高技术通讯 2010年 第20卷 第12期:1224~1228

doi:10.3772/j.issn.1002-0470.2010.12.003

结合文本聚类和文本检索的语料选取方法①

何 峰② 丁晓青③

(清华大学电子工程系 北京 100084)

摘 要 为了克服用应用相关的文本数据进行语音识别、智能输入等各种自然语言处理中在有些情况下因很难收集到充足的相关数据和缺乏应用相关的训练数据带来的困难,提出了一种通过结合非监督文本聚类和文本检索技术实现相关语料选取的新方法。该方法仅使用少量与特定应用相关的文本,即可从未经整理的大规模语料库中发现更多与此应用相关的文本。利用该方法在手机短信文本和未经整理的大规模语料库上进行了实验,实验结果表明该方法能够有效提取应用相关的文本。

关键词 文本聚类,文本检索,Kullback-Leibler距离,统计语言模型

0 引 言

应用相关或领域相关的文本在各种自然语言处理中非常有用。在诸如自动语音识别(automatic speech recognition, ASR)、利用上下文信息的光学字符识别(optical character recognition, OCR)以及机器翻译等各种应用中,应用相关文本或领域相关文本能够有效改进词典选择和统计语言模型的性能。目前,在这方面已经有一些相关的研究工作。文献[1]先形成内容主题聚类,基于数量相对较少的聚类集合得到分量语言模型,通过对各个分量语言模型分配权重得到混合语言模型。利用已经整理好并有手工标注的语料,文献[2]基于与特定主题相关的文本建立了树形主题体系以及适应于此主题的语言模型。通过利用初始识别结果检索相关文档,一些ASR系统[3,4]进行了在线语言模型调整,并利用调整得到的新语言模型重新进行语音识别。文献[5]在机器翻译的文本语料的帮助下,利用语言模型自适应技术改善了基于少量的任务相关文本建立的语言模型,并利用英语到冰岛语的翻译得到的数据进行冰岛语语音识别实验,实验表明显著减少了词错误率。

然而,在一些情况下,由于过于繁重的人力劳动负担或者出于保护隐私的考虑,难以收集到充足的应用相关文本。在本文的情况下,仅有少量中文手机短信文本作为应用相关文本,同时,大规模语料库是未经整理的,没有内容领域标记,没有内容关键词,甚至文章之间也没有分隔标记。我们知道,手机短信内容的主题各种各样,其主要特征在于口语化的语言风格,而我们手头能够获得的大规模语料库的文本数据主要来自于中文书籍、报刊,我们需要从这些语言风格更书面化的大量文本数据中挑选出风格与手机短信相近的文本数据,从而为改进手机中的智能输入法提供更多的训练数据。为解决上述问题,本文提出了一种结合文本聚类和文本检索实现相关语料选取的新方法。在手机短信文本和未经整理的大规模语料库上进行的实验表明,本方法能够有效选取应用相关的文本。

1 文本选取方法的整体设计

首先,将未经整理的大规模语料库机械地分隔为大量长度相等的小文本块(此后称为文档)。然后,语料选取方法主要包括两个步骤:第一步是文本聚类,采用大规模文本聚类方法[6]对大量文档进行聚类,形成数百个文本类;第二步,以少量手机短信文本作为查询语料,在聚类得到的文本类中进行检索,按照相对于短信的相关度,将这些文本类进行排序,从而实现从大规模语料中选取出与手机短信相关的文本。

由于聚类的类别数量比较大,聚类得到的各个

① 973计划(2007CB311004)资助项目。
② 男,1979年生,博士生;研究方向:自然语言处理,模式识别,人工智能;E-mail:hefeng97@mails.tsinghua.edu.cn
③ 通讯作者,E-mail:dxq@ocrserv.ee.tsinghua.edu.cn
(收稿日期:2009-08-11)

类别具有较好的内部一致性,因此,在文本检索步骤中,相比直接对单个文档进行检索,对文本类进行检索在统计上更加鲁棒,同时,又实现了弹性,即,将检索种子语料由本文中使用的手机短信替换为其它应用相关或领域相关的文本数据,本方法就可以方便地扩展到将来可能的其它应用领域。

