自下而上的抽象概述

塞巴斯蒂安·格尔曼　　邓云天　　亚力山大·拉什
哈佛大学工程与应用科学学院
{gehrmann, dengyuntian, srush} @ seas.harvard.edu

摘要

基于神经网络的抽象摘要方法产生的输出比其他技术更流畅，但在内容选择方面可能较差。 这项工作提出了一种解决此问题的简单技术：使用数据有效的内容选择器来过度确定源文档中应成为摘要一部分的短语。 我们使用此选择器作为自下而上的关注步骤，将模型约束为可能的短语。 我们表明，这种方法提高了压缩文本的能力，同时仍然生成流畅的摘要。 这个两步过程比其他端到端内容选择模型更简单，性能更高，从而显着改善了CNN-DM和NYT语料库的ROUGE。 此外，内容选择器可以用少至1,000个句子进行训练，从而可以轻松地将训练有素的摘要器传送到新域。

1 介绍

文本摘要系统旨在生成以较长文本压缩信息的自然语言摘要。 使用神经网络的方法已经在这个任务上展示了有希望的结果，其中端到端模型编码源文档，然后将其解码为抽象摘要。当前最先进的神经抽象摘要模型通过使用指针生成器样式模型来组合提取和抽象技术，这些模型可以复制源文档中的单词（顾等人。，2016; 看到 等。，2017）。 这些端到端模型产生了流畅的抽象摘要，但在内容选择方面取得了不同的成功，即与完全提取模型相比，决定了总结内容。

从建模角度来看，端到端模型具有吸引力；然而，有证据表明，在总结人们时，请遵循两步

来源文件

德国总理安吉拉·梅克尔[她]在[年度]复活节假期[在意大利]期间对[天气]的温度高达21c，温度高达21c，梅克尔夫人和她的丈夫[天]不感到高兴。 ，化学教授joachim sauer，]不得不沉寂12度。 英国财政大臣和她的[配偶]一直在伊斯基亚岛上的小岛上度过复活节，在地中海的那不勒斯附近停留了十多年。

[不那么阳光明媚：] angela merkel [和]她的丈夫[，化学教授joachim sauer，]被发现在他们的[年度]复活节之旅中，在那不勒斯[，]附近的那不勒斯[。] 这对夫妇[传统上]在岛屿南部的五星级miramare温泉酒店度过他们的假期[来自]，拥有自己的私人海滩[和俯瞰海洋的阳台[。] ...

参考

- angela merkel和丈夫在意大利语时发现岛屿度假。

...

基线方法

- 安格拉梅克尔和她的丈夫，化学教授约阿希姆·绍尔（joachim sauer）被发现在他们一年一度的复活节之旅中，在那不勒斯附近的那不勒斯岛上。
- 安格拉梅克尔和她的丈夫被发现在他们的身上复活节前往伊斯基亚岛，靠近那不勒斯。

...

图1：有和没有自下而上关注的两个句子摘要的示例。 该模型不允许复制[灰色]中的单词，尽管它可以生成单词。 随着自下而上的注意，我们看到更明确的句子压缩，而没有它整个句子被逐字复制。

首先选择重要短语然后解释它们的方法（安德森和希迪，1988; Jing和McKeown，1999）。对图像字幕进行了类似的论证。 在— 儿子等人。 （2017）使用两步法开发最先进的模型，首先预先计算分段对象的边界框，然后将注意力应用于这些区域。 这种所谓的自下而上的关注受到神经科学研究的启发，该研究基于以下属性描述注意力 -

刺激（布施曼和米勒,2007).

在这种方法的推动下，我们考虑自下而上的关注神经抽象概括。 我们的方法首先为源文档选择一个选择掩码，然后通过该掩码约束标准神经模型。 这种方法可以更好地决定模型应该包含在摘要中的哪些短语，而不会牺牲神经抽象总和的流畅性优势。 此外，它需要更少的数据来训练，这使其更适应新的领域。

我们的完整模型包含一个单独的内容选择系统，以决定源文档的相关方面。 我们将此选择任务框定为序列标记问题，目标是从文档中识别作为其摘要一部分的标记。 我们展示了一个基于上下文嵌入的内容选择模型（彼得斯等人。，2018）可以识别正确的令牌，召回率超过60%，精度超过50%。 为了将自下而上的注意力结合到抽象概括模型中，我们采用掩蔽来将单词复制到文本的选定部分，从而产生语法输出。 我们还尝试了多种方法，通过多任务学习或直接结合完全可区分的掩模，将类似的约束结合到更复杂的端到端抽象概括模型的训练过程中。

我们的实验将自下而上的注意力与其他一些最先进的抽象系统进行了比较。 与我们的基线模型相比 看到 等。 (2017自下而上的注意力导致CNN-Daily Mail（CNN-DM）语料库的ROUGE-L得分从36.4提高到38.3，同时训练更简单。 我们也看到了与我们的MLE训练系统最近基于强化学习的方法相比或更好的结果。 此外，我们发现内容选择模型具有非常高的数据效率，并且可以使用少于原始训练数据的1%进行训练。 这为域转移和低资源摘要提供了机会。 我们证明了在CNN-DM上训练并在NYT语料库上进行评估的摘要模型可以在ROUGE-L中提高超过5个点，其中内容选择器仅在1,000个域内句子中进行训练。

