一种话题敏感的抽取式多文档摘要方法[1]

应文豪，李素建，穗志方

(北京大学 计算语言学教育部重点实验室，北京市 100871)

摘要：抽取式摘要的核心问题在于合理地建模句子，正确地判断句子重要性。本文提出一种计算句子话题重要性的方法，通过分析句子与话题的语义关系，判断句子是否描述话题的重要信息。针对自动摘要任务缺乏参考摘要作为训练数据的问题，本文提出一种基于排序学习的半监督训练框架，利用大规模未标注新闻语料训练模型。在 DUC2004 多文档摘要任务上的实验结果表明，本文提出的话题重要性特征能够作为传统启发式特征的有效补充，改进摘要质量。

关键词：文本自动摘要；卷积神经网络；排序学习

中图分类号：TP391　　**文献标识码：**A

A Topic-sensitive Extractive Method for Multi-document Summarization

Wenhao Ying, Suijian Li, Zhifang Sui

(MOE Key Laboratory of Computational Linguistics, Peking University,
Beijing 100871, China)

Abstract: The key of extractive summarization is to judge the importance of sentences, which depends on reasonable sentence modelling. This paper proposes a method to model the relations between a sentence and its topics, and evaluates the topical importance of the sentences. To deal with the lack of gold references, the paper proposes a semi-supervised training framework based on learning-to-rank, which is able to exploit a lot of unlabeled news documents. The experiments on DUC2004 multi-document summarization data verify that the proposed feature of topical importance is an effective supplement to heuristic features and can improve the summarization performance.

Key words: Text Summarization; Convolutional Neural Network; Learning to Rank

1　引言

抽取式多文档摘要的核心问题是合理地建模句子，正确地判断句子重要性。为了能够正确地判断句子内容的重要性，研究者通过分析句子在原文出现的位置、句子词汇的词频等方面[1]，抽取启发式特征建模句子，度量句子的重要性。尽管这些启发式特征十分简单，但是在抽取式摘要任务上表现出了良好的效果。

另一方面，句子与话题的语义关联对于句子重要性的计算也是十分重要的。话题是对一组相关事件的抽象概述，文章通过组织若干主、次要话题表达其主旨。在抽取摘要时，分析句子所属话题以及句子与话题之间的关系，有助于找到能够概述原文的摘要句。Chambers 等指出事件具有框架结构，在事件框架下存在若干事件槽或事件角色，描述了该框架最主要的几个方面[2,3]。例如，在“爆炸”话题对应的事件框架下，“犯罪者”“作案工具”“伤亡”等就是爆炸事件框架下的事件槽，表达爆炸事件最核心的信息。因此，在建模句子时，有必要从语义上分析句子是否描述了话题关注的内容，考虑句子在特定话题下的重要性。

---

[1] * **收稿日期：**　　**定稿日期：**

基金项目：国家 973 计划课题（2014CB340504）；国家自然科学基金（61375074）；国家自然科学基金（61572049）

作者简介：应文豪（1991—），男，硕士研究生，主要研究方向为计算语言学；李素建（1975—），女，副教授，主要研究方向为自然语言处理、自动文摘和篇章分析；穗志方（1970—），女，教授，主要研究方向为计算语言学、知识工程。

2 话题敏感的抽取式多文档摘要

2.1 基本思想

话题所概述的相关事件具有事件框架结构，在框架结构内存在若干事件槽，这些事件槽描述了事件框架最核心的内容。表 1 给出了爆炸话题对应的事件框架和部分事件槽的示例。

在爆炸话题对应的事件框架下，“犯罪者”和“作案工具”事件槽提供了是谁使用哪种工具造成爆炸事件的信息，“目标、受害者”事件槽给出了爆炸事件导致的影响，这三个事件槽基本覆盖了爆炸事件的主要内容。事件的框架结构说明，某个话题下存在一些重要的事件槽，这些事件槽是整个话题的核心组成部分，因而可以选择描述这些事件槽下行为的句子概述文章中该话题的主要内容。

