链滴

玻森新闻自动摘要算法简介

作者：BosonNLP

原文链接：https://ld246.com/article/1452149221136

来源网站：链滴

自动摘要（或摘要技术Automatic Summarization），顾名思义，是指从单篇或者多篇文章中，摘要点来概括文章大意的技术。它在机器学习和数据挖掘中有着重要的地位。

在这篇文章中，将要详细谈一谈自动摘要算法实现，以及玻森进行的优化。（对于自动摘要概念有兴想深入了解的读者可以自行搜索，该篇文章中在这方面不作详细介绍。）

自动摘要可以分为笼统式和查询式。由于查询式摘要的应用场景较为特定，笼统式摘要成为现阶段主，它也能在很大程度上满足人们对摘要的需求。因此玻森的新闻摘要是笼统式抽取摘要。

玻森采用的是最大边缘相关模型（Maximal Marginal Relevance）的一个变种。MMR是无监督学习型，它的提出是为了提高信息检索（Information Retrieval）系统的表现。例如搜索引擎就是目前大最常用的信息检索系统。大家可能经常会碰到，对于我们输入的一个关键词，搜索引擎通常会给出重的或者内容太接近的检索的情况。为了避免这个现象，搜索引擎可以通过MMR来增加内容的多样性给出多方面考虑的检索结果，以此来提高表现。

这样的思想是可以被借鉴用来做摘要的，因为它是符合摘要的基本要求的，即权衡相关性和多样性。难理解，摘要结果与原文的相关性越高，它就接近全文中心意思。而考虑多样性则使得摘要内容更加全面。非常的直观和简单是该模型的一个优点。

相比于其他无监督学习方法，如TextRank（TR），PageRank（PR）等，MMR是考虑了信息的多性来避免重复结果。TR，PR是基于图（Graph）的学习方法，每个句子看成点，每两个点之间都有条带权重（Weighted）的无向边。边的权重隐式定义了不同句子间的游走概率。这些方法把做摘要问题看成随机游走来找出稳态分布（Stable Distribution）下的高概率（重要）的句子集，但缺点之便是无法避免选出来的句子相互之间的相似度极高的现象。

而MMR方法可以较好地解决句子选择多样性的问题。具体地说，在MMR模型中，同时将相关性和多性进行衡量。因此，可以方便的调节相关性和多样性的权重来满足偏向“需要相似的内容”或者偏向需要不同方面的内容”的要求。对于相关性和多样性的具体评估，玻森是通过定义句子之间的语义相度实现。句子相似度越高，则相关性越高而多样性越低。

自动摘要的核心便是要从原文句子中选一个句子集合，使得该集合在相关性与多样性的评测标准下，分最高。数学表达式如下。

$$\mathrm{MMR} := \arg \max_{D_i \in R \setminus S} [\lambda(\mathrm{Sim}_1(D_i, Q)) - (1-\lambda)(\max_{D_j \in S} \mathrm{Sim}_2(D_i, D_j))]$$

需要注意的是，D，Q，R，S都为句子集，其中，D表示当前文章，Q表示当前中心意思，R表示当前摘要，S表示当前摘要。

可以看出，在给定句子相似度的情况下，上述MMR的求解为一个标准的最优化问题。但是，上述无督学习的MMR所得摘要准确性较低，因为全文的结构信息难以被建模，如段落首句应当有更高的权等。为了提高新闻自动摘要的表现，玻森在模型中加入了全文结构特

征，将MMR改为有监督学习方法。从而模型便可以通过训练从“标准摘要”中学习特征以提高准确。

玻森采用摘要公认的Bi-gram ROUGE F1方法来判断自动生成的摘要和“标准摘要”的接近程度。经训练，玻森在训练数集上的表现相对于未学习的摘要结果有了明显的提升——训练后的摘要系统F1提了30%。值得一提的是，在特征训练中，为了改善摘要结果的可读性，玻森加指代关系特征，使得模表现提高了8%。

摘要引擎的具体调用API可以参见文档