2 相关工作

在靠近源文档和允许压缩或抽象修改之间的文档摘要中存在紧张。 许多非神经系统采用选择和压缩方法。 例如， 多尔等人。 (2003）引入了一个系统，该系统首先从新闻文章的第一句中提取名词和动词短语，并使用迭代缩短算法对其进行压缩。 最近的系统如Durrett等人。 (2016）还学习一个模型来选择句子，然后压缩它们。

相比之下，基于神经网络的数据驱动的提取摘要的最近工作集中于提取和排序完整的句子（程 和拉帕塔，2016; Dlikman和Last，2016). NAL— 拉帕蒂等人。 (2016b）使用分类器来确定是否包括一个句子和一个选择器，对正分类的那些进行排名。 这些方法经常过度提取，但在单词级别提取需要保持语法正确的输出（程和拉帕塔,2016），这很难。 有趣的是，在不符合语法的情况下，关键短语提取通常与人为生成的摘要内容紧密匹配（Bui等人。,2016).

第三种方法是使用序列到序列模型的神经抽象摘要（Sutskever等人。，2014; Bahdanau等。,2014）。 这些方法已应用于标题生成等任务（拉什等人。，2015）和文章摘要（Nallapati等。，2016a). 乔普拉 等。 (2016）表明对摘要更具体的注意方法可以进一步提高模型的性能。 顾等人。 (2016）是第一个展示复制机制的人 Vinyals等。 (2015），可以通过从源复制单词来结合提取和抽象摘要的优点。 见等。 (2017）改进这种指针生成器方法并使用额外的覆盖机制（Tu等人。，2016）使模型意识到其注意历史，以防止反复注意。

最近，强化学习（RL）方法优化了除最大似然之外的摘要目标，已经证明可以进一步提高这些任务的性能（保罗斯等人。,2017;李等人。,2018b; Celiky— ilmaz等。,2018). 保罗斯等人。 (2017）通过内部注意来接近覆盖问题，其中解码器关注先前生成的单词。 但是，基于RL的培训可以

很难调整和慢慢训练。 我们的方法不使用RL训练，尽管在理论上这种方法可以适用于RL方法。

一些论文还探讨了多遍提取 - 抽象概括。 Nalla— 帕蒂等人。 (2017）创建一个新的源文档，包含来自源的重要句子，然后训练一个抽象系统。 刘 等。 (2018描述提取完整段落的提取阶段和确定其顺序的抽象阶段。 最后 曾等人。 (2016）引入一种机制，在两次传递中读取源文档，并使用第一次传递的信息来偏置第二次传递。 我们的方法的不同之处在于我们使用完全抽象的模型，偏向于强大的内容选择器。

最近的其他工作探讨了内容选择的替代方法。 例如， 科汉 等。 (2018）使用分层注意来检测文档中的相关部分， 李等人。 (2018a)生成一组用于指导摘要过程的关键字，以及Pasunuru和Bansal (2018根据是否在摘要中包含显着关键字来开发损失函数。 其他方法调查句子级别的内容选择。 Tan等人。 (2017）描述一个基于图表的注意力，一次一个句子，陈和班萨尔 (2018）首先从文档中提取完整的句子然后压缩它们，然后 Hsu等人。 (2018根据句子在摘要中的可能性来调节注意力。

3 背景：神经总结

在本文中，我们考虑一组对文本 （X， Y），其中$x \in X$对应于源令牌$x_1, \ldots, x_n$和$y \in Y$到摘要$y_1, \ldots, y_m$，$m << n$。

抽象摘要生成一个单词一次。 在每个时间步，模型都知道先前生成的单词。 问题是学习由 θ 参数化的函数f（x），其最大化生成正确序列的概率。 在之前的工作之后，我们使用注意序列到序列模型对抽象概括进行建模。 在神经网络内计算的用于解码步骤j的注意分布p（a_jx，$y_{1:j-1}$）表示在所有源令牌上的嵌入式软分布，并且可以被解释为模型的当前焦点. .

该模型还有一个复制机制 -

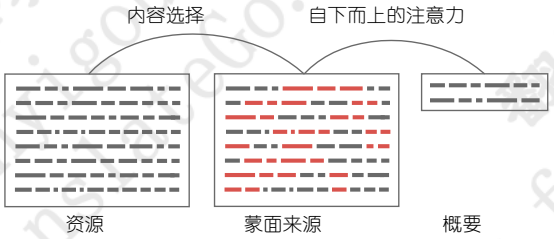

图2：整个部分描述的选择和生成过程概述 4.

nism （Vinyals等。，2015）从源头复制单词。复制模型通过预测确定模型是复制还是生成的二进制软开关z_j来扩展解码器。 复制分布是源文本上的概率分布，联合分布被计算为模型的两个部分的凸组合，

$$p(y_j \mid y_{1:j-1}, x) = p(z_j = 1 \mid y_{1:j-1}, x) \times p(y_j \mid z_j = 1, y_{1:j-1}, x) + p(z_j = 0 \mid y(TF218), x) \times p(y_j \mid z_j = 0, y_{1:j-1}, x) \quad (1)$$

其中两部分分别代表复制和生成分布。 遵循指针生成器模型 见等。 (2017），我们重用注意力p（a_jx，$y_{1:j-1}$）分布作为复制分布，即通过复制注意力在源w中的令牌的复制概率被计算为对所有出现的关注的总和 W上。 在训练期间，我们使用潜在开关变量最大化边际可能性。

4 自下而上的注意力

接下来我们将考虑将内容选择合并到抽象摘要中的技术，如图所示 2.