表 1 爆炸话题的事件框架

事件槽	说明(对象/行为)
犯罪者	人：放置、引爆、拘留
作案工具	物体：爆炸、拆除、引爆
目标、受害者	人物：摧毁、破坏、受伤

由此，本文提出一种话题敏感的抽取式多文档摘要方法，建模句子是否描述话题的重要内容，从而判断句子的话题重要性。本文不显式地抽取事件的框架结构，而是直接分析句子语义，判断句子是否描述话题的重要内容。由于诸如词频、位置等启发式特征缺乏语义信息，在建模句子时首先使用卷积神经网络学习句子的语义表示，然后通过隐狄利克雷分配模型[4]推断句子的话题，并在此基础上计算句子的话题重要性。

2.2 基于卷积神经网络的句子语义表示

近年来，通过卷积神经网络建模句子语义，在文本分类[5]、自动摘要[6]、情感分析[7]等众多自然语言处理任务上取得了良好的效果。本文在文献[8]的基础上，使用多个不同大小的卷积窗口识别短语概念，学习句子的语义表示。

对于长度为 n 的句子，$\boldsymbol{v}_j \in \mathrm{R}^d$ 是句子第 j 个词的词向量，$s_{i:i+h-1} \in \mathrm{R}^{h\times d}$ 是连续 h 个单词的词向量矩阵。窗口大小为 h 的卷积操作定义为:

$$c_i = f(\mathbf{W}\cdot s_{i:i+h-1} + b)$$

其中，c_i 是学习到的卷积特征，$\mathbf{W}\in \mathrm{R}^{h\times d}$ 是卷积核，b 是偏置项，f 表示用于抽取特征的非线性变换函数，$\cdot$ 是元素乘操作。卷积核可以视作特征识别器，发现滑动窗口内出现的某类短语概念。依次施加卷积在句子所有可能的窗口位置上，得到句子在卷积核下的概念特征向量$[c_1, c_2, ..., c_n]$，接着使用最大池化操作得到句子在当前卷积核下最显著的概念特征，其形式化描述为:

$$c_{max} = \max\{c_1, c_2, ..., c_n\}$$

为了识别尽可能多的特征，采用多个卷积核 $\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_k$ 抽取概念特征，这些概念特征构成了句子在给定窗口下的特征向量$\boldsymbol{c}^h = [c^1_{max}, c^2_{max}, ..., c^k_{max}] \in \mathrm{R}^k$。本文使用多个卷积窗口识别不同长度短语描述的概念特征，句子的语义表示$\boldsymbol{r} \in \mathrm{R}^l$通过组合这些不同卷积窗口下的特征向量得到，$\mathbf{U} \in \mathrm{R}^{l\times Km}$ 是转换矩阵，$[c^{h_1}, c^{h_2}, ..., c^{h_m}] \in \mathrm{R}^{km}$ 是向量的拼接：

$$\boldsymbol{r} = f(\mathbf{U}[c^{h_1}, c^{h_2}, ..., c^{h_m}]^{\mathrm{T}} + b)$$

2.3 基于 LDA 的句子话题推断

本文采用隐狄利克雷分配模型(Latent Dirichlet Allocation, LDA)推断句子的话题分布。LDA 是一类用于挖掘数据内隐含组结构的统计生成模型。对于文本而言，LDA 模型能够发现句子内的话题结构。LDA 模型建模话题时，使用一组预定义数目的话题组织文本内容，每个句子都有关于这些话题的概率分布。这种话题建模方式类似概率化的隐语义分析[9]。不同之处在于，LDA 在话题分布上额外增加了狄利克雷共轭先验，描述话题结构的先验知识。同样地，对于话题下的词汇分布也有类似的共轭先验存在。

LDA 模型描述了文本的生成过程。每个句子都有自己的话题分布，然后根据话题分布选择每个词位所属的话题，最后根据话题的词汇分布生成句子的单词。图 1 给出了 LDA 模型的概率图模型表示。

