

抛开模型，探究文本自动摘要的本质——ACL2019 论文佳作研读系列



邱震宇

华泰证券股份有限公司 算法工程师


113 人赞同了该文章

文本自动摘要任务作为NLP领域中一个富有挑战性的任务，同时也是很多研究团队关注的问题。而在工业界，文本摘要任务的应用也非常广泛，除了直接提供文本摘要结果供用户阅读外，在很多其他下游任务中都充当着重要角色。例如长文本情感分析、搜索引擎、推荐系统等，相较于直接使用原文，使用好的摘要能够在提升性能的同时又不会损失太多信息。

本人对文本自动摘要任务有一些项目经验。在浏览本年度ACL论文列表时，有一篇对文本摘要的本质进行分析论证的论文进入了我的视线，也就是我接下来要来分享的论文：

A Simple Theroretical Model of Importance for Summarization

www.aclweb.org



这篇论文我最喜欢的一点就是它并不涉及复杂模型，也不涉及庞大的语料和训练成本。它以信息学最基本的理论为基础，来探究什么样的摘要才是一个好摘要。论文中使用到的数学理论很直观，没有很复杂的推导，但是读完后，有种“原来如此”的感觉。下面就切入正题，结合我做摘要的一些体会，来分享这篇佳作。

论文理论的依据和前提

本篇论文采用的是信息论中的基本理论。信息论简单说就是一种描述万物信息量的理论。而摘要的主要目的就是在损失最小信息量的情况下，最大限度表达原文信息量。因为基于信息论来研究摘要任务是合适的。然而信息论着重于研究信息的不确定性，容易忽略人类语言中的语义信息，因此直接应用信息论也是不适合的。因此，论文里将文本切分成最基本的语义单元，语义单元负责语义部分，而信息论只需要关注由语义单元构成的文本信息即可。这个语义单元可以是字符，也可以是词，也可以是n-gram词，或者具有更复杂语义语法内容的单元。文本则是用这些基本语义单元的概率分布来表示。这个其实也是与language model的定义相类似。目前NLP领域中的词向量，language model都是基于的语义单元理论进行研究。

论文中，引入的术语符号如下：

Ω ：表示语义单元集合

ω_i ：表示一个语义单元

X ：表示从语义单元集合中抽语义单元组合成的文本

P_X ：表示每个文本基于语义单元组合的概率分布

P_D ：表示原文的基于语义单元组合的概率分布，D表示原文档

P_S ：表示候选摘要的基于语言单元组合的概率分布，S表示候选摘要

为了后续描述方便，会将 P_D ， P_S 等概率分布简写为D,S。

如何判断一个摘要是好的？

论文从四个不同的角度，在本质上对摘要本身做了分析。分别是冗余度(redundancy)，相关性

(relevance), informativeness, 重要性(importance)。其中，重要性这个概念是论文新突出的理念，它结合了其余三个概念的内容，并进行了公式化。下面分别对这个四个概念做论述：

相关性

首先说相关性，目前大部分模型对摘要抽取或生成的目标都可近似为相关性。对于有监督学习训练来说，抽取式摘要的训练数据标注了哪些句子是摘要句。最后任务转化为对每个句子做二分类问题，而生成式摘要的seq2seq模型中，也是与标注的人工摘要进行语义单元上的差异计算。对于无监督学习来说，大部分的方法的建模目标都是相关性。下面按照大致的类别举几个无监督学习的例子：

- 基于关键词的方法。先使用关键词抽取模型抽取关键词，然后统计包含关键词最多的句子作为候选摘要。关键词抽取应用比较广泛的就是基于tfidf方法。
- 基于主题模型，如LDA，LSA等，分析文档隐含的主题，然后分析句子和主题的相关性。
- 将句子向量化表示，然后对句子进行聚类，隐含的每个聚类代表某个主题，然后从这些主题中挑选摘要句。
- graph-based方法，以textrank为经典方法，将句子作为节点，句子之间的相似度关系作为边，构建有权图，利用图论中的算法，得到每个句子的权重分数。这种方法，相比较于前面三个不太直观，实质上，它挑选的句子通常是相似性最强的一堆句子中的一个。即textrank认为一个句子如果与它相似的句子数越多，表明这个句子与文档主题内容越相关。

上面几种方法，虽然表面上看建模的目标都不一样，但是实际上都是在朝着相关性的方向走。他们认为文档中覆盖面最大的主题是最能概括原文内容的，因此最终的目标即是找这些主题对应的句子。

