入实体信息增强来实现对摘要性能的提升，为了探究实体信息对摘要性能的影响，本文做了相关的消融对比实验。表3给出了不同粒度对实验结果的影响，其中without sent representation表示在多粒度融合中包含词信息和实体信息融合的模型，without entity representation表示在多粒度融合中包含词信息和句子信息融合的模型，其中MGNIE-BERT表示将预训练模型换成BERT模型，不加入外部实体信息的模型。MGNIE表示我们在文中提出的模型，加入外部实体信息增强的多粒度融合模型。

Model	ROUGE-1	ROUGE-2	ROUGE-L
without sent representation	45.51	16.32	19.57
without entity representation	44.02	15.98	19.23
MGNIE-BERT	45.75	16.01	20.39
MGNIE	**46.80**	**17.22**	**22.85**

Table 3: 消融实验对比

从表3的结果中我们可以发现without entity representation不加入实体信息表示表现不佳，表明加入实体表征的是非常有效的。并且MGNIE的表现比MGNIE-BERT要好，表明在预训练模型过程中进行外部实体信息嵌入式对预训练模型的效果有重要提升。通过在预训练模型中加入外部知识图谱的结构化的实体关系信息，以及在自然语言编码是加入实体表征可以充分的挖掘文本中的实体信息以及上下文语义信息，从而是模型获得更好的效果。

4.5 人工评测：

由于评估摘要生成的流畅性以及事实一致性在摘要生成中是十分重要的，所以进行人工评测是必不可少的。具体来说，本文选择5名研究生来对本文生成的摘要进行评估，在MultiNews数据集中随机选择50个样本，为了评估模型的质量，本文选择MGSum模型作为基线模型来进行对比，并从三个方面来进行评估：流畅性(fluency)，信息量(Informativeness)以及与原文本的忠实度(faithfulness)，流畅性是指文本的可读性，包括语法、名词短语和逻辑上的一致性。信息量表示摘要与原文包含的关键内容相关性的数量。忠实度是指摘要与原文的事实一致的相关性。本文选取的评分标准为1-5分，分数越大说明性能越好。

Model	Fluency	Informativeness	Faithfulness
MGSum	3.48	3.26	3.31
MGNIE	**3.83**	**3.58**	**3.96**

Table 4: 人工评测结果

表4是在MGSum模型与本文的模型的人工评估结果，从表中可以观察到本文的模型在流畅性、信息量以及忠实度方面较MGSum都有提升，尤其在忠实度方面的评估结果表明本文提出的基于实体信息增强来提升生成摘要的事实一致性是有效的。

4.6 摘要生成实例分析：

Source Text
Document 1: san francisco (marketwatch)-trading in all nasdaq-listed stocks and options was halted on thursday due to technical problems on the bourse, according to nasdaq omx group (nasdaq:ndaq). **the exchange sent out a series of emails alerting investors that it was experiencing issues with " quote submissions. "** in response, the new york stock exchange has also stopped trading in all nasdaq securities at the request of nasdaq omx. " **all orders in those securities have been cancelled back to customers , " said nyse in a statement.** the nasdaq composite index (nasdaq:comp) was last at 3631.17, up 31.38 points, before trading was suspended. there was no immediate word on when transactions will resume. Document 2 : updated with nyse developments. **"a technical glitch knocked out trading in all nasdaq stock market securities for three hours thursday afternoon, an unprecedented meltdown for a u.s. exchange that paralyzed a broad swath of markets and highlighted the fragility of the financial world's electronic backbone. "nasdaq officials scrambled to figure out what happened and resume trading.** they shared few of their findings with trading firms or the public during regular trading hours, sowing confusion across wall street and leaving many investors frustrated. ", "the decision to reopen trading with about 35 minutes to go before the close came after exchange officials were sure that banks ... "
Gold
nasdaq is back in business after an apparent technical glitch brought the exchange to a rare halt this afternoon for more than three hours, reports the wall street journal. **the exchange hasn't fully explained what happened,** but trading of all nasdaq securities ground to a halt just after noon today, reports marketwatch. other exchanges quickly suspended trading of nasdaq stocks. " **all orders in those securities have been canceled back to customers, " says the new york stock exchange in a statement. nasdaq blamed " quote submissions " in an email to investors.**
Our Model
Nasdaq officials are scrambling to figure out what happened and resume trading in all nasdaq stocks and options, reports marketwatch. **the glitch knocked out of all nasdaq stock market securities for three hours thursday afternoon, an unprecedented meltdown for a us exchange that paralyzed a broad swath of markets and " highlighted the fragility of the financial world's electronic backbone, "** reports the wall street journal. the move came after officials were sure that banks would reopen trading with 35 minutes to go before the close.**"the exchange sent out a series of emails alerting investors that it was working. "** said nasdaq omx group.