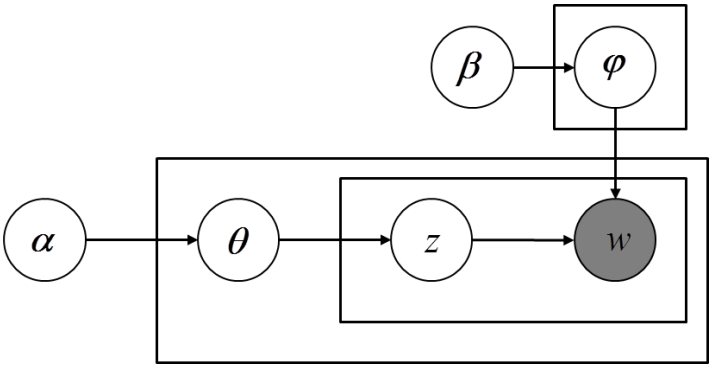

图 1 LDA 模型的概率图表示

在 LDA 的概率图表示中，α 是句子话题分布的狄利克雷共轭先验分布的参数，β 是话题下词汇分布的狄利克雷共轭先验分布的参数，θ 表示句子的话题分布，φ 是话题的词汇分布，z 和 w 分别是词位的话题标签和对应的单词。图中矩形框表示重复的图结构，嵌套结构中最外层的矩形框对应句子，内部的矩形框对应每个词汇和它的话题。最上层的矩形框是多个话题的词汇分布。

在估计 LDA 模型参数时，以文章为单位同时推断文章的话题分布、话题的词汇分布以及文章词汇的话题。而在推断句子的话题时，固定先前学习到话题的词汇分布，只需要推断句子的话题分布以及句子内词汇的话题标签。推断得到的句子话题分布 $[\mathrm{P}(t_1 \mid s), \mathrm{P}(t_2 \mid s), ..., \mathrm{P}(t_K \mid s)] \in \mathrm{R}^K$ 将用于计算句子的话题重要性。

2.4 句子话题重要性

话题的事件框架结构表明，事件槽对应的内容是话题下更加重要的内容，给出了话题的主要信息。一篇文章通过组织若干话题表达其主旨，如果找到描述话题下事件槽对应内容的句子，则可以作为摘要句用于概述原文。但是，相比语义角色标注、句法分析等句级语义分析任务，抽取事件框架结构需要对文档集内文章进行篇章级语义分析，推断相关事件以及事件之间的框架结构，是十分困难的。对于抽取式摘要任务，其目标是选择表达重要信息的句子概述原文，因而只要能够判断句子是否描述了话题重要的内容，并不需要直接抽取话题的事件框架结构来确定句子所属的事件槽。

为此，本文设计了一个神经网络 $\mathrm{NET}_{\mathrm{topic}}$ 计算句子的话题重要性。$\mathrm{NET}_{\mathrm{topic}}$ 首先通过 2.2 节定义的卷积神经网络架构学习句子的语义表示 $\boldsymbol{r}$，然后基于 LDA 推断得到的句子话题分布，计算句子话题重要性。网络对每个话题 t 学习一个用于给句子打分的权重向量 $\boldsymbol{w}_t$，判断句子内容对于当前话题是否重要。每个话题重要性得分通过计算句子语义向量 $\boldsymbol{r}$ 和权重向量 $\boldsymbol{w}_t$ 的内积得到，句子的话题重要性得分由下式计算:

$$score_{topic} = [< \boldsymbol{r}, \boldsymbol{w}_1 >, < \boldsymbol{r}, \boldsymbol{w}_2 >, ..., < \boldsymbol{r}, \boldsymbol{w}_K >]$$

