文本摘要中的句子抽取方法研究*

张龙凯，王厚峰
北京大学 计算语言学教育部重点实验室，北京 100871
E-mail: zlk0825@gmail.com

摘　要：抽取式摘要是从正文中按照一定策略抽取重要句子组成摘要。本文提出了一种句子抽取方法。基本思想是将句子的抽取看作序列标注问题，采用条件随机场模型对句子进行二类标注，根据标注结果抽出句子以生成摘要。由于不在摘要中的句子的数量远大于摘要中的句子数，标注过程倾向于拒绝将句子标注为摘要句，针对此问题本文引入了修正因子进行修正。实验表明该方法具有较好的效果。

关键词：文本摘要；句子抽取；条件随机场

Research on Sentence Extraction Methods in Text Summarization

Zhang Longkai, Wang Houfeng
Key Laboratory of Computational Linguistics (Peking University) Ministry of Education, Beijing 100871
E-mail: zlk0825@gmail.com

Abstract: Extractive summarization attempts to extract important sentences from the original text and re-organize them in a summary. In this paper we propose a method to automatically identify significant sentences. The basic idea of this method is to label each sentence with one of two tags and a sequence labeling model can be employed here. We use Conditional Random Fields model. Considering that many sentences are tend to be rejected due to the fact that sentences in summarization is much less than the opposite, we introduce a correction factor to smooth.The method. Experiment results show that our method gets a good performance.

Keywords: text summarization; sentence extraction; CRF

1 概述

随着电子文本数量的剧增，快速获取文本信息的需求越来越强烈。作为浓缩文本信息的技术，自动摘要可以扮演重要的角色。自动摘要的宗旨是为用户提供简短的文本表示。在保留尽可能多的原文信息的同时，形成尽可能短的摘要。对于一个理想的抽取式摘要而言，具有三个基本特征：源自文本、保留重要信息、长度短[1]。

按照摘要源自的文本个数，可分为单文本摘要和多文本摘要。按照摘要的方式，又分成生成式摘要和抽取式摘要。本文研究单文本、抽取式摘要问题。在抽取式摘要中，从文本中选取代表性句子是难点所在。IBM 的 Luhn 在 1958 年提出一种基于高频词的方法，将高频词列出并给包含这些高频词的句子打分，得分高的句子被认为是摘要句[2]。Baxendale 则引入句子位置作为判断句子重要性的一种特征，该特征被后来大部分机器学习算法所借鉴[3]。Edmundson 整合了 Luhn 和 Baxendale 的方法，并在科技文献中取得了较好的应用效果[4]。Kupiee et al.在 1995 年提出一种基于朴素贝叶斯算法的方法，在 Edmundson 的基础上增加了句长等特征[5]。同样使用朴素贝叶斯算法的还有 1999 年 Aone et al.，其中考虑了 TF-IDF 等多个特征[6]。同时期的还有 1999 年 Lin 的方法。不同于朴素贝叶斯算法的独立性假设，Lin 采用决策树算法并取得了较好的效果[7]。2001 年 Conroy 和 O'leary 提出一种基于隐马尔可夫模型的方法，由于隐马尔可夫模型有较强的独立性假设，该方法仍存在不足[8]。Osborne 于 2002 年提出一种基于最大熵模型的方法，结果表明，通过增加先验概

* 基金资助：国家自然科学基金（资助号：91024009，60973053）

率，该方法优于基于朴素贝叶斯模型的方法。文献[9]提出一种基于条件随机场模型的方法选择句子，并在英文测试语料上有着较好的效果。

在基于机器学习的文本摘要中，对代表性句子的选择大多将句子作为分类问题。本文考虑了句子之间的依赖关系，将摘要句的提取过程看作一个序列标注问题。基本思想是，将文本看作是句子的序列，如果某个句子出现在摘要中，则标为“在”，否则，标为“不在”。利用“带标”的文本集合，可以训练一个序列标注模型。由于条件随机场（CRF）属于全局优化的序列标注模型，本文采用CRF模型标识句子。一般而言，摘要远远短于原始文本，因此，原文本中的大多数句子都将被排除在摘要之外。这样，训练的模型会倾向于将句子标为非摘要句。本文引入修正因子来平滑这一现象。