1.1　大规模文本聚类

由于中文词项之间没有天然界限,先对中文句子进行分词,以获得词项序列。然后,滤除在整个文档集合中出现次数极少或者仅在极少量文档中出现的词项。

在文本聚类中采用广泛使用的向量空间模型,将各个文档表示为词项权重向量。假设有 $|V|$ 个词项,V 表示选择出来的词典,则将各个文档表示为 $|V|$ 维的向量。使用词频-逆文档频率(term frequency-inverse document frequency, TF-IDF)权重计算方法[7],文档 j 中的词项 i 的 TF-IDF 权重表示如下:

$$TFIDF_{ij} = \frac{TFIDF'_{ij}}{\sqrt{\sum_{k}^{|V|}(TFIDF'_{kj})^2}}$$

$$TFIDF'_{ij} = tf_{ij} \times \log(idf_i)$$

$$idf_i = \frac{|D|}{df(w_i)}$$

其中,tf_{ij} 称为词项频率,即,词项 i 在文档 j 中出现的次数,$|D|$ 是整个集合中的文档数,而 $df(w_i)$ 是其中出现了词项 i 的文档数目,并且,在信息检索领域中,将 idf_i 这一项称为倒文档频率。这样,得到文档 d_j 的向量表示:

$$d_j = (TFIDF_{1j}, TFIDF_{2j}, \cdots, TFIDF_{|V|j})$$

使用文档向量的内积作为文档之间的相似度度量:

$$similarity(d_i, d_j) = \sum_{k=1}^{|V|}(TFIDF_{k,i} \times TFIDF_{k,j})$$

由于已经对文档向量进行了归一化,此相似度量也即是两个文档向量之间夹角的余弦。

然后,利用 k 均值聚类方法对这些文档进行聚类。此处使用的方法与文献[8]中提出的类似。然而,由于 k 均值方法对于初始类别分配比较敏感,其仅得到局部最优的聚类结果。在本文中,在每次 k 均值聚类的初始化阶段,随机为各个文档分配类别,但是,进行多次 k 均值聚类,选择使得以下目标函数 Q 具有最大值的那一次聚类作为最终的聚类结果。目标函数 Q 定义如下:

$$Q(\{\pi_j\}_{j=1}^{k}) = \sum_{j=1}^{k}\sum_{x_i \in \pi_j} x_i^{\mathrm{T}} c_j$$

其中,π_j 表示类别 j,x_i 是文档 i 的向量,而 c_j 是类别 j 的中心向量,定义为

$$\vec{c}_j = \frac{1}{|\pi_j|}\sum_{x_i \in \pi_j}\vec{x}_i$$

函数 Q 表示所得到的聚类类别的类内聚合度,因此,其实质上可以作为聚类结果的性能评价。

由以上文本聚类处理,我们将大量文档聚类形成数百个文本类。

1.2　文本检索

经过以上文本聚类后,将每个文本类别作为文本检索的基本单元,使用少量应用相关的文本作为查询文本,采用两种不同的文本检索方法从文本聚类集合中检索出相关文本。基于文本类与应用相关的查询文本的相似度,降序排列文本类集合中的各文本类别。

1.2.1　TF-IDF 方法

一种检索方法是 TF-IDF 方法,此处,文档的表示以及文档之间的相似度量和聚类步骤相同。将每个文本类别视为单个文档,用 TF-IDF 加权的向量进行表示,并将文档向量之间的内积定义为文档之间的相似度。与上述文本聚类处理不同的是,在此步骤中,我们仅将各文本类与查询文本进行比较,然后,根据文本类与查询文本的相似度对这些文本类进行排序。

1.2.2　Kullback-Leibler(KL)方法

另一种检索方法采用 Kullback-Leibler(KL)散度作为文档之间的相似度量。KL 散度,也称为交叉熵或相对信息,其度量两种概率分布的差异。KL 散度越小,则表明两种概率分布越相近。在我们的方法中,KL 方法计算各个文本类与作为查询语料的应用相关文本这两者的一元词项概率分布之间的交叉熵。即,认为 KL 散度是查询文本和文本类之间的相似度量如下:

$$dissimilarity(query, cluster) = \sum_{k=1}^{|V|}\{p(w_k \mid q)\log\frac{p(w_k \mid q)}{p(w_k \mid cluster)}\}$$

其中,V 是词典,$p(w_k \mid q)$ 是词项 w_k 在查询文本中的概率,$p(w_k \mid cluster)$ 是词项 w_k 在文本类中的概率。

由于词典 V 中的很多词项在查询文本和单个文本类中没有出现,直接用词项在文本单元中出现的频率来估计词项概率将导致词项概率为零。为了避免一元词项概率为零,并确保对应的词项概率总和为 1,在本文中,采用 Good-Turing[9] 平滑算法估计

各文本单元中的一元词项概率。Good-Turing平滑算法对文本单元中出现的词项频率进行折扣,并为没有出现的词项赋以很小的概率。此平滑算法公式表示如下:

$$P_{GT}(w_i)=\begin{cases}\dfrac{r^*}{N},\ r^*=\dfrac{(r+1)N_{r+1}}{N_r},\\ \qquad 0<r=C(w_i)\leqslant k\dfrac{r}{N},r>k\\ \dfrac{1-\sum\limits_{C(w_i)>0}N_rP_{GT}(w_i)}{N_0}\approx\dfrac{N_1}{N_0N},r=0\end{cases}$$