另一方面，由于句子的话题是通过无监督的 LDA 模型推断得到，只有部分话题对应实际上重要的事件。为了区分这些主、次要话题，增加额外的一组参数 $\boldsymbol{u} \in \mathrm{R}^K$ 表示话题本身的重要性，使用话题权重给句子的话题重要性得分加权，得到：

$$weighted\_score_{topic} = [u_1 * score_1, u_2 * score_2, ..., u_K * score_K]$$

由于句子可以同时表达多个话题，使用 LDA 模型推断得到话题分布,规范化话题的重要性得分，得到句子最终的综合话题重要性:

$$overall\_score_{topic} = \sum_{k=1}^{K} \mathrm{P}(t_k \mid s) * weight\_score_{\mathrm{k}}$$

2.5 特征集设计与打分模型

除了句子的话题重要性特征，位置、词频等启发式特征也能够有效地用于度量句子的重要性。因此，本文设计了一组启发式特征，并结合话题重要性得分给句子打分。打分模型采用的启发式特征集详见表 2。

表 2 打分模型采用的启发式特征

特征	说明
IS_FIRST	是否出现在文章的第一句
POSITION	句子在文章出现的位置
LENGTH	句子长度
POS_RATIO	名词、动词、形容词和副词在句子的比例
ENTITY_RATIO	命名实体的比例
NUM_RATIO	数字的比例
SP_RATIO	停用词的比例
POS_POLAR_RATIO	强正向极性词的比例
NEG_POLAR_RATIO	强负向极性词的比例
AVG_DF	句子的词平均在文档集内多少文档出现
AVG_IDF	句子内词的平均逆文档频率
AVG_TFIDF	句子内词的平均 TFIDF
ALL_TFIDF	句子内词的 TFIDF 总和

为了判断句子的重要性，本文使用一个简单的两层网络 $\mathrm{NET}_{\mathrm{score}}$ 作为打分模型。这个两层网络仅包含输入层和输出层，构成一个线性函数。$\mathrm{NET}_{\mathrm{score}}$ 的输入是句子的话题重要性得分和启发式特征，句子的重要性得分由下式计算得到，其中 n 是特征数：

$$score = \sum_{i=1}^{n} w_i * feature_i$$

本文使用一个统一的神经网络组合 $\mathrm{NET}_{\mathrm{topic}}$ 和 $\mathrm{NET}_{\mathrm{score}}$，学习句子语义表示并计算话题重要性以及给句子打分。在训练时，通过反向传播算法同时调整网络各个部分的参数。

3 半监督模型训练

选择摘要句的过程本质上是排序句子，因此本文采用一种基于梯度的排序学习方法[10]训练神经网络模型。具体来说，每个训练样例包含句对$(x_i^{(1)}, x_i^{(2)})$，每个句对带有一个排序概率P_i，排序概率P_i给出了句子$x_i^{(1)}$的优先级高于句子$x_i^{(2)}$的概率。训练时模型会给每个句子打分，利用句对之间的分差s_i估计排序概率：

$$\overline{\mathrm{P}_i} = \frac{exp(s_i)}{1+exp(s_i)}$$

为了使得模型输出的句子得分符合句子之间的优先级关系，可以通过交叉熵比较预测排序概率$\overline{\mathrm{P}_i}$和真实排序概率P_i的差异程度，最小化训练集上的交叉熵，通过反向传播算法调整模型参数。

整个神经网络模型包含两个待学习的部分，分别是计算话题重要性的网络 $\mathrm{NET}_{\mathrm{topic}}$ 和线性打分网络 $\mathrm{NET}_{\mathrm{score}}$。由于现有摘要任务的参考摘要规模有限，为了能够有效地训练神经网络模型，本文提出一种利用大规模未标注新闻语料的半监督训练方法。在半监督训练框架下，神经网络的训练分成两个阶段。首先，使用少量人工标注和大量未标注的新闻语料预训练计算句子话题重要性的 $\mathrm{NET}_{\mathrm{topic}}$。然后，针对具体多文档摘要任务，使用 DUC 提供的语料生成训练数据，同时调整两个网络的参数。