4.1 内容选择

我们将内容选择问题定义为单词级提取摘要任务。 虽然在自定义提取摘要方面做了大量工作（参见相关工作），但我们做了一个简化的假设并将其视为序列标记问题。 设$t_1, \ldots, t_n$表示每个源令牌的二进制标签，即如果在目标序列中复制一个字则为1，否则为0。

虽然此任务没有监督数据，但我们可以通过将摘要与文档对齐来生成培训数据。 我们将单词x_i定义为

复制如果（1）它是标记s = $x_{I-J:I:I+K}$的最长可能子序列的一部分，对于整数ji; k（xi），如果sx和sy，和∈ ∈
（2）没有先前的序列u，s = u。

我们使用标准双向LSTM模型训练，最大可能性为序列标记问题。 最近的结果表明，更好的单词表示可以显着改善序列标记任务的性能（彼得斯等人。，2017）。 因此，我们首先将每个标记w_i映射到两个嵌入通道

$e_i^{(w)}$ 和$e^{(c)}$。 $e^{(w)}$嵌入表示a

预训练词嵌入的静态通道，
例如GLoVE（Pennington等。，2014）。 $e^{(c)}$是来自预训练语言的上下文嵌入

模型，例如ELMo（彼得斯等人。，2018）使用字符识别标记嵌入（Kim等人。，2016）接下来是两个双向LSTM铺设 -

ers $h^{(1)}$和$h^{(2)}$。 上下文嵌入是

微调以学习任务特定的嵌入$e^{(c)}$作为每个LSTM层的状态和令牌嵌入的线性组合，

$$e^{(c)} = \gamma \times s_j \times h^{(R)}, \quad R=0$$

以γ和$s_{0,1,2}$为可训练参数。 由于这些嵌入仅向标记器添加了四个附加参数，因此尽管具有高维嵌入空间，但它仍然具有非常高的数据效率。

两个嵌入都连接成单个向量，用作双向LSTM的输入，双向LSTM计算单词w_i的表示h_i。然后，我们可以计算出具有可训练参数W_s和b_s的单词被选择为 σ（$W_s h_i + b_s$）的概率q_i。

4.2 自下而上复制注意

灵感来自自下而上的图像工作（安德森等人。，2017）它限制了对图像内预定边界框的注意，我们使用这些注意掩模来限制指针 - 发生器模型的可用选择。 如图所示 1，神经复制模型的一个常见错误是复制很长的序列甚至整个句子。 在基线模型中，超过50%的复制令牌是超过10个令牌的复制序列的一部分，而参考摘要的此数字仅为10%。 虽然自下而上的注意力也可用于修改源编码器表示，但我们发现了标准编码器全文在聚合方面是有效的，因此将自下而上的步骤限制为注意力掩盖。

具体地说，我们首先在完整数据集以及上面定义的内容选择器上训练指针生成器模型。在推理时，为了生成掩码，内容选择器计算源文档中的每个标记的选择概率$q_{1:n}$。 选择概率用于修改复制关注分布以仅包括由选择器标识的标记。 设a^i表示解码步骤j对编码器的注意

我是一个字。 给定阈值E，选择是

作为一个硬面具，这样

$$p(a^i \sim | x, y_{1:j-1}) = \begin{cases} p(a^i_j / x, y) & q\lambda_{j-1} \\ 0 & \end{cases}$$

流。

确保方程式 1 仍然产生正确的概率

我们首先乘以p（$\sim_j$x，$y_{1:j-1}$）
通过归一化参数λ然后重新归一化分布。 得到的归一化分布可用于直接替换a作为新的复制概率。

4.3 端到端替代方案

两步BOTTOM-UP注意力具有培训简单性的优点。但理论上，标准副本应该能够学习如何在端到端培训中执行内容选择。 我们考虑将其他内容选择纳入神经训练的其他几种端到端方法。

方法1 :(仅限MASK)：我们首先考虑自下而上方法中使用的对齐是否有助于标准摘要系统。 灵感来自 Nallapati等。(2017)，我们调查在训练期间是否对齐摘要和来源并修复黄金副本注意选择“正确”源词是有益的。 我们可以将这种方法视为将可能的副本集限制为固定的源字。 此处培训已更改，但在测试时未使用任何掩码。

方法2 (MULTI-TASK)：接下来，我们研究内容选择器是否可以与抽象系统一起训练。我们首先通过将摘要作为多任务问题进行测试并使用相同的特征训练标记器和摘要模型来测试该假设。 对于此设置，我们使用共享编码器进行抽象摘要和内容选择。 在测试时，我们

应用与自下而上注意相同的掩蔽方法。

方法3（DIFFMASK）：最后我们考虑在训练期间使用面罩端对端地训练整个系统。 在这里，我们共同优化两个目标，但使用预测的选择概率来轻柔地掩盖复制注意力 $p(a_j^{\sim} \mid x, y_{1:j-1}) = p(a_j \mid x, y_{1:j-1}) \times q_i$，其中 j j 完全可区分的模型。 该模型在测试时使用相同的软掩模。

5 推理

一些作者已经注意到，长形神经生成仍然存在长度不正确和重复单词的重要问题，而不是像翻译这样的短期问题。 建议的解决方案包括修改具有扩展的模型，例如覆盖机制（Tu等人。，2016; 见等。，2017）或句内注意（Cheng等。，2016; 保罗斯等人。，2017）。 我们坚持修改推理的主题，并修改评分函数以包括长度惩罚lp和覆盖惩罚cp，并定义为 $s(x, y) = \log p(y \mid x) / lp(x) + cp(x; y)$。