接下来的几部分详细介绍了本文所采用的思路，主要包括特征的选择和模型的训练。最后，同已有方法作了对比，测试表明，本文所述方法有着较好的效果。

2 特征选择

一个句子是否被抽取作为摘要句，受多个因素的影响。总体上可以分成两类，其一是句子自身，其二是上下文信息：这里，我们称为单句特征和关联特征。

2.1 单句特征

单句特征是指句子自身体现出来的特征，不涉及上下文因素。本文使用的单句特征包括以下几种：

句子长度：过长或过短的句子通常较少地出现在文本摘要中，本文在计算句子长度时首先过滤掉停用此，然后以词为单位计算长度。通过对语料观测后，本文最终选取最长与最短长度的阈值分别为36与5。

特定线索词语：一些特殊词所在的句子被选入摘要的概率要大于其他句子，如“表示”。我们称这类词为摘要句线索词。统计表明，有26%的句子含有线索词。本文利用这些词作为判断摘要句的标记。

位置特征：位置是一个重要特征，特别是在新闻语料中。一般而言，文章的首段和尾段以及段落的首句和尾句相比而言更为重要。文献[9]着重考虑了首句和尾句信息。本文采用了是否在首段、是否在段首、是否在段尾、是否在前2段的位置特征。

高频词：高频词是指在文章中出现频率较高的词，一般而言，词频越高，词的重要程度越大，所在的句子也可能更有代表性，但虚词例外。本文在利用停用词表进行过滤处理之后，再度量高频词。

数字、时间及专有名词：命名实体经常成为文章的焦点。本文在选择句子时，也使用了相关特征，包括数字、时间以及专有名词。

2.2 关联特征

一个句子是否选为摘要句，除了自身的特征外，也受到上下文的影响。关联特征是指影响摘要句选择的上下文信息。本文使用了如下几种关联特征：

与前一句的关系：该特征主要考察前一句是否是摘要句。根据观察，摘要句通常不会密集出现，相邻两句同为摘要句的概率较低。

与标题的相似度：标题包含了文本的重要信息，句子与标题相似度越大，则通常更可能出现在摘要中。

与其他句子的相似度：同高频词原理类似，该特征可以看作是寻找“高频句”，计算公式见公式(1)。

$$f_i = \sum_{j \neq i} sim(i, j) \tag{1}$$

$sim(i, j)$ 为句子 i 与句子 j 的相似度，通过 i 和 j 的词集的交集衡量。

与前后 2 句的相似度：句子与周围句子的相似度在一定程度上反应了句子在局部的重要性。

3 实现算法

3.1 CRF 模型

条件随机场（Conditional Random Fields, CRFs）模型是一种判别模型，其主要模型思想来源于最大熵模型。CRF 模型是在给定观察序列的前提下，计算整个标记序列的概率。CRF 模型可以较好地解决序列标注问题，在词性标注、命名实体识别、语块分析中都得到了很好的应用。

在 CRF 模型中，令 $\vec{x}$ 表示待标记的观察序列，$\vec{y}$ 表示对应的标注序列，则条件概率 $p(\vec{y} \mid \vec{x})$ 的计算方法见公式(2)。

$$p(\vec{y} \mid \vec{x}) = \frac{1}{Z(\vec{x})} \prod_{c \in C} \Psi_c(\vec{x}_c, \vec{y}_c) \tag{2}$$

Ψ_c 为对应最大团的势函数[10]。在 Linear-chain CRF 中，势函数 Ψ_c 的形式见公式(3)。

$$\Psi_j(\vec{x}, \vec{y}) = \exp(\sum_{1}^{m} \lambda_i f_i(y_{j-1}, y_j, x, j)) \tag{3}$$

$Z(\vec{x})$ 为归一化因子。

3.2 特征函数

在上面第 2 节已经描述了特征模板。在模型使用中，本文使用了二值特征值，通常定义为由输入变量 x 以及输出变量 y 的函数 f(x,y)，取值为 0 或 1。

3.3 修正因子

文本摘要中句子的数量远少于原文本中句子的数量，这样，被选句子的特征出现频率将会偏低。在使用序列标注时，原文本中的句子也倾向于不被选择，从而，导致实验结果中准确率较高而召回率较低的现象。