Table 5: MGNIE模型摘要生成实例

表5展示了在MGNIE模型上摘要生成的实例与原文本以及标准摘要的对比，加粗的文字表示与原文的关键内容重合的部分。从表中我们可以观察到本文提出的模型在重合度方面与原文内容高度重合，在信息量方面也提取了大量关键信息，捕捉到了“three hours thursday afternoon”这个时间点，以及“the glitch”这个信息点，并且与标准摘要进行对比可以发现，本文模型生成的摘要在内容信息量以及内容重合度都很高。在事实一致性方面，可以发现生成的摘要无论是与原文的对比，还是对标准摘要的对比上都是保持事实一致的。在摘要的流畅度方面，我们可以发现生成的文本是可读的，并且句子之间的连接词使得文本承接上下文具有连贯性以及逻辑性。

5 结论

在本文中，我们针对生成式多文档摘要中存在的缺乏结构化信息的嵌入以及生成摘要的事实不一致性提出了基于实体信息增强以及多粒度信息融合的多文档摘要模型，具体来说，我们加入了预训练模型ERNIE并且将外部知识图中的实体-关系信息嵌入预训练模型以丰富语义信息，与此同时，我们还获取了词信息、实体信息以及句子信息层面的信息融合来编码语言表征，实现了更深层次的信息挖掘，实现了信息增强。最后进行了大量的对比实验表明，本文提出的方法在多文档摘要中取得了有效影响，在一定程度上实现了信息增加解决了事实一致性问题。

在未来的工作中，我们还将考虑将文本与外部知识图结构进行相互融合，将文本转换为结构化的图与外部知识图相连从而获得结构化的信息，来实现基于图结构的多文档摘要模型的性能提升。

参考文献

Alambo A, Lohstroh C, Madaus E, et al. 2020. Alternation. *Topic-centric unsupervised multi-document summarization of scientific and news articles[C]//2020 IEEE International Conference on Big Data (Big Data). IEEE*, 2020: 591-596.

Bordes A, Usunier N, Garcia-Duran A, et al. 2013. Alternation. *Translating embeddings for modeling multi-relational data[J]. Advances in neural information processing systems.*

Carbonell J, Goldstein J. 1998. *The use of MMR, diversity-based reranking for reordering documents and producing summaries[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998: 335-336.

Christensen J, Soderland S, Etzioni O. 2013. *Towards coherent multi-document summarization[C]//Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2013: 1163-1173.

Dong Y, Wieting J, Verga P. 2022. *Faithful to the Document or to the World? Mitigating Hallucinations via Entity-linked Knowledge in Abstractive Summarization[J]*, arXiv preprint arXiv:2204.13761, 2022.

Erkan G, Radev D R. 2004. *Lexrank: Graph-based lexical centrality as salience in text summarization[J]. Journal of artificial intelligence research*, 2004,22: 457-479.

Fabbri A R, Li I, She T, et al. 2019. *Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model[J]*, arXiv preprint arXiv:1906.01749, 2019.

Hendrycks D, Gimpel K. 2016. *Gaussian error linear units (gelus)[J]*, arXiv preprint arXiv:1606.08415.

Jin H, Wang T, Wan X. 2020. *Multi-granularity interaction network for extractive and abstractive multi-document summarization[C]//Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020: 6244-6254.

Kingma D P, Ba J. 2014. *Kingma D P, Ba J. Adam: A method for stochastic optimization[J]*, arXiv preprint arXiv:1412.6980.

Lee J D M C K, Toutanova K. 2018. *Pre-training of deep bidirectional transformers for language understanding[J]*, arXiv preprint arXiv:1810.04805.

Liu Y, Lapata M. 2019. *Hierarchical transformers for multi-document summarization[J]*, arXiv preprint arXiv:1905.13164.

Pasunuru R, Celikyilmaz A, Galley M, et al. 2021. *Data augmentation for abstractive query-focused multi-document summarization[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021)*, 2021: 13666-13674.

Perez-Beltrachini L, Lapata M. 2021. *Multi-document summarization with determinantal point process attention[J]. Journal of Artificial Intelligence Research*, 2021, 71: 371-399.

See A, Liu P J, Manning C D. 2017. *Get to the point: Summarization with pointer-generator networks[J]*, arXiv preprint arXiv:1704.04368.

Szegedy C, Vanhoucke V, Ioffe S, et al. 2016. *Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016: 2818-2826.

Vaswani A, Shazeer N, Parmar N, et al. 2017 *Attention is all you need[J]. Advances in neural information processing systems.*

Yasunaga M, Zhang R, Meelu K, et al. 2017. *Graph-based neural multi-document summarization[J]*, arXiv preprint arXiv:1706.06681.

Zhang X, Wei F, Zhou M. 2019. *HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization[J]*, arXiv preprint arXiv:1905.06566.

Wang Y, Sun Y, Ma Z, et al. 2019. *Zhang Y. Evaluating the factual correctness for abstractive summarization[J]. CS230 Project.*

Zhang Z, Han X, Liu Z, et al. 2019. *ERNIE: Enhanced language representation with informative entities[J]*, arXiv preprint arXiv:1905.07129

Zhao J, Liu M, Gao L, et al. 2019. *Summpip: Unsupervised multi-document summarization with sentence graph compression[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020: 1949-1952.

Zheng X, Sun A, Li J, et al. 2019. *Subtopic-driven multi-document summarization[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019: 3153-3162.