3.1 预训练模型

本文使用 Cheng 等[11]自动标注的 CNN 新闻语料预训练 $\mathrm{NET}_{\mathrm{topic}}$。他们在一个规模约几百篇的人工标注新闻语料上训练分类器，判断句子是否应该包含在摘要内。然后，他们使用这个分类器自动标注了约八万篇 CNN 的新闻文章，文章中的每个句子被分为应该(标记 1)、可能(标记 2)和不应该(标记 0)包含在摘要这三类。为了构造训练数据，随机地从每篇文章标记为 1 和标记为 0 的句子中采样，分别将标记 1 和标记 0 的句子作为用于排序学习的训练句对的正、负例。本文使用这种方式生成约 600 万个句对，训练 $\mathrm{NET}_{\mathrm{topic}}$。预训练时，直接使用 $\mathrm{NET}_{\mathrm{topic}}$ 输出的话题重要性作为句子得分，计算句对的排序概率，然后最小化交叉熵调整网络参数。在预训练阶段的损失函数如下所示：

$$loss_{pre-training}(\theta) = \frac{1}{m}(\sum_{i=1}^{m} -\overline{\mathrm{P}}_i log\mathrm{P}_i - (1-\overline{\mathrm{P}}_i) log(1-\mathrm{P}_i)) + \lambda \|\theta\|^2$$

其中，m 是用于排序学习的句对数，损失函数的第一部分是排序损失，第二部分是 L2 正则化项。

3.2 精调模型参数

DUC2001-DUC2004 是经典的多文档摘要任务，每年提供文章和参考摘要作为评测数据。但是这些语料规模十分有限，每年的数据仅包含几十个文档集，约几百篇文章。本文仅使用这些语料作为精调 $\mathrm{NET}_{\mathrm{topic}}$ 和 $\mathrm{NET}_{\mathrm{score}}$ 参数的训练数据和评测系统的测试数据。在实验中，DUC2001 和 DUC2002 两年的文章用于生成训练数据，根据句子关于参考摘要的 unigram 召回率，以 0.05 为区间，不同区间的两个句子构成一个训练句对。训练时，使用上一阶段预训练的 $\mathrm{NET}_{\mathrm{topic}}$ 初始化网络参数。DUC2003 被用作开发集，采用 early stopping 策略，当开发集上的句对分类准确率持续下降即停止训练。

为了避免在精调阶段模型参数变化过大，精调阶段的损失函数额外带有一个正则化项 $\|\theta - \theta_{pre-train}\|$。其中，$\theta_{pre-train}$ 是预训练结束时的网络参数，θ 是当前模型参数。这个正则化项的目的在于避免精调时参数变化过大，丢失了 $\mathrm{NET}_{\mathrm{topic}}$ 在预训练时学习到的信息。完整

的损失函数由下式定义

$$loss_{fine-tuning}(\theta)=\frac{1}{n}(\sum_{i=1}^{n}-\bar{\mathrm{P}}_i log\mathrm{P}_i - (1-\bar{\mathrm{P}}_i)log(1-\mathrm{P}_i))+\lambda\|\theta\|+\delta\left\|\theta-\theta_{pre-train}\right\|^2$$

4 实验设计与结果分析

4.1 实验设计

本文使用 Word2Vec 在 Gigawords 语料上训练词向量，词典包含约 74 万个单词。由于整个词向量的参数规模很大，在训练神经网络模型时，固定词向量减少需要训练的模型参数避免导致过拟合。训练采用 ADAM 算法[12]，同时使用 Dropout[13]减少模型在小规模数据集上过拟合的风险，完整的模型配置见表 3。

表 3 模型超参数

特征	配置	说明
WORD_EMB_DIM	50	词向量维度
WIN_SIZES	[1,2,3]	滑动窗口大小
NUM_FILTERS	32 * 3	卷积核数
SENT_REPR_DIM	48	句子语义向量维度
NUM_TOPIC	50	LDA 话题数
LEARNING_RATE	1e-3	学习率
L2_LAMBDA	1e-3	L2 正则化系数
DELTA	1e-2	偏离预训练结果的正则化系数