长度：鼓励产生更长时间在序列中，我们在波束搜索期间应用长度归一化。 我们使用长度惩罚 吴 等。 (2016)，制定为

$$lp(y) = \frac{(5 + |y|)^{\alpha}}{(5 + 1)^{\alpha}},$$

使用可调参数α，其中增加α会导致更长的摘要。 我们还根据训练数据设置了最小长度。

重复：复制模型经常重复使用相同的源令牌，多次生成相同的短语。 我们引入一个新的摘要特定覆盖惩罚，

$$cp(x; y) = \beta \sum_{i=1}^{n} \max\left(\sum_{j=1}^{m} a_{ij}, 1.0\right).$$

直观地，只要解码器将序列内的总注意力超过1.0指向单个编码令牌，该惩罚就会增加。 通过选择足够高的 β，这种惩罚会阻止摘要，只要它们会导致重复。 另外，我们遵循（保罗斯等人。，2017并限制光束搜索从不重复三元组。

6 数据和实验

我们评估我们在CNN-DM语料库上的方法（赫尔曼等人。，2015; Nallapati等。，2016a）和NYT语料库（桑德豪斯，2008），它们都是新闻摘要的标准语料库。 CNN-DM语料库的摘要是各自所示文章的要点网站，而NYT语料库包含总和 - 由图书馆学家撰写的maries。 CNN-DM摘要是完整的句子，平均66个令牌（σ = 26）和4.9个子弹点。 NYT摘要并不总是完整的句子而且更短，平均有40个令牌（σ = 27）和1.9个要点。 最近的工作使用了CNN-DM上的匿名和非匿名版本，因此直接比较可能很困难。以下 见等。 (2017)，我们使用此语料库的非匿名版本并将源文档截断为400个令牌，并将目标摘要截断为训练和验证集中的100个令牌。 对于使用NYT语料库的实验，我们使用了描述的预处理 保罗斯等人。 (2017此外，删除作者信息并将源文档截断为400个令牌而不是800.这些更改导致每篇文章平均326个令牌，比使用800个令牌截断文章的549个令牌减少。所有型号的目标（非复制）词汇量限制为50,000个令牌。

内容选择模型使用预先训练的 GloVe嵌入大小为100，ELMo为 大小为1024.双-LSTM有两层，隐藏大小为256. Dropout设置为0.5，模型用Adagrad训练，初始学习率为0.15，初始累加器值为0.1。 我们将语料库中的训练样例数量限制为100,000，这对性能影响很小。 对于联合训练的内容选择模型，我们使用与抽象模型相同的配置。

对于基础模型，我们重新实现了 如下所述的指针生成器模型 见等。 (2017)。 要有相同数量的参数 - 对于以前的工作，我们在单层LSTM中使用具有256个隐藏状态的编码器，对于单层解码器使用512个隐藏状态。 嵌入大小设置为128. 我们发现增加模型大小或将模型更改为Transformer（Vaswani等。，2017）可以导致性能略有提高，但代价是增加培训时间和参数。 该

方法	R-1	R-2	RL
指针生成器（见等。，2017）	36.44	15.66	33.42
指针生成器+覆盖范围（见等。，2017）	39.53	17.28	36.38
ML +内部注意力（保罗斯等人。，2017）	38.30	14.81	35.49
ML + RL（保罗斯等人。，2017）	39.87	15.82	36.90
显着性+蕴涵奖励（Pasunuru和Bansal，2018）	40.43	18.00	37.10
关键信息指南网络（李等人。，2018a）	38.95	17.12	35.68
不一致性损失（Hsu等人。，2018）	40.68	17.97	37.13
句子重写（陈和班萨尔，2018）	40.88	17.80	**38.54**
指针生成器（我们的实现）	36.25	16.17	33.41
指针生成器+覆盖惩罚	39.12	17.35	36.12
指针生成器+仅掩码	37.70	15.63	35.49
指针生成器+多任务	37.67	15.59	35.47
指针生成器+ DiffMask	38.45	16.88	35.81
自下而上的总结	**41.22**	**18.68**	38.34

表1：CNN-DM数据集上抽象摘要的结果。[2] 第一部分显示了用交叉熵训练的编码器 - 解码器抽象基线。第二部分描述了基于强化学习的方法。 第三部分介绍了我们的基线和本工作中描述的注意力掩蔽方法。

使用与内容选择器相同的Adagrad配置训练模型。 另外，一旦验证困惑在一个纪元之后没有减少，学习速率在每个纪元之后减半。 我们不使用dropout并使用最大范数为2的渐变裁剪。

所有推理参数都在验证集的200个句子子集上进行调整。 长度罚分参数 α 和复制掩模E在模型和基线之间不同， α 范围从0.6到1.4，E范围从0.1到0.2。 CNN-DM的生成摘要的最小长度设置为35，NYT的最小长度设置为6。虽然指针生成器使用5的光束尺寸并且没有用更大的光束改善，但我们发现自下而上的注意力需要更大的光束尺寸并将其设置为10.覆盖惩罚参数 β 设置为10，并且对于两种方法，将注意归一化参数 λ 复制到2。

我们使用AllenNLP（Gardner等人。，2018）对于内容选择器，抽象模型在OpenNMT-py中实现（Klein等。，2017）.[3].

