文章编号:1007-9432(2011)02-0133-05

利用统计量和语言学规则提取多字词表达

刘 荣,王奕凯
(太原理工大学 外国语学院,太原 030024)

摘 要:基于特定领域的语料库,利用统计和语言学规则相结合的方法提取多字词表达(Multiword expressions)。首先利用领域高频词作为种子词提取候选串,进一步利用各种统计量、多字词表达边界过滤规则对候选串进行噪声剔除,得到多字词表达。实验结果表明,该方法对于处理大规模真实文本效率很高,可以有效提高多字词表达的获取,可以更有针对性地在特定领域提取多字词表达。

关键词:多字词表达;互信息;熵;语言学规则
中图分类号:TP391.1 **文献标识码**:A

多字词表达可以定义为不可根据部分组成成分而知其意的具有句法或语义特质的任意词语组合[1]。多字词表达包括动词短语(爆发、取决于),复合名词(机器翻译、警用车辆),成语(雨后春笋、三人行必有我师)等等。在日常语言生活中,多字词表达大量存在。多字词表达在日常生活中使用频率很高,Jackendoff[2]推测在一个人的词典中多字词表达的数量和单个词的数量级是一样的。在 WordNet 中几乎一半的词条是多字词表达。Biber et al.[3]指出,在英语口语中大约 30 %到 45 %的内容,学术文章中 21 %的内容都是由多字词表达构成。温端政[4]指出"事实上,汉语里语的数量并不比词少,而是相反,至少是平分秋色"。由此可见,在汉语中,语的数量也是非常庞大的。汉语中,除了传统语言学定义的成语、谚语、歇后语、惯用语之外,没有被词典收录的应该就是多字词表达。对于特定领域而言,专业词汇组成了大量的多字词表达,并且新的多字词表达层出不穷。Helena de Medeiros Caseli et al.[5]指出在这个意义上,在特定领域多字词表达的数量有可能被低估了。在自然语言理解和自然语言处理的许多实际应用中,多字词表达有着很重要的作用。如果不能有效识别多字词表达,就会对自然语言处理带来很多困难,特别是涉及到语义处理的问题。对于句法分析任务而言,Baldwin et al.[6]发现,在英国国家语料库中随机选取 20,000 句,在所有句法分析错误中,因为不能识别多字词表达而产生的错误高达 8 %。因此,自然语言处理需要鲁棒的自动或半自动方法提取多字词表达,进而进行资源建设。此外,由于多字词表达与所属语言和所属文化有关,识别合适的翻译对是机器翻译的一个难题。本文所提取的多字词表达有广泛的应用前景,它对于词典编纂、中文词语的歧义消解、提高中文文本自动分类的准确率、提高搜索引擎的效率、中文信息处理的浅层句法分析、自动文摘、信息抽取、对外汉语教学的教材更新、机器翻译等方面都会有所帮助。

1 多字词表达提取的基本方法

在提取多字词表达时,最简单的方法就是统计词串的数量,这种统计方法虽然简单而有效,但是会产生大量的噪音,从而引起准确率不高的问题。以往的词串统计方法只计算了词串的频度信息,在抽取结果中会包含很多不合语法和语义的词串。为了消除词串统计方法的缺陷,就必须将更多的统计量综合计算。在各种统计量中,有的统计量可以度量多字词表达内部的结合紧密程度,例如:t-score、互信息(MI)、x^2 值、对数可能性值(Log-likelihood)等。我们可以把这些方法称作内部方法[7]。有的统计量可以度量多字词表达的独立性即外部边界,例如左右熵的方法[8]。我们把这些方法称作外部边界方法。

收稿日期:2010-10-20
基金项目:国家自然科学基金项目(60663008);山西省出国留学人员项目(2009 - 33);太原理工大学青年基金项目(900103 - 03010255)
作者简介:刘荣(1972 -),男,山西太原人,讲师,主要从事外语教学与研究,(Tel)13935157502

根据文献和本文处理的数据规模,我们选取了频次、互信息、左右熵这三个统计量,判别候选多字词表达的内部结合紧密程度和外部边界独立性;随后还借助语言学规则和停用词表对候选多字词表达进行过滤,去筛选统计方法不易判断的候选多字词表达确认多字词表达的合法性。