在选择句子生成摘要时，采用类似 MMR[14]的贪心句子选择方法，在上一步打分结果的基础上对句子重排序，其形式化描述如下式:

$$s_{重排序得分}=s_{原始得分}-\lambda_{mmr}*max_{s_i\in S}\{sim(s,s_i)\}$$

其中，S 是当前已生成的摘要，s 和 s_i 分别是待选择的句子和摘要句。sim 是句子相似度函数，实验中采用句子之间的词共现率。λ_{mmr} 是相似度罚项系数，本文直接选择 $\lambda_{mmr}=0.5$。

4.2 结果分析

实验中比较了多个常用的基准多文档摘要模型(FreqSum, TsSum, Centroid, LexRank, Greedy-KL)[1,15,16,17]以及 DUC2004 任务上当年参赛的最好结果(Peer 65)[18]。由于不同方法在评估时采用了不同的数据集或者使用了不同的评测指标，为了能够合理地与这些方法进行比较，本文采用 Hong 等[19]在 DUC2004 数据集上的评测结果，他们复现或直接引用这些基准模型的结果，使用统一的 ROUGE[20]参数评测各方法。

为了验证本文提出的句子话题重要性特征的作用，实验中采用多组不同的特征集训练模型。模型一仅使用全部启发式特征直接在 DUC2003 和 DUC2004 上训练，得到基于启发式特征的基本摘要模型 BasicSum。模型二结合话题重要性特征和启发式特征,然后直接在 DUC2003 和 DUC2004 上训练，得到话题敏感的摘要模型 TSSM。模型三直接使用 LDA 推断得到的话题分布计算句子的话题重要性，然后结合启发式特征在两年 DUC 数据集上训练，得到 TSSM(Topic_Dist)。模型四在半监督框架下使用全部特征训练，得到半监督话题敏感的摘要模型 Semi-TSSM。本文直接引用 Hong 等论文中的 ROUGE-1 和 ROUGE-2 召回率评测结果，使用相同的 ROUGE 配置评测 HSM、TSSM、TSSM(Topic_Dist)和 semi-TSSM，具体实验结果见表

4。

Peer65 是 DUC2004 正式参赛队伍中的最好结果，采用隐马尔科夫模型(HMM)建模句子之间的关系选择摘要句。尽管没有使用复杂的句子选择方法，BasicSum 已经在 ROUGE-2 召回率上超过了 Peer65。在实验中除了常用的词频特征，我们还考虑了命名实体、数字等词和短语的类型，实验结果验证了这类特征的有效性。

可以看到，半监督话题敏感的 Semi-TSSM 模型取得了最高的 ROUGE-2 召回率，同时在 ROUGE-1 召回率也有接近最好的结果。模型采用的启发式特征主要通过统计词的类型、频率等作为句子内容重要性的指示。而本文提出的话题重要性特征则是从另一个角度，考虑句子与所属话题的语义关系，判断句子是否描述了话题的重要内容。对比仅使用启发式特征的 BasicSum，增加话题重要性特征后明显地提高了摘要系统的性能，说明话题重要性特征能够作为传统启发式特征的有效补充。

表 4 DUC2004 多文档摘要结果

系统	ROUGE-1-R	ROUGE-2-R
LexRank	35.95	7.47
Centroid	36.41	7.97
FreqSum	35.30	8.11
TsSum	35.88	8.15
Greedy-KL	**37.98**	8.53
Peer 65	37.62	8.96
BasicSum	36.64(+0.0)	9.05(+0.0)
TSSM	36.85(+0.21)	8.93(-0.12)
TSSM(Topic_Dist)	36.98(+0.34)	9.06(+0.01)
Semi-TSSM	37.75(+1.11)	**9.24(+0.19)**