[3]可以在以下位置找到代码和复制指令 https://github.com/sebastianGehrmann/ 自下而上，汇总

[3]这些结果比较了这个语料库的非匿名化版本（见等。，2017）。 匿名版本的最佳结果是R1：41.69 R2：19.47 RL：37.92 from（Celikyilmaz等。，2018）。 我们将它们与NYT语料库中的DCA模型进行比较。

7 结果

表 1 显示我们在CNN-DM语料库上的主要结果，顶部显示抽象模型，底部显示自下而上注意方法。 我们首先观察到使用覆盖推断惩罚分数与完全覆盖机制相同，而不需要任何其他模型参数。 我们发现，我们的端到端模型都没有导致改进，这表明在训练期间很难应用掩蔽而不会损害训练过程。 仅对Mask模型具有更强的复制机制监控功能，与MultiTask模型非常相似。 另一方面，自下而上的注意力导致所有三个分数的重大改善。 虽然我们期望更好的内容选择主要用于改进ROUGE-1，但事实上三者都增加了一些暗示流畅性没有受到特别伤害的暗示。 我们的交叉熵训练方法甚至优于ROUGE-1和2中所有基于强化学习的方法，而最高报告的ROUGE-L得分由 陈和班萨尔（2018）落在我们结果的95％置信区间内。

表 2 显示了在NYT语料库中使用相同系统的实验。 我们看到，与基线指针生成器最大似然方法相比，2点改进延续到此数据集。 在这里，模型表现优异

方法	R-1	R-2	RL
ML*	44.26	27.43	40.41
ML+RL*	47.03	30.72	**43.10**
DCA†	**48.08**	31.19	42.33
Point.Gen。 +覆盖笔。	45.13	30.13	39.67
自下而上的总结	47.38	**31.23**	41.81

表2：NYT语料库的结果，我们将其与RL训练的模型进行比较。 *标记模型和结果 保罗斯等人。(2017）和†的结果 Celikyilmaz 等。 (2018).

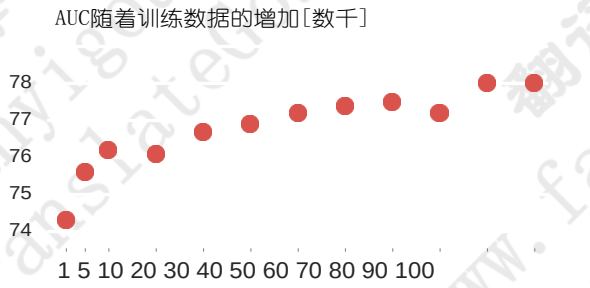

图3：在CNN-DM上训练的内容选择器的AUC，其不同的训练集大小范围从1,000到100,000个数据点。

	AUC	R-1	R-2	RL
CNNDM		25.63	11.40	20.55
+1k	80.7	30.62	16.10	25.32
+10k	83.6	32.07	17.60	26.75
+100k	86.6	33.11	18.57	27.69

表3：域转移实验的结果。 显示内容选择器的AUC号码。 ROUGE分数代表在CNN-DM上训练并在NYT上评估的抽象模型，其中附加的复制约束训练在NYT语料库的1/10 / 100k训练样本上。

基于RL的模型 保罗斯等人。 (2017) in ROUGE-1和2，但不是L，并且具有可比性到（的结果）Celikyilmaz等。，2018） 除对于ROUGE-L。 同样可以观察到比较ML和我们的指针生成器。 我们怀疑总结长度有所不同

我们的推理参数选择会导致这种差异，但无法访问他们的模型或摘要来调查此声明。 这表明，自下而上的方法即使对于针对概要特定目标进行培训的模型也能获得有竞争力的结果。

自下而上总结的主要好处似乎来自减少错误复制的单词。 使用最佳的Pointer-Generator模型，复制单词的精度与参考相比为50.0%。 这种精度提高到52.8%，这主要推动了R1的增加。 独立样本t检验显示，这种改善在统计学上显着，$t = 14.7$ ($p < 10^{-5}$)。我们还观察到，与指针生成器相比，添加内容选择时，摘要的平均句子长度从13个字减少到12个字，同时保持所有其他推理参数不变。

域名转移虽然端到端培训已经很普遍，但两步骤方法也有好处。 由于内容选择器只需要解决预训练向量的二进制标记问题，即使训练数据非常有限，它也能很好地运行。 如图所示 3该模型只有1000个句子，达到了超过74的AUC。超过这个大小，模型的AUC随着训练数据的增加而略有增加。

为了进一步评估内容选择，我们考虑应用于域名转移。 在本实验中，我们应用指针生成器在CNN-DM上接受了NYT语料库的培训。 此外，我们在NYT集的1,10和10万个句子中训练三个内容选择器，并在自下而上的摘要中使用它们。 结果如表所示 3，表明即使是在最小子集上训练的模型也能比模型提高近5个点而无需自下而上的注意力。 这种改进随着较大的子集而增加，最多可达7个点。 虽然这种方法没有达到与直接在NYT数据集上训练的模型相当的性能，但它仍然代表了未增加的CNN-DM模型的显着增加，并产生了相当可读的摘要。 我们在附录中显示了两个示例摘要 A。此技术可用于低资源域和有限数据可用性的问题。

8 分析与讨论

内容选择的摘录摘要？ 鉴于内容选择器与抽象模型一起有效，有趣的是知道它是否已经自己学习了有效的提取摘要系统。 表 4 显示比较内容选择的实验

方法	R-1	R-2	RL
LEAD-3	40.1	17.5	36.3
NEUSUM (周等人。, 2018)	41.6	19.0	38.0
前三名（续。选择。）	40.7	**18.0**	37.0
Oracle短语选择器	67.2	37.8	58.2
内容选择器	**42.0**	15.9	**37.3**