本文在 CRF 基础上，引入了修正因子。于是，一个句子 s 究竟是“在”还是“不在”，便由公式(4)决定。

$$label(s) = \arg\max_{c \in C} adjust(c) P(c \mid s) \tag{4}$$

其中 $adjust(c)$ 为类别 c 的修正因子，$P(c \mid s)$ 是由 CRF 模型计算出的类别 c 的条件概率。由于原文本中的句子要么被抽取，要么不被抽取，只需要将原文中的句子标记为“在”和“不在”两个标记之一即可，这样，$adjust(\text{在}) = 1 - adjust(\text{不在})$。

通过统计发现，在我们使用的训练语料中摘要句数不超过 3 句的比例高达 98.7%，这样，越长的文章其句子越容易被标记为“不在”。考虑这一因素，本文采用如下公式(5)计算修正因子。

$$adjust(x) = \frac{1}{\sqrt{Length(x)}} \tag{5}$$

其中 x 表示某个文本，a、b 为常数。Length(x)文本的长度，即文本所包含的句子数。

4 实验评测

4.1 实验语料

训练以及测试的语料来自对网易新闻。我们从网易上采集了大量带“核心提示”的新闻文本，核心提示可以看成摘要。在采集后，对其进行适当处理，最终得到 17610 篇新闻文本作为实验语料。所作的处理主要是剔除摘要句较少的文本。

此外，我们还发现，在“核心提示”中的有些句子并非完整地出现在原文之中，有的做了修改，有的由不同句子拼接而成。也就是说，“核心提示”并不完全是抽取式摘要，我们对这种情况做了剔除处理。为了判断是否完全对应，本文使用了基于字的最长公共子序列的算法实现核心提示中的句子与原文句子的对齐，以保证其正确性。

所有 17610 个新闻文本平均分为 5 部分，采用交叉验证方法评测，4 份用于训练，1 份用于测试。

4.2 评测准则

为评价摘要效果，本文采用准确率、召回率和 F 值 3 个标准来衡量，其中以 F 值为最重要的指标。这 3 个指标的计算公式公式(6)。由于摘要句按照原文出现的先后顺序出现在摘要中，因此语序并不是评测要求。

$$\mathrm{P}=\frac{a}{a+c} \quad R=\frac{a}{a+b} \quad F=\frac{2\mathrm{P}\times R}{\mathrm{P}+R} \tag{6}$$

其中：

P：准确率

R：召回率

a：在摘要中、同时被标记为摘要句的句子数

b：在摘要中、但是没有被标记为摘要句的句子数

c：不在摘要中、但是被标记为摘要句的句子数

4.3 实验设计

本文实验分为两组。第一组使用基本的线性 CRF 序列标注模型，同时与朴素贝叶斯和最大熵两种以往实验中用于摘要抽取的分类模型作为对比；第二组在线性 CRF 基础上增加了修正因子。

4.4 实验结果

表 1 列出了第一组实验的结果。朴素贝叶斯模型、最大熵模型、线性 CRF 模型均在同样的训练数据和测试数据下做的实验。为了比较，朴素贝叶斯模型与最大熵模型尽量采用了与 CRF 模型相同的特征，但考虑到这两种模型与 CRF 模型的差异，所采用的特征不包括与前一句的标记有关的特征。

表 1　不同模型下准确率、召回率以及 F 值

	准确率	召回率	F 值
朴素贝叶斯模型	53.5%	57.2%	55.3%
最大熵模型	66.2%	53.7%	59.2%
条件随机场模型	68.1%	53.8%	60.1%

表 1 显示，CRF 模型的综合效果(F 值)好于其他两种模型，但召回率不如朴素贝叶斯模型。我们观察标注的结果发现，当文本过长时，摘要句分布过于分散。位于文本中部的句子其位置特征

相对较弱，容易造成误判。文本长度对实验效果有着重要的影响。对实验语料统计发现，50%的文本长度超过 10 句，20%的文本长度超过 21 句。文本越长，句子抽取效果相对越差。