其中,$P_{GT}(w_i)$是词项w_i在文本单元中的概率,使用Good-Turing平滑方法进行估计,r和$C(w_i)$是词项w_i在此文本单元中的频率,k是预定义的常数(通常选为7),N是此文本单元中所有词项的出现次数之和,N_r表示在文本单元中出现r次的词项的个数,例如,N_0是词典中没有在文本单元中出现过的词项的个数,N_1则是仅在文本单元中出现过一次的词项的个数。

与TF-IDF方法类似,采用此KL散度作为相似度量,根据各文本类相对于感兴趣的应用相关文本的KL散度对这些文本类进行相关排序。

如上所述,通过有效地结合文本聚类和文本检索,将文本类按照其与应用相关文本的相似进行排序,从而选取出更多与应用相关的文本。

2 实验及结果分析

2.1 实验数据描述

在少量中文手机短信文本以及大规模语料库上进行了实验,其中,手机短信文本作为应用相关语料,而大规模语料库主要由从中文书籍报刊收集到的文本组成。

手机短信文本包含42606条短消息,大约25.5万个中文字符,而大规模语料库包含约4.53亿个中文字符。

2.2 实验及结果

将大规模语料机械地顺序分割为140230个大小几乎相等的文本文档,每个文档约3000个中文字符。

由于中文词之间没有天然的分界符号,作为预处理步骤,进行中文分词由句子得到词项序列。

由于短信文本中有大量停顿词,这也是短信文本的一种语言风格特征,因此,和通常的文本检索应用不同,此处不滤除停顿词。

在文本聚类步骤中,滤除了在整个文本集合中出现次数少于10次的词项以及在少于10个文档中出现的词项之后,向量空间模型中剩余75872个不同的词项。

对文档集合进行5次k均值聚类,形成200个文本类别,自动地选取使得前述目标函数Q最大的那次聚类作为最终的结果。在所得到的200个文本类别中,最大的类包含8070个文档,最小的包含222个文档。

在文本检索步骤中,将每个文本类作为一个基本的检索单元,使用手机短信文本作为查询文本,在利用前述的TF-IDF方法或KL方法进行检索之后,根据相对于短消息文本的相似度降序排列这200个文本类。

为了验证所提出的语料选取方法的性能,采用广泛使用的n元语言模型的复杂度作为评价标准。语言模型的复杂度度量从训练文本建立的语言模型与短信文本的匹配程度。复杂度越低,则说明语言模型越匹配短信文本,也就说明训练文本与短信文本相关度越好。

为建立统计可靠的n元语言模型,将利用KL和TF-IDF检索方法获得的200个文本类依照相关排序的顺序集合成具有相近大小的10个文本数据集。使用SRILM程序包[10],由各个文本数据集训练得到采用Good-Turing折扣法、Katz后退法的3元语言模型。为了纯粹检验语料选取方法,在n元语言模型训练过程中没有进行n元项剪除。

使用短信文本作为测试语料进行测试,得到这10个3元语言模型的每个汉字的平均复杂度(perplexity per character,PPC),从而评价各个文本数据集与短信文本之间的相关程度。

由于大规模语料库与短信文本并不非常匹配,所得到的语言模型的PPC相对较高。在用KL方法检索到的文本数据集训练得到的10个语言模型中,由相关排序最靠前的文本集训练出的语言模型的PPC最低,为99.2,而由相关排序最靠后的文本集训练出的语言模型的PPC最高,为227.1,并且,随着文本集相关排序的降低,其对应的语言模型的PPC逐渐增高。用TF-IDF检索方法替换KL检索方法,可得到类似结果,其中,最低的PPC为98.8,最高的PPC为224.8。这说明本文的语料选取方法确实能够根据手机短信文本很好地区分和组织大规模语料库,从而从语料库中选取出与短信数据最相关的语料。

为了进一步测试所述语料选取方法,我们比较从通过语料选取方法得到的语料建立的语言模型以及从原始的未整理的语料建立的语言模型在手机短信测试文本上的表现。将短信文本分为两个部分:一部分作为验证集,其包含 80% 的短信;另一部分作为测试集,其包含剩余的 20% 的短信。