表4：CNN-DM数据集上的提取方法的结果。 第一部分显示句子提取分数。 如果内容选择器根据我们的匹配启发法选择了所有正确的单词，则第二部分首先显示oracle分数。 最后，我们在内容选择器提取高于选择概率阈值的所有短语时显示结果。

提取基线。 LEAD-3基线是新闻摘要中常用的基线，它从文章中提取前三个句子。 前3个显示了当我们从选择器中通过平均复制概率提取前三个句子时的性能。 有趣的是，使用这种方法，前三个句子中只有7.1%不在前三个范围内，进一步增强了LEAD-3基线的强度。 我们的天真句子提取器的表现略差于最高报告的提取分数 周等人。 (2018）专门训练以评分句子组合。 最后的条目显示当提取高于阈值的所有单词时的性能，使得得到的摘要大约是参考摘要的长度。 如果我们的模型具有完美的准确性，oracle得分代表结果，并且表明内容选择器在产生竞争结果的同时，在未来的工作中有进一步改进的空间。 这个结果表明该模型在找到重要单词（ROUGE-1）方面非常有效，但在将它们链接在一起效果较差（ROUGE-2）。 如同 保罗斯等人。 (2017），我们发现ROUGE-2的减少表明生成的摘要缺乏流畅性和语法性。 一个典型示例如下所示：

> 36年来，他的第一个汉堡包食错了。现年69岁的迈克尔·汉纳因1980年因法官指控射杀卡车司机麦克加里而被判谋杀罪。

这个特殊的不合语法的例子的ROUGE-1为29.3。这进一步突出了组合方法的好处，

数据	%小说	动词	名词	副官
参考	14.8	30.9	35.5	12.3
香草S2S	6.6	14.5	19.7	5.1
指针发电机	2.2	25.7	39.3	13.9
自下而上的注意力	0.5	53.3	24.8	6.5

表5：%Novel显示了a中单词的百分比

不在源文档中的摘要。 最后三列显示了生成的摘要中的新词的词性标签分布。

预测被抽象系统流畅地链接在一起。 但是，我们还注意到抽象系统需要访问完整的源文档。 我们尝试使用内容选择的输出作为抽象模型的训练输入的蒸馏实验表明模型性能急剧下降。

复制分析虽然Pointer-Generator模型具有摘要摘要的能力，但复制机制的使用导致摘要主要是提取的。 表 5 表明，通过复制，不在源文档中的生成单词的百分比从6.6%降低到2.2%，而参考摘要更具抽象性，14.8%的新单词。 自下而上的注意力进一步减少到只有百分之五。 然而，由于生成的摘要通常不超过40-50个单词，因此有和没有自下而上注意的抽象系统之间的差异小于每个摘要的一个新单词。 这表明抽象模型的好处在于它们产生更好的释义的能力较少，但更多的是能够从大多数提取过程中创建流畅的摘要。

表 5 还显示了小说生成的单词的词性标签，我们可以观察到有趣的效果。 自下而上注意力的应用导致新形容词和名词的急剧减少，而动词新词的比例急剧增加。 在查看正在生成的新动词时，我们注意到非常高的时态或数字变化百分比，由“说”这个词的变化表示，例如“说”或“说”，而新名词主要是形态变体源中的单词。

数字 4 显示正在复制的短语的长度。 虽然大多数复制短语在

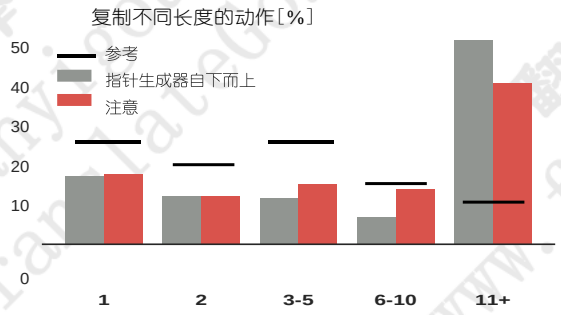

图4：对于所有复制的单词，我们显示了它们所属的复制短语长度的分布。 黑色线条表示参考摘要，条形图摘要有和没有自下而上的注意。

参考摘要以1到5个字为一组，指针生成器复制许多非常长的序列和超过11个单词的完整句子。 由于内容选择掩码中断了大多数长拷贝序列，因此模型必须仅使用生成概率生成未选择的单词，或者使用不同的单词。 虽然我们在生成的摘要中经常观察到这两种情况，但非常长的复制短语的比例会减少。 然而，无论是否有自下而上的注意，复制短语的长度分布仍然与参考完全不同。

推理惩罚分析我们接下来分析推理时间损失函数的影响。 表 6 当一次添加一个惩罚时，表示对简单指针生成器的边际改进。 我们观察到，即使在其他两个分数之上添加，所有三个分数都会提高所有三个分数。 这进一步表明未修改的指针生成器模型已经学习了抽象概括问题的适当表示，但受到无效内容选择和推理方法的限制。

9 结论

这项工作提供了一个简单但准确的摘要内容选择模型，用于识别文档中可能包含在摘要中的短语。 我们发现这个内容选择器可以用于自下而上的注意，这限制了抽象的和函数从源复制单词的能力。 自下而上的综合摘要系统导致CNN-DM和NYT语料库的ROUGE得分均超过两分。 一个