另一方面，直接在 DUC2001 和 DUC2002 上训练话题敏感的摘要模型 TSSM 效果不佳，仅在 ROUGE-1 召回率上有略微的提升，而 ROUGE-2 召回率有所下降。可能的原因是句子内容与话题的语义关系比较复杂，在小规模数据集上无法学习到通用的模式，因而容易导致模型训练时出现过拟合。

注意到直接使用 LDA 的话题分布计算句子话题重要性的 TSSM(Topic_Dist)同样效果并不理想。相比仅使用启发式特征的 BasicSum，在 ROUGE-1 和 ROUGE-2 的召回率上仅有略微提升。这是因为话题本身是十分抽象的概念，不考虑句子的内容而直接使用话题分布估计话题重要性，无法体现同一个话题下不同内容的重要程度。因此，在建模句子的话题重要性时，需要同时考虑句子所属的话题以及句子的内容，判断句子是否描述了重要话题下的重要内容。

为了直观地说明话题重要性特征的作用，表 5 给出了 Semi-TSSM 的线性打分模型权重最高的前 6 个特征。从表 5 可以看到，话题重要性对句子打分的结果影响很大，其在打分模型的权重位于所有特征中的第三位。注意到词汇在文档集内的 TFIDF 是一类十分有效的特征，事实上具有高 TFIDF 的词汇可以视作文档集的主题词。这些词汇在文档集内反复出现，而又较少出现在其他文章中，因而这些词汇通常描述了文档集主题的相关信息。另外，位置特征在识别摘要句时也起到了重要的作用，这是因为从写作习惯上作者往往会在文章的首段中说明主旨。

表 5 线性打分模型权重最高的前 6 个特征

特征	权重	说明
ALL_D_TFIDF	2.222	句子词汇在文档集下的 TFIDF 总和
ENTITY_RATIO	1.252	命名实体比例
TOPIC_IMPORTANCE	0.919	话题重要性
IS_FIRST	0.899	是否出现在文章的首句
NUM_RATIO	0.753	数字的比例
POSITION	0.659	句子在文章出现的位置

本文提出的话题重要性特征可以视为一种度量句子的话题先验重要性的方法，这里的先验指的是仅根据单独的句子判断是否描述了话题的重要内容，事先不需要知道句子所在的篇章信息。另一方面，生成多文档摘要时句子与文档集的话题一致性也值得关注，因为适合作为摘要的句子同时应描述了文档集的主要话题。为此，我们使用 LDA 同时建模句子和文档集的话题概率分布，并通过 KL 散度度量两者的差异。表 6 给出了增加话题一致性的摘要模型 Semi-TSSM 在 DUC2004 数据上的实验结果。

表 6 句子与文档集的话题一致性对摘要结果的影响

系统	ROUGE-1-R	ROUGE-2-R
BasicSum	36.64	9.05
Semi-TSSM	**37.75**	9.24
Semi-TSSM (with topic consistency)	37.6	**9.61**

尽管相比只使用话题先验重要性特征，模型在 ROUGE-1 召回率上略有下降。但是，考虑句子与文档集之间的话题一致性后明显提高了 ROUGE-2 召回率。这说明生成摘要时从话题层面需要同时考虑两方面的内容。一方面，句子是否描述了某些话题的重要信息。另一方面，句子的话题是否与文档集的话题一致也值得关注。

在句子选择过程中，本文采用贪心的句子选择算法，为了分析不同 λ_{mmr} 对摘要结果的影响，分别设置 λ_{mmr} 从 0.1 到 1.0 抽取了 DUC2004 数据集的多文档摘要，图 2 给出不同 λ_{mmr} 设置下的 ROUGE-2 召回率变化曲线。