这篇论文则是从一个新的角度来建模相关性：交叉熵。

论文认为，通过阅读摘要，应该降低对原文的不确定感，摘要文本应当以最小的信息损失来推断原文文档。而上一小节提到过，摘要和文档都可以当作是一个概率分布，那么做深度学习的同学们应该都比较了解，可以通过交叉熵来建立损失函数，让模型去拟合一个真实的概率分布。这里摘要代表我们的模型，文档代表真实的概率分布。公式如下：

$$REL(S, D) = -CE(S, D)$$

这里CE指的是交叉熵的函数。注意到这里有个负号。因为交叉熵越小，表示摘要和文档的差异越小，那么相关性应当越强。

冗余度

实际生产环境中，使用以相似度为目标建模的方法配以一些人工规则通常能够符合需求。然而，当处理较长的文本时，只以相关性为目标做摘要会遇到一个很明显的问题：模型会倾向于生成一堆相似度很高的摘要句，而丢失了一些小众主题的信息。这些信息虽然在文档中的比例不高，但是不代表它不重要。尤其是金融领域中的研报或者财报等类型的文本，有些重要信息信息并不会在文本中出现很多次。因此，需要考虑提取尽可能多样化的摘要内容。

历史上，也有一些学者对摘要的冗余度进行研究的例子。下面分享一些我之前在工作中调研实践过的方法。

- MMR，全称Maximal Marginal Relevance。将相关性和冗余度放在一个目标函数中，使用贪心方法优化目标函数。每次挑选摘要时，除了建模其相关性分数外，还要扣除其与当前已有的摘要集合的冗余度分数，最后挑选综合分数最高的那个加入到摘要集合中。这个方法简单高效，且是无监督的方法，在生产环境中，如果缺少高质量的标注数据，可以使用这个方法。
- 利用submodular函数原理来建模冗余度。关于submodular函数的原理，这里就不展开了，有兴趣的同学可以自行谷歌。它在建模冗余度时，不是惩罚冗余性，而是奖励多样性。当挑选一个与摘要集合不太相似的句子时会奖励一个较高分数，但是后续如果有与之类似的句子被挑选了，则这个奖励会呈非线性递减，借鉴了经济学中的边际效益思想。之所以奖励多样性，是为了与相关性分数的变化关联一致，维持submodular函数的单调性质。

论文中，对冗余度的建模方法简单到不可思议。它认为，摘要包含的信息量表明了其本身的多样性程度。因此直接使用信息论中的熵来度量其信息量，即冗余程度。简化后的公式如下：

$$Red(S) = -H(S)$$

注意到等式右边有一个负号，表示熵越大，文本不确定性越高，信息量也越大，那么其冗余度也越小。

informativeness

对于这个单词，我之所以没给中文翻译是因为我还不知道如何确切翻译它，如果翻译成信息量，我觉得还不是很够味。根据论文的叙述，这个概念假设当前有一个背景知识库K，此时需要对文档D进行摘要抽取，那么候选摘要S对于K来说，应当新增尽可能多的信息，才能让读者在阅读摘要后获取最多的新信息。如果摘要句子说的都是用户早就知道的事情，那么阅读摘要没有给用户产生任何价值。

相关性和冗余度只是在当前处理文档的范围内进行建模，但是人类的语言是有庞大的常识库的。只使用相关性和冗余度有其局限性，因此才引入了informativeness的概念。那么如何度量这个概念呢？论文给出的方法也很简单，informativeness的目标是让S尽可能与K不同，同时K也是由语义单元组成的文本语料集合，因此也可以用 P_K 来表示K的概率分布，那么剩下的工作就很明显了，与相关性类似，使用交叉熵来衡量两个概率分布的差异性。简化后的公式如下：

$$Inf(S, K) = CE(S, K)$$

注意等式右边没有负号，表示S和K的差异越大，我从摘要中获取的新的信息越多，则informativeness就应当越大。

其实，informativeness的内涵与tfidf的思想比较像。tfidf找的是词频较大，但是又不是每篇文章都包含的词。最典型的实例就是会过滤掉停止词，这些词的信息量是很少的。

有的同学会问了，这个K从哪里来呢？有条件的，可以通过构建大规模的知识库语料，如训练BERT使用的多领域的文本语料；没有条件的，可以直接将需要提取摘要的所有文档集作为K。

重要性

importance是论文提出的一个新的概念。首先要明确一点，它针对的是语义单元，目标是计算每个语义单元的重要性分数，在构造摘要时，根据每个语义单元的评分来丢弃不需要的语义单元。

论文中，将importance作为相关性和informativeness的整合产物。既要摘要保留足够多的D的信息，又要尽量产生与K不同的新的信息。因此，其给出了一个建模函数的模板 $f(P_D(\omega_i), P_K(\omega_i))$ 表示一个语义单元的重要性分数，它必须满足如下条件：