数据	R-1	R-2	RL
指针生成器	36.3	16.2	33.4
+长度惩罚	38.0	16.8	35.0
+承保罚款	38.9	17.2	35.9
+ Trigram重复	39.1	17.4	36.1

表6：一次添加一个推理惩罚时CNN-DM的结果。

与端到端训练方法的比较表明，这个特定问题不能用单个模型轻易解决，而是需要微调推理限制。 最后，我们发现这种技术由于其数据效率，可用于调整具有少量数据点的训练模型，从而可以轻松地转移到新域。 在需要内容选择的其他领域（如语法修正或数据到文本生成）中研究类似的自下而上方法的初步工作已显示出一些前景，并将在未来的工作中进行调查。

致谢

我们要感谢Barbara J. Grosz就这项工作的早期阶段提供了有益的讨论和反馈。 我们进一步感谢三位匿名审稿人。 这项工作得到了三星研究奖的支持。 YD的部分资金来自彭博研究奖。 SG的部分资金来自NIH资助5R01CA204585-02。

参考

Peter Anderson，何晓东，Chris Buehler，Damien Teney，Mark Johnson，Stephen Gould和张磊。 2017.图像字幕和vqa的自下而上和自上而下的关注。 arXiv preprint arXiv：1707.07998。

Valerie Anderson和Suzanne Hidi。 1988.教学生总结。 教育领导，46（4）：26-28。

Dzmitry Bahdanau，Kyunghyun Cho和Yoshua Bengio。 通过联合学习对齐和翻译的神经机器翻译。 arXiv preprint arXiv：1409.0473。

Duy Duc An Bui，Guilherme Del Fiol，John F Hurdle和Siddhartha Jonnalagadda。 2016.摘要文本摘要系统，以帮助系统评价开发中的全文数据提取。 Journal of生物医学信息学，64：265-272。

Timothy J Buschman和Earl K Miller。 自上而下与自下而上的自我关注控制

前额叶和后顶叶皮质。 科学，315（5820）：1860-1862。

Asli Celikyilmaz，Antoine Bosselut，Xiaodong He和Yejin Choi。 2018.用于抽象摘要的深层沟通代理。 在2018年计算语言学协会北美分会会议记录：人类语言技术，第1卷（长篇论文），第1卷，第1662-1675页。

陈仁春和莫希特班萨尔。 2018.快速抽象摘要与强化选择的句子重写。 arXiv preprint arXiv：1805.11080。

Jianpeng Cheng，Li Dong和Mirella Lapata。 2016.用于机器读取的长期短期记忆网络。 arXiv preprint arXiv：1601.06733。

Jianpeng Cheng和Mirella Lapata。 2016.通过提取句子和单词进行神经总结。 arXiv preprint arXiv：1603.07252。

Sumit Chopra，Michael Auli和Alexander M Rush。2016.用细心的递归神经网络进行抽象句子总结。在计算语言学协会北美分会2016年会议论文集：人类语言技术，第93-98页。

Arman Cohan，Franck Dernoncourt，Doo Soon Kim，Trung Bui，Seokhwan Kim，Walter Chang和Nazli Goharian。 2018.一种用于长文档抽象概括的话语意识关注模型。 在2018年计算语言学协会北美分会会议记录：人类语言技术，第2卷（短文），第2卷，第615-621页。

Alexander Dlikman和Mark Last。 2016.在单文档提取摘要中使用机器学习方法和语言特征。在DMNLP @ PKDD / ECML，第1-8页。

Bonnie Dorr，David Zajic和Richard Schwartz。对冲修剪器：标题生成的解析和修剪方法。 在文本摘要研讨会上的HLT-NAACL 03论文集 - 第5卷，第1-8页。 计算语言学协会。

Greg Durrett，Taylor Berg-Kirkpatrick和Dan Klein。 2016.基于学习的单文档摘要，具有压缩和回发约束。 arXiv preprint arXiv：1603.08887。

Matt Gardner，Joel Grus，Mark Neumann，Oyvind Tafjord，Pradeep Dasigi，Nelson Liu，Matthew Peters，Michael Schmitz和Luke Zettlemoyer。 2018. Allennlp：深层语义自然语言处理平台。 arXiv preprint arXiv：1803.07640。

顾嘉涛，郑正东，李航和Victor OK Li。 2016.在序列到序列学习中加入复制机制。 arXiv preprint arXiv：1603.06393。

Karl Moritz Hermann，Tomas Kocisky，Edward Grefenstette，Lasse Espeholt，Will Kay，Mustafa Suleyman和Phil Blunsom。 2015.教学机器阅读和理解。 在神经信息处理系统的进展，第1693-1701页。

徐婉婷，林杰凯，李明英，柯瑞敏，唐静，孙敏。2018.使用不一致性损失的抽取和抽象概括的统一模型。 arXiv preprint arXiv：1805.06266。

Hongyan Jing和Kathleen R McKeown。 1999.人类写的总结句的分解。 在第22届年度国际ACM SIGIR信息检索研究与开发会议论文集中，第129-136页。

Yoon Kim，Yacine Jernite，David Sontag和Alexander M Rush。 2016.字符感知神经语言模型。 在AAAI，第2741-2749页。

Guillaume Klein，Yoon Kim，Yuntian Deng，Jean Senellart和Alexander M Rush。2017.Opennmt：用于神经机器翻译的开源工具包。 arXiv preprint arXiv：1701.02810。