如上所述,本文的技术路线是采用各种统计量结合语言规则。图1是本文汉语多字词表达提取的流程。

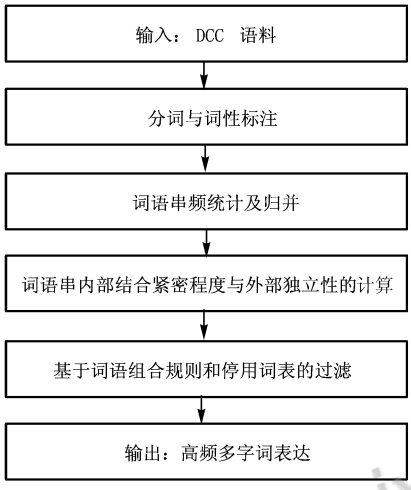

图1 中文多字词表达的提取流程

2 多字词表达提取的技术路线

2.1 特定领域种子词的确认

为了方便教学以及考察提取的多字词表达的效果,笔者选择特定领域的文本作为处理对象。本文所用语料是北京语言大学DCC语料库2007年教育类文本,大小是154 MB。首先,我们对教育类文本进行分词和词性标注。

在某个特定领域高频出现的词语有可能表现这个领域的特色。在笔者所选取的教育领域文本中,“学校、学生”等反映了教育领域文本的特点,但是频次排序最高的“的、是、在、一、了”等词并不能代表教育领域的特点。因而,我们采用位序比的方法去提取教育领域特色的词语。位序比值的计算公式可以表示为:

$$C_I = \frac{L(a,I)}{L(b,I)}. \qquad (1)$$

式中:I代表某个词;a,b代表不同的表;$L(a,I)$代表词I在表a中的频次排列的位序;$L(b,I)$代表词I在表b中的频次排列的位序;C_I代表词I在表a和表b中的位序的比值。

具体的操作步骤是先将教育领域和其他领域词表中的词按频次排序,再将教育领域词语表同其他领域词表导入同一个数据表中,对同一个词在两表中的排序位次值做除法,在教育领域词表中按照位序比值设定阈值抽取词语。

2.2 利用特定领域高频种子词提取词语串

通过上述位序比的方法,抽取领域种子词。最后我们将教育类和其它领域共有的按照频次排序前5 000词作为领域种子词。接着,以领域种子词为锚点进行开窗口运算并合并相同的词语串。通过对数据的考察,我们发现二元的词语串数量最多,占所有词语串的65 %。下面我们对这些二元词语串进行统计量的计算。

3 基于统计量的筛选方法

利用领域种子词提取出的词串有很多不具备语法和语义意义。对于这些词串,我们利用统计量进行筛选。我们选取的统计量分别从词串内部的结合紧密程度和词串外部的边界度量入手。具体的筛选方法包括互信息和左右熵这两个统计量的计算和阈值的设定。

3.1 内部边界的判定方法

从统计学的视角而言,多字词表达内部词语之间的结合紧密程度依赖于词语的共现频度。如果某些互相成搭配的词语串反复大量出现,即它们的共现频度越高,那么词语串的结合紧密性越强。因此,高频的词语串可能是一个完整的多字词表达。

罗盛芬等[10]对9种常用统计量在汉语自动抽词中的表现进行了考察,对汉语二字词的自动抽词实验结果表明,这9种常用统计量中,互信息的抽词能力最强。他们指出这9种统计量并不具备良好的互补性;一般情况下,可以直接选用互信息这种简单有效的统计量进行二元自动抽词。

根据罗盛芬等的实验结果,笔者利用互信息这个统计量判断词串内部结合紧密程度。

互信息体现了两个变量之间的相互依赖程度。二元互信息是指两个事件同时发生的概率函数:

$$\mathrm{MI}(X,Y) = \log_2 \frac{P(X,Y)}{P(X)P(Y)}. \qquad (2)$$

互信息值越高,表明X和Y相关性越高,则X和Y组成短语的可能性越大;反之,互信息值越低,X和Y之间相关性越低,则X和Y之间存在短语边界的可能性越大。

3.2 外部边界的判定方法

为了确保多字词表达具有合法的语义,必须保证多字词表达是独立的语言单位。笔者采用另外一

种统计方法来判定多字词表达的边界。Hung et al.[11]对自然语言处理中的各个统计量进行了比较,指出左右熵能够确定多字词表达的非解构性。本文使用左右熵的方法(Left and Right Entropy Measures)来判别候选多字词表达的独立性和边界。