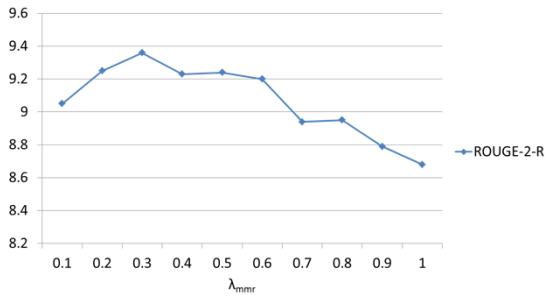

图 2 不同 λ_{mmr} 的 ROUGE-2 召回率变化

从图 2 中发现，不同 λ_{mmr} 的设置对最终摘要结果的影响很大，过高或者过低的 λ_{mmr} 都将降低摘要质量。这是因为过低的 λ_{mmr} 会在摘要内引入更多的冗余内容，而过高的 λ_{mmr} 可能会排除包含重要信息但是又与当前摘要部分内容重合的长句。

五 结论

本文提出了一种计算句子话题重要性的方法，区别于传统启发式特征，该方法从另一个角度分析句子与话题之间的语义关系，判断句子是否描述话题的重要内容。实验结果表明话题重要性特征能够作为传统启发式特征的有效补充，提升文档摘要质量。在自动摘要任务，模型训练面临缺乏参考摘要的问题，结合排序学习和大规模未标注语料的半监督训练框架是一种可行的解决方法。生成摘要时，从话题层面需要同时考虑句子的话题重要性以及句子与文档集的话题一致性。目前，我们仅直接使用线性模型结合这两类信息。实际上，文章与句子的话题是比较复杂的。话题存在层次结构，同时相互之间也存在依赖关系。因此，如何有效地结合这些信息值得进一步的研究。此外，我们发现在贪心的句子选择算法中相似度惩罚系数λ对最终摘要结果影响很大，因此在后续的工作中有必要进一步研究有效的全局优化方法，改进贪心的句子选择过程。

参考文献

[1] Nenkova A, Vanderwende L, McKeown K. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 573-580.

[2] Chambers N, Jurafsky D. Template-based information extraction without the templates[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 976-986.

[3] Chambers N. Event Schema Induction with a Probabilistic Entity-Driven Model[C]//EMNLP. 2013, 13: 1797-1807.

[4] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[5] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[6] Cao Z, Wei F, Li S, et al. Learning Summary Prior Representation for Extractive Summarization[C]//ACL (2). 2015: 829-833.

[7] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188, 2014.

[8] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(Aug): 2493-2537.

[9] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 50-57.

[10] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent[C]//Proceedings of the 22nd international conference on Machine learning. ACM, 2005: 89-96.

[11] Cheng J, Lapata M. Neural summarization by extracting sentences and words[J]. arXiv preprint arXiv:1603.07252, 2016.

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

[13] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural

networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[14] Carbonell J, Goldstein J. 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. ACM, 1998: 335-336.
[15] Lin C Y, Hovy E. The automated acquisition of topic signatures for text summarization[C]//Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2000: 495-501.
[16] Radev D R, Allison T, Blair-Goldensohn S, et al. MEAD-A Platform for Multidocument Multilingual Text Summarization[C]//LREC. 2004.
[17] Erkan G, Radev D R. Lexrank: Graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.
[18] Conroy J M, Schlesinger J D, Goldstein J, et al. Left-brain/right-brain multi-document summarization[C]//Proceedings of the Document Understanding Conference (DUC 2004). 2004.
[19] Hong K, Conroy J M, Favre B, et al. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization[C]//LREC. 2014: 1608-1616.
[20] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out: Proceedings of the ACL-04 workshop. 2004, 8.

应文豪(1991-)，男，硕士研究生，北京大学信息科学与技术学院，主要研究方向为计算语言学，E-mail: yingwh123@sina.com

李素建（1975-），女，副教授，北京大学信息科学与技术学院，主要研究方向为自然语言处理、自动文摘和篇章分析，E-mail: lisujian@pku.edu.cn

穗志方（1970-），女，教授，北京大学信息科学与技术学院，主要研究方向为计算语言学与知识工程，E-mail: szf@pku.edu.cn