李晨亮，徐伟然，司力，盛高。 2018A。 基于关键信息引导网络引导抽象文本摘要生成。 在2018年计算语言学协会北美分会会议记录：人类语言技术，第2卷（短文），第2卷，第55-60页。

李碧芝，李立兵，和林。 2018B。 基于演员评论的抽象概括培训框架。 arXiv preprint arXiv：1803.11070。

Peter J Liu，Mohammad Saleh，Etienne Pot，Ben Goodrich，Ryan Sepassi，Lukasz Kaiser和Noam Shazeer。 2018.通过总结长序列生成维基百科。 arXiv preprint arXiv：1801.10198。

Ramesh Nallapati，Feifei Zhai和Bowen Zhou。2017. Summarunner：一种基于递归神经网络的序列模型，用于文档的提取摘要。 在AAAI，第3075-3081页。

Ramesh Nallapati，Bowen Zhou，Caglar Gulcehre，Bing Xiang，et al。 2016a。使用序列到序列rnns及以后的抽象文本摘要。 arXiv preprint arXiv：1602.06023。

Ramesh Nallapati，Bowen Zhou和Mingbo Ma。2016B。 分类或选择：用于提取文档摘要的神经架构。 arXiv preprint arXiv：1611.04244。

Ramakanth Pasunuru和Mohit Bansal。 2018.多奖励强化摘要，具有显着性和蕴涵性。 在2018年计算语言学协会北美分会会议论文集：人类语言

Technologies, Volume 2 (Short Papers)，*第2卷，第646-653页。*

Romain Paulus, Caiming Xiong和Richard Socher。2017.一个深度强化的抽象概括模型。 arXiv preprint arXiv: 1705.04304。

Jeffrey Pennington，Richard Socher 和 Christopher Manning。 手套：词汇表示的全球载体。 在2014年自然语言处理经验方法会议论文集 (EMNLP)，第1532-1543页。

Matthew E Peters，Waleed Ammar，Chandra Bhagavatula和Russell Power。 2017.使用双向语言模型的半监督序列标记。 arXiv preprint arXiv: 1705.00108。

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee和Luke Zettlemoyer。 2018.深层语境化词汇表示。 arXiv preprint arXiv: 1802.05365。

Alexander M Rush, Sumit Chopra和Jason Weston。用于抽象句子摘要的神经注意模型。 arXiv preprint arXiv: 1509.00685。

埃文桑德豪斯。 2008年。纽约时报注释语料库。 语言数据联盟，费城，6 (12) : e26752。

Abigail See，Peter J Liu和Christopher D Manning。 2017.重点：使用指针生成器网络进行汇总。 arXiv preprint arXiv: 1704.04368。

Ilya Sutskever, Oriol Vinyals和Quoc V Le。用神经网络进行序列学习的序列。 在神经信息处理系统的进展中，第3104-3112页。

谭继伟，万晓军，肖建国 2017.使用基于图的注意神经模型的抽象文档摘要。 在计算语言学协会第55届年会论文集 (第1卷：长篇论文)，第1卷，第1171-1181页。

Zhaopeng Tu, Lu Zhengdong Lu, Yang Liu, Xiaohua Liu和Hang Li。 2016.神经机器翻译的建模覆盖范围。 arXiv preprint arXiv: 1601.04811。

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit，Llion Jones，Aidan N Gomez，ŁukaszKaiser和Illia Polosukhin。2017.注意力就是你所需要的。 在神经信息处理系统的进展，第6000-6010页。

Oriol Vinyals，Meire Fortunato 和 Navdeep Jaitly。 2015.指针网络。 在神经信息处理系统的进展，第2692-2700页。

Wu Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le，Mohammad Norouzi，Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al。 2016.谷歌的神经机器翻译系统：缩小人机翻译的差距。 arXiv preprint arXiv: 1609.08144。

曾文元，罗文杰，Sanja Fidler和Raquel Urtasun。2016.使用read-again和copy机制进行高效的汇总。 arXiv preprint arXiv: 1611.03382。

周庆宇，南阳，傅茹，黄少汉，周明，赵铁军。2018.通过联合学习评分和选择句子的神经文档摘要。 在计算语言学协会第56届年会论文集 (第1卷：长篇论文)，第1卷，第654-663页。

例子	生成的摘要
参考	绿湾包装工队成功的赛季很大程度上归功于四分卫布雷特
S2S	艾曼·格林在00-00战胜巨人队时冲上了000码。 真正，dorsey levens，足以让大多数球队开始，但现在是绿色的替补，贡献了00,00和00码的开球回报。
内容选择	季后赛的绿湾包装工队在00-00胜利中击败了巨人队。 包装工队赢了三场比赛，六场比赛。
参考	保罗，视觉人类学的先驱，在00岁去世
S2S	保罗的早期修炼者保罗在十二月去世。 00在他在曼哈顿的家中。他入伍海军，训练他作为密码分析师并将他安置在澳大利亚。
内容选择	保罗人，早期的人类学实践者，开创了玛格丽特·米德。

表7：域转移示例。

A 域转移示例

我们在表中提供了两个生成的CNN-DM到NYT域转移实验的摘要 7。 S2S指的是在CNN-DM上训练的具有覆盖惩罚的指针生成器，其在NYT数据集上得分为20.6 ROUGE-L。 内容选择将其改进为27.7 ROUGE-L，而无需对S2S模型进行任何微调。