由于在整个大规模语料库上利用 SRILM 程序包进行语言模型训练过于耗费内存,速度极慢,我们不能由整个语料库直接训练得到一个整体的语言模型。为此,首先将原始的大规模语料库随机划分为具有相等大小的 10 个文本数据集。由这 10 个数据集训练得到经过 Good-Turing 折扣和 Katz 后退处理的 3 元统计语言模型。在词项级别对这 10 个语言模型进行插值组合,构建出混合语言模型,其中,使用期望最大(expectation-maximization, EM)算法[11]从短信验证集进行最优估计得出各分量语言模型的权重。将此混合语言模型作为基准模型,其在短信测试集上得到 PPC 为 115.4 这样的测试结果。

然后,对两组分别由 KL 检索方法和 TF-IDF 检索方法得到的 10 个文本集训练出来的 10 个语言模型在词项级别进行插值组合,得到两个混合语言模型,其中,同样利用 EM 算法在验证集上进行最优估值得到各分量模型的权重。在两个混合语言模型中,排序居前两位的两个分量语言模型的权重之和均约为 0.9,而居后的 8 个分量模型的权重总和才 0.1,这也说明了经过选取整理的语料很好地适配了手机短信应用。表 1 给出了这两个混合语言模型与基准语言模型在测试集上的 PPC 结果的比较。

表 1　各语言模型的 PPC 结果

	基准	KL	TF-IDF
PPC	115.4	94.9	94.7

由表 1 可见,分别利用通过 KL 检索方法以及 TF-IDF 检索方法获取的整理语料建立的两个混合语言模型在测试集上具有相近的 PPC,相比基准混合语言模型,它们的 PPC 结果减小了 17.8%,这表明了本文结合文本聚类和文本检索进行语料选取的方法的有效性。

3　结　论

本文中,为了处理各种自然语言应用中经常出现的缺乏应用相关训练语料的问题,提出了一种新的语料选取方法。本文方法结合大规模文本聚类和文本检索技术,根据少量的应用相关文本对未经整理的大规模原始语料进行组织和排序,从而从中选取出与应用比较相关的语料。文中使用 n 元统计语言模型在应用相关语料上的复杂度结果评价此语料选取方法。实验结果显示,由经过整理排序的文本集训练得到的语言模型在应用相关文本上的复杂度结果具有明显区别,这表明了本文的语料选取方法的有效性。

在以后的研究中,将使用其它领域或应用相关文本数据检验所提出的语料选取方法,考察和分析聚类对语料选取性能的影响。对于已标注样本较少的文本分类问题,进一步工作还可根据特征词项在各类文本中的分布对特征词项进行有监督的聚类,得到语义相近的词项类,并以此作为统计单元进行模型统计,可减少待估计参数。

参考文献

[1] Clarkson P R, Robinson A J. Language model adaptation using mixtures and an exponentially decaying cache. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, 1997. 799-802

[2] Seymore K, Rosenfeld R. Using story topics for language model adaptation. In: Proceedings of the European Conference on Speech Communication and Technology, Rhodes, Greece, 1997

[3] Souvignier B, Kellner A. Online adaptation for language models in spoken dialogue systems. In: Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, 1998

[4] Nanjo H, Kawahara T. Unsupervised language model adaptation for lecture speech recognition. In: Proceedings of the International Conference on Spoken Language Processing, Denver, USA, 2002

[5] Jensson A T, Iwano K, Furui S. Language model adaptation using machine-translated text for resource-deficient languages. *EURASIP Journal on Audio, Speech, and Music Processing*, 2008

[6] Manning C, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press, 1999. 495-500

[7] Manning C, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press, 1999. 539-544

[8] Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 2001,

42(1/2), 143-175

[9] Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, San Francisco, USA, 1996. 310-318

[10] Stolcke A. SRILM: An extensible language modeling toolkit. In: Proceedings of International. Conference on Spoken Language Processing, Denver, USA, 2002

[11] Jelinek F. Readings in Speech Recognition. Cambridge: Morgan Kaufman Publishers, 1990. 450-506

Combining text clustering and text retrieval for corpus adaptation

He Feng, Ding Xiaoqing

(Department of Electronic Engineering, Tsinghua University, Beijing 100084)

Abstract

In order to solve the difficulties brought about in some situations when using the application-relevant text data to do various natural language processings, such as automatic speech recognition and intelligent input due to the hard collection of relevant data and the scarcity of application-relevant training texts, this paper presents a novel method for corpus adaptation by combining the unsupervised text clustering and text retrieval techniques. The method only uses a small set of application specific text to find the relevant text from a large scale of unorganized corpus, thereby, it adapts training corpus towards the application area of interest. The performance of the *n*-gram statistical language model, which was trained from the text retrieved and tested on the application-specific text, was used to evaluate the relevance of the text acquired. The preliminary experiments on short message texts and unorganized large corpus demonstrated the good performance of the proposed method.

Key words: text clustering, text retrieval, Kullback-Leibler distance, statistical language model