熵这个术语表示随机变量不确定性的量度。具体表述如下:一般地,设 X 是取有限个值的随机变量(或者说 X 是有限个离散事件的概率场),X 取值 x 的概率为 $P(x)$,则 X 的熵定义为:

$$H(X) = -\sum_{(x \in X)} P(x) \cdot \log_2 P(x). \quad (3)$$

左右熵是指多字词表达的左边界的熵和右边界的熵。左右熵的公式如下:

$$E_L(W) = -\sum_{\forall a \in A} P(aW \mid W) \cdot \log_2 P(aW \mid W). \quad (4)$$

$$E_R(W) = -\sum_{\forall b \in B} P(Wb \mid W) \cdot \log_2 P(Wb \mid W). \quad (5)$$

在上面的公式中,E_L 和 E_R 分别表示词串的左熵(Left Entropy)和右熵(Right Entropy);W 表示 N-gram 的词语串,$W = \{w_1, w_2, \cdots, w_n\}$;$A$ 表示词串左边出现的所有词语的集合,a 表示左边出现的某一个词语;B 表示词串右边出现的所有词语的集合,b 表示右边出现的某一个词语。如果词串的 E_L 和 E_R 数值越大,即词串 W 左右出现的词语越多,W 就更有可能是一个完整的多字词表达。

4　基于语言学规则的筛选方法

虽然统计方法能大规模地对词串进行筛选,但是统计方法不是万能的。满足统计值的词串并不一定是具有语义完备性的多字词表达。例如"下都满足"。因此,我们考虑引入语言学规则进行筛选。

4.1　停用词表法

通过互信息计算,我们可以确定一些停用词。停用词指的是不能出现在一个词串首部和尾部的词。例如在词性序列把/家长、把/孩子、是/家长、是/孩子等词串中,"把"、"是"就作为停用词被选出来。停用词表完全通过人工筛选来完成,并可根据数据的实际情况增加。

借助词性标注和停用词表,我们对词串进行简单的过滤,可以将首部和尾部含有停用词的词串过滤掉。算法执行如下:

BEGIN

Step1　读入一条词串,如果词串首部和尾部没有停用词,到 step3,否则删除该词条并且到 step2

Setp2　读入下一个词串

Step3　输出词串

Step4　如果读完最后一个词串,就退出;否则到 Step1,读入下一条词串。

4.2　句法规则法

汉语的构词方式(偏正、联合等)决定了汉语的多字词表达至少有一个中心词,中心词位于多字词表达的左端或右端。通过对汉语词类和多字词表达边界的考察,我们发现有些词类是不能出现在多字词表达的前边界或后边界。例如"助词、量词"不能位于多字词表达的前边界;"副词、连词"不能出现在多字词表达的后边界。根据我们的语料,在总结了这些语言学规则之后,我们利用这些规则对词串进行筛选。

基于句法规则,过滤掉不符合句法规则的词串。算法执行过程如下:

Step1　读入一条词串,如果词串的最左边没有不能出现在多字词表达左边界的词类或词串最右边没有不能出现在多字词表达右边界的词类,到 step3,否则删除该词条并且到 step2;

Setp2　读入下一个词串;

Step3　输出多字词表达;

Step4　如果读完最后一个词串,就退出;否则到 Step1,读入下一条词串。

5　实验数据及实验结果

5.1　特定领域文本的提取

本文的语料来自北京语言大学应用语言研究所动态流通语料库 2007 年的文本,采用"中文新闻信息分类与代码"进行分类,我们选取了领域特定较为不同的 4 类文本,它们分别是:

教育类文本 154 MB;经济类文本 288 MB;体育类文本 172 MB;娱乐类文本 207 MB。

5.2　各种统计量阈值的设定

当求出一个词串的互信息、左熵和右熵后,如何判定这个词串是不是一个词语,并无一定的标准,需要试探不同的阈值。根据所用语料的规模和特点,经过反复实验,我们确定了互信息和左右熵的阈值。结合紧密的词语必然搭配强度大,其互信息值越高。根据互信息公式,我们设定了互信息值大于 0 作为词串内部结合紧密程度的一个判定标准。对于全部数据的左熵和右熵综合考察,通过对不同阈值的观察,我们发现左熵和右熵都大于或等于 3 的词串提取效果较好,因而将左熵和右熵的阈值设定为 3。

5.3 实验结果及分析(见表1)

表1 部分实验结果表

二字格	网/瘾、家/长、真/题,招/办、课/改、插/班、买/书、审/题、写/手、加/分、降/分、读/研、查/分、估/分、停/招、考/录
三字格	理工/类、冬令/营、月/收入、预科/班、外语/类、重点/线、大学/城、少年/班、初中/部、高职/类、录取线
四字格	名牌/大学、基础/知识、信息/技术、重点/中学、职业/生涯、校园/文化、高中/阶段、思维/方式、师资/队伍、技术/人员、科研/成果、师资/力量、电子/邮件、科研/项目
五字格	贫困/大学生、著名/科学家、著名/商学院、毕业生/就业、大学生/就业、公务员/考试、大学生/创业、公务员/招考、经济/全球化、多媒体/教学、文化课/考试、研究生/入学、志愿者/招募、研究生/复试
六字格	普通/高等院校、著名/经济学家、毕业生/就业率、教育/行政部门、出国/留学人员、爱国主义/教育、精神文明/建设、基础科学/研究
七字格	社会主义/荣辱观、马克思主义/哲学、马克思主义/理论
八字格	社会主义/市场经济、社会主义/精神文明

实验结果表明,本文所采用的统计量结合语言学规则对多字词表达的提取是行之有效的,效果比较好。在全部数据中,四字格的最多,八字格的最少。但是,如果左右熵阈值设定小于3时,除了合乎句法语义要求的多字词表达之外,在抽取结果中,也有部分不合适的词串。造成这一现象的原因主要是因为本文所采用的筛选规则还不完备,不能全部判定多字词表达内部结构和外部边界的合法性。

6 结语

笔者提出了一种统计和规则相结合的多字词表达抽取方法。首先用位序比的方法选取特定领域的高频词,利用高频词抽取词串,继而利用统计方法和语言学规则对词串进行筛选。最后提取出的多字词表达取得了比较令人满意的结果。选用高频词提取词串避免了数据稀疏的问题。在统计方法中,互信息和左右熵相结合的方法能够度量词串的内部结合紧密程度和外部边界。此外,语言学规则和停用词对词串的过滤有一定的帮助。实验结果表明统计和规则相结合混合的方法比完全借助统计的方法更好。根据需要,可以设定不同阈值并辅以人工校对,提取的多字词表达准确率会更高。本文所提出的技术路线易于操作,可以很好地应用于大规模真实文本语料库。

参考文献:

[1] Ivan A Sag, Timothy Baldwin, Francis Bond, et al. Multiword Expressions: A Pain in the Neck for NLP[C]//Proceedings of the Third International conference on computational linguistics and intelligent text processing, 2002: 1-15.

[2] Ray Jackendoff. Twistin' the night away[J]. Language, 1997, 73: 534-539.

[3] Douglas Biber, Stig Johansson, Geoffrey Leech, et al. Grammar of Spoken and Written English[M]. London: Longman, 1997.

[4] 温端政. 汉语语汇学[M]. 北京: 商务印书馆, 2005.

[5] Helena M Caseli, Carlos Ramisch, Maria G V Nunes, et al. Alignment-based extraction of multiword expressions[C]//Language Resources and Evaluation to appear, 2009.

[6] Timothy Baldwin, Emily M Bender, Dan Flickinger Ara Kim, et al. Road-testing the English Resource Grammar over the British National Corpus[C]//Proceedings of the fourth international conference on language resources and evaluation, 2004: 2047-2050.