- 在文档中的不同语义单元 $\omega_i, \omega_j, i \neq j$ ，如果两个单元的在D中的概率分布相同。在K中的概率分布不同，且 $P_K(\omega_i) > P_K(\omega_j)$ ，则应有 $f(P_D(\omega_i), P_K(\omega_i)) < f(P_D(\omega_j), P_K(\omega_j))$ 。在保持相关性一致时，两个语义单元的重要性分数应当与informativeness呈负相关关系。
- 在文档中的不同语义单元 $\omega_i, \omega_j, i \neq j$ ，如果两个单元的在K中的概率分布相同。在D中的概率分布不同，且 $P_D(\omega_i) > P_D(\omega_j)$ ，则应有 $f(P_D(\omega_i), P_K(\omega_i)) > f(P_D(\omega_j), P_K(\omega_j))$ 。在保持informativeness一致时，两个语义单元的重要性分数应当与相关性呈正相关关系。
- 这个度量函数应当保持信息论中度量方法的加性性质，因此需要满足：
 $I(f(P_D(\omega_i), P_K(\omega_i))) \equiv \alpha I(P_D(\omega_i)) + \beta I(P_K(\omega_i))$ 。I表示香农理论中的信息度量方法。
- 保证f得到的结果是一个有效的概率分布，即 $\sum_i f = 1$ 。这里，我个人认为之所以要保证其是一个有效的概率分布，是因为要将所有语义单元的importance分数都放在同一个维度去对比。

为了同时满足上面的四个条件，论文给出一个函数实例，如下：

$$\mathbb{P}_{\frac{D}{K}}(\omega_i) = \frac{1}{C} \cdot \frac{d_i^\alpha}{k_i^\beta}$$

$$C = \sum_i \frac{d_i^\alpha}{k_i^\beta}, \alpha, \beta \in \mathbb{R}^+$$

知乎 @邱震宇

其中， $d_i^\alpha = P_D^\alpha(\omega_i)$ ， $k_i^\beta = P_K^\beta(\omega_i)$ 。而 α, β 则是控制相关性和informativeness在计算importance时的权重。根据实际情况，可调整两个权重，看重哪个就调高哪个。在真实的模型设计和训练时，可以将其作为超参数，使用验证集来调参。

论文最后还附上了上述公式满足四个基本条件的证明，证明很简单，只要将 $I = \log_2$ 函数代入推导就可以，另外关注一点，就是将 $-\beta$ 替代了 β 。因为减法在广义上也是一种加法。

如何整合四个维度？

通过上面几节的论述，得到了四个不同的维度，来勾勒出一个好的摘要的轮廓。但是，最终我们是要为每个摘要句进行评分，因此需要将所有摘要建模维度进行统一，因此需要设计一个方法将上述四个维度进行整合。

整合相关性和冗余度

我们先关注一下相关性和冗余度两个维度。之后我们会由此引申到整合四个维度。

还记得上章节得到相关性和冗余度的两个公式吗？

$$REL(S, D) = -CE(S, D)$$

$$Red(S) = -H(S)$$

将上述两个公式组合一下，就能得到另外一个衡量两个概率分布差异的方法：**KL散度**。

$$KL(S||D) = CE(S, D) - H(S)$$

$$-KL(S||D) = Rel(S, D) - Red(S)$$

其实，深度学习中经常使用的交叉熵损失与KL散度的关系非常密切，因为深度学习中的训练数据要求是独立同分布的，因此对于深度学习的数据而言，H(X)是常量，每条训练数据的信息量是确定的，因此做优化时通常会不考虑H(x)，只看CE。而这里情况就不同，我们需要关注摘要的信息量。

KL散度很好的整合了相关性和冗余度。当S与D的KL散度很小的时候，说明摘要拟合D的效果非常好，此时相关性和冗余度的综合分数就比较高，而对应的，Rel就需要取较大的值，而Red需要取较小的值。这个与之前的论述是一致的。

整合所有维度

首先，对于importance来说，它整合了相关性和informativeness两个维度，它得到的理想摘要应当尽可能得拟合概率分布 $P_{\frac{D}{K}}$ 。这里仍然使用交叉熵来度量分布的差异。因此有：

$$Importance(S, \frac{D}{K}) = -CE(S, \frac{D}{K})$$

注意等式右边的负号，表示两个分布越接近，交叉熵越小，对应的摘要句的importance分数越高。

然后，将冗余度添加进来，添加方法参考上一小节整合相关性和冗余度的方式，使用KL散度来表示最终的度量函数：

$$\theta_I(S, D, K) = -CE(S, \frac{D}{K}) + H(S) = Importance(S, \frac{D}{K}) - H(S) = -KL(S||\frac{D}{K})$$