基于上述分析，本文进一步通过引入修正因子来比较了各项指标的变化。图 1 显示了不同修正因子下准确率、召回率以及 F 值的变化曲线。图中横轴所示为标记为拒绝为摘要句的修正因子（下面简称“修正因子”）。从图中可见，随着修正因子的增大，准确率升高，这是因为越趋向于拒绝，未被拒绝的句子更可能为摘要句。但从图中也可以发现，随着修正因子增大，尽管准确率提高了，但是召回率随之降低，并在修正因子取 0.45 之后，F 值也会下降。因此，确定合理的修正因子也是一个重要的问题。

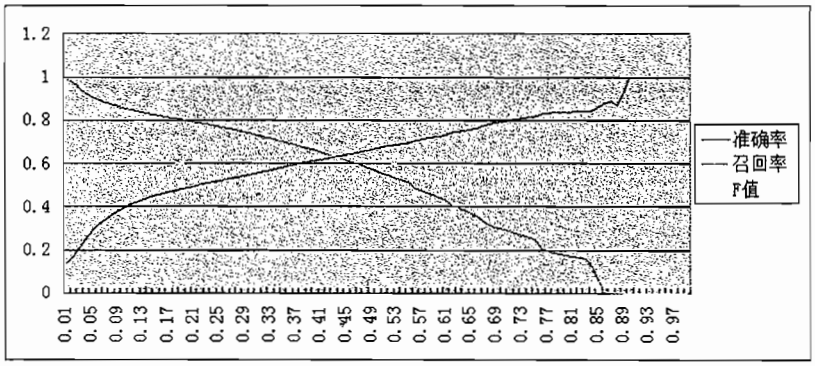

图 1　不同修正因子下准确率、召回率以及 F 值的变化

表 2 比较了根据公式 7，不加修正因子以及增加修正因子对交叉验证的实验结果的影响。

表 2　有无修正因子各项指标对比

	无修正因子	有修正因子
准确率	68.1%	57.1%
召回率	53.8%	69.2%
F 值	60.1%	62.5%

从表 2 中可以看出，采用了适当的修正因子后，F 值与无修正因子相比提高了 4.2%。总体来看，适当增加修正因子可以在一定程度上提高 F 值，具有更好的效果。

5　总结

本文所提出的方法是作者在总结已有的文本摘要算法后做的一个初步尝试，虽然取得了较好的结果，但是仍有许多值得商榷和改进。比如考虑到文本长度的影响，增加修正因子以提高相应指标，虽然对实验效果有一定的提高，但仍有改进余地，今后可以考虑多种因素的修正因子。在训练及测试语料方面依赖于网易新闻所提供的新闻数据，由于时间仓促、数量巨大，标注过程中难免会有疏漏，这都是今后可以改进的地方。

参 考 文 献

[1] Dipanjan Das, Andre F.T.Martins. A survey on Automatic Text Summarization. Literature Survey for the Language and Statistics II. Nov 21, 2007.

[2] Luhn, H. P. The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2): 159-165. 1958.

[3] Baxendale, P. Machine-made index for technical literature - an experiment. IBM Journal of Research Development, 2(4): 354-361. 1958.

[4] Edmundson, H. P. New methods in automatic extracting. Journal of the ACM, 16(2): 264-285. 1999.

[5] Kupiec, J., Pedersen, J., and Chen, F. A trainable document summarizer. In Proceedings SIGIR '95, pages 68-73, New York, NY, USA. 1995.

[6] Aone, C., Okurowski, M. E., Gorlinsky, J., and Larsen, B. A trainable summarizer with knowledge acquired from robust nlp techniques. In Mani, I. and Maybury, M. T., editors, Advances in Automatic Text Summarization, pages 71-80. MIT Press. 1999.

[7] Lin, C. -Y. Training a selection function for extraction. In Proceedings of CIKM '99, pages 55-62, New York, NY, USA. 1999.

[8] Conroy, J. M. and O'leary, D. P. Text summarization via hidden markov models. In Proceedings of SIGIR '01, pages 406-407, New York, NY, USA. 2001.

[9] D. Shen, J. T. Sun, H. Li, Q. Yang, Z. Chen, Document Summarization using Conditional Random Fields[C], In IJCAI, pages 1805-1813. 2007.

[10] Kschischang, Frank; Frey, Brendan J.; Loeliger, Hans-Andrea: Factor Graphs and the Sum-Product Algorithm. In: IEEE Transactions on Information Theory 47 (2001), No. 2, pages 498-519. 2001.