[7] Magerman D, Marcus M. Parsing a natural language using mutual information statistics[C]//Proceedings of AAAI'90, 1990: 984-989.

[8] 谌贻荣. 中文术语自动提取技术研究[D]. 北京: 北京大学, 2005.

[9] 姜柄圭, 张秦龙, 谌贻荣, 等. 面向机器辅助翻译的汉语语块自动抽取研究[J]. 中文信息学报, 2007, 21(1): 9-16.

[10] 罗盛芬, 孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报, 2003, 17(3): 9-14.

[11] Hung Huu Hoang, Su Nam Kim, Min-Yen Kan. A Re-examination of Lexical Association Measures[C]//Proceedings of the 2009 Workshop on Multiword Expressions, 2009: 31-39.

[12] Shailaja Venkastsubramanyan, Jose Perez-Carballo. Multiword Expression Filtering for Building Knowledge Maps[C]//Second ACL workshop on multiword expressions, 2002: 40-47.

Extracting Multiword Expressions with Statistics and Linguistic Rules

LIU Rong, WANG Yi-kai

(College of Foreign Languages, TUT, Taiyuan 030024, China)

Abstract: Multiword Expressions (MWEs) are one of the bottlenecks for more precise Natural Language Processing (NLP) systems. Particularly, the lack of coverage of MWEs in resources can impact negatively on the performance of tasks and applications. For special domains, a significant portion of the vocabulary is composed of MWEs. This paper puts forwards an automatic method for extracting Chinese MWEs with help of statistics and linguistic rules. Seed words of high frequency in special domain are selected to extract candidate strings. By means of statistical measures and linguistic rules, noises in candidate strings are filtered. After filtering, Chinese MWEs are obtained finally. The result of our experiment shows that the method in this paper is efficient to deal with large-scale real texts. The method can extract Chinese MWEs rapidly. Chinese MWEs extracted in this way can be used in many application fields.

Key words: multiword expressions (MWEs); MI; entropy; linguistic rules

(编辑:贾丽红)

(上接第 110 页)

参考文献:

[1] Lahouari G, Ahmed B, Mohammad K I, et al. Digital image watermarking using balanced multiwavelets[J]. IEEE Trans on Signal Processing, 2006, 54(4): 1519-1536.

[2] Yuan H, Zhang X P. Multiscale fragile watermarking based on the gaussian mixture model[J]. IEEE Trans on Image Processing, 2006, 15(10): 3189-3200.

[3] 叶天语,钮心忻,杨义先. 多功能双水印算法[J]. 电子与信息学报,2009, 31(3):546-551.

[4] Schlauweg M, Profrock D, Palfner T, Müller E. Quantization-based semi-fragile public-key watermarking for secure image authentication[C]//Proc of SPIE, California, USA, 2005: 41-51.

[5] 马朝阳,张雪英,李高云. 模糊 RBF 神经网络在音频零水印中的应用[J]. 数学的实践与认识,2010, 40(13): 81-87.

[6] 钟西,唐向宏. 基于音频特征的小波域零水印算法[J]. 杭州电子科技大学学报, 2007, 27(2): 33-36.

[7] 廖琬名,张玉贤等. 基于小波变换的脆弱-鲁棒双重音频水印[J]. 浙江大学学报,2009, 43(4):722-726.

A Digital Audio Dual-Watermarking Scheme Used for Tamper Location

MA Zhao-yang, ZHANG Xue-ying, LI Wen-lu

(College of Information Engineering of TUT, Taiyuan 030024, China)

Abstract: Aiming at the problems of dual watermarking algorithm as poor robustness and inaccurate positioning, a digital audio dual watermarking was presented. A zero-watermarking scheme system based on RBF neural network was constructed as robust watermarking. Since the scheme did not change the original audio data, it had a good transparency. The fragile watermarking embedded watermark image block by double bipolar, it had a good sensitivity. The experiment shows that the watermarks are robust and sensitive to many signal operations. Moreover, it has a good accuracy of tamper location.

Key words: tamper localization; zero-watermarking; RBF; double bipolar

(编辑:贾丽红)