这里， $S, \frac{D}{K}$ 都对应各自的概率分布 $P_S, P_{\frac{D}{K}}$ 。

最终得到的摘要句要最大化我们的 θ_I ，也就是要最小化我们的 $KL(S||\frac{D}{K})$ 。

另外，论文额外备注了一个容易忽视的问题，就是作为求解目标的概率分布 $P_{\frac{D}{K}}$ ，它本身做一个概率值，它的量纲变化范围比较狭窄。例如，会出现很多 $P_{\frac{D}{K}}$ 值很接近的句子，此时就不太好选择摘要句。因此可以使用通过对概率分布取熵的方法，扩大其量纲，而 \log_2 函数能够很好地完成这个任务。

Potential Information

论文在理论部分的最后还提到了一个概念叫**potential information**。这个概念的目标主要是为建模informativeness目标提供上界。具体来说，potential information 建模的是D与K的分布差异。简单来说就是在背景知识K的条件下，我们能从D获取的新的信息量。其同样使用交叉熵来建模两个分布的差异，公式如下：

$$PI_K(D) = CE(D, K)$$

注意公式等式的右边没有负号，表示文档D和背景知识K的分布差异越大，potential information 就应当越高。

对于摘要S来说，它本身的目标就是以D为信息源的条件下，能够获取尽可能多的不同于K的信息，其上限值就是这个 $PI_K(D)$ 。

个人认为这个概念的提出，有助于设计模型的性能上界，方便指导模型训练时的终止时机，调参的性价比等等。当前，所有这些工作的前提是K和D需要有差异性，需要设置大量的基础语料作为背景知识库K。

实验部分

至此，论文的理论部分已经分享结束。后面还有作者验证理论所设计的简单实验，这里就不多做介绍了。大致介绍一下，就是论文并没有使用一些传统的rouge、bleu等评测方法做摘要评测，而是使用了一种金字塔评分方式，由机器和人工各自生成摘要集合，然后分别用金字塔方法计算评测分数，最后计算机器生成摘要和人工摘要的相关系数，以评测机器生成的摘要是否达到了人工的水平。另外，作者也做了机器生成摘要和标准摘要的关于 θ_I 分数的t-test。

小结

读完这篇论文，我大概能明白它能够得奖的原因。因为它使用的理论都很简单，没有复杂的推导，相反，它的所有理论的提出都很符合直觉，但是又有种恍然大悟的感觉。我个人觉得它的价值比提出一个复杂模型，在榜上刷上了几个点要高。当然，我并不反对研究模型刷榜，但是我更希望能多一点这样的探究某个NLP领域任务本质的研究出现。

最后，我认为本文的价值在于为后续研究人员设计自动摘要模型，提出了理论上的指导。后续的模式设计时，可以借鉴这篇论文的思想设计合适的目标函数。另外，可以结合当前比较火热的大规模预训练语言模型，好好设计论文中提到的K，也许会有更好的思想被提出来。

发布于 2019-08-04

文章被以下专栏收录



我的ai之路

进入专栏




PaperWeekly

PaperWeekly是一个推荐、解读、讨论和报道人工智能前沿论文成果的学术平台， ...

进入专栏

写下你的评论...


😊

 **kappa** 6 个月前

博主想問個文中提到的P_K,P_D,P_S具體而言該如何求得啊？

計算出現的單詞在總資料中的機率相乘嗎？

👍 赞

 **邱震宇 (作者)** 回复 **kappa** 6 个月前


我认为它是指某种将文本概率分布化的一种思想，你说的计算方法也是可以的，其实很多language model也是通过计算词的联合概率来得到某段文本的概率

👍 2

 **斯多歌** 6 个月前


文本摘要其实还有一个维度是连贯性，不知道paper里有没有考虑

👍 赞

 **邱震宇 (作者)** 回复 **斯多歌** 6 个月前

我理解连贯性应该和通顺度差不多意思吧，那这篇文章是没有考虑的，这个维度确实是很量化的，感觉可以单独出篇论文研究了


👍 赞

 **斯多歌** 回复 **邱震宇 (作者)** 6 个月前

是有量化的方法的，好像是用二部图做的，之前有篇论文


👍 赞

[查看全部 13 条回复](#)

 **知乎用户** 3 个月前

问一下现在文本摘要的比较好的实践是用什么方案

👍 赞

 **邱震宇 (作者)** 回复 **知乎用户** 3 个月前

具体场景是什么呢，如果是抽取式，现在一般都是结合bert，如果是生成式，就是一个nlg问题，一般也有结合gpt2或者用强化学习思想直接优化rouge或者bleu指标的，不过这样的摘要可能不会很好

👍 2

 **知乎用户** 回复 **邱震宇 (作者)** 3 个月前

抽取式结合bert做分类吗

👍 赞

[展开其他 2 条回复](#)

 **篠原悠夜** 2 个月前

感谢您！非常好的论文解读！

👍 赞