基于网络的英文缩略语全称挖掘

杨华　孙常龙　赵明明　葛运东　姚建民
苏州大学江苏省计算机信息处理重点实验室 苏州 215006
E-mail:20084227065085@suda.edu.cn, jyao@suda.edu.cn

摘　要：本文提出了一种新颖的缩略语全称挖掘方法，分别利用Google和Wikipedia挖掘英文缩略语全称，然后比较两者的正确率和召回率。具体而言，首先利用基于Google的方法挖掘英文缩略语的全称，然后与利用基于Wikipedia的方法得到的全称从正确率和召回率方面做比较。实验结果显示我们的方法比基于网络搜索结果的方法正确率要高，同时可以看出基于Google的方法在正确率和召回率上优越于Wikipedia。

关键词：网络挖掘，Wikipedia，缩略语全称抽取，字母匹配，句法规则

Web-Based Full-Name Extraction For English Abbreviation

Yang Hua, Sun Changlong, Zhao Mingming, Ge yundong, Yao Jianmin
Provincial Key Laboratory of Computer Information Processing Technology
Soochow University, Suzhou 215006
E-mail:20084227065085@suda.edu.cn, jyao@suda.edu.cn

Abstract: This paper presents a novel approach to mine the full-name, namely respectively using Google and Wikipedia for mining the abbreviations full-name, and then we compare the accuracy and coverage. Specifically, we firstly use Google to mining the full-name, then the Wik0ipedia based. The experimental results show that our approach is much better than Search-Result-Based Abbreviation in the degree of accuracy, and at the same time we can see that Google is much better than Wikipedia both in the degree of accuracy and in the degree of cover.

Key words: Web mining, Wikipedia, Abbreviation full-name mining, Letters matching, Syntactic cues

1　引言

缩略语是压缩固定短语或事物名词而形成的语言结构，是相对于没有简化的语言结构而言的，其意义相当于全称。例如："ACL"代 表"the Association for Computational Linguistics"。同时由于人们运用语言的特点，决定了缩略语的实时性和一对多的特点。缩略语的实时性就是缩略语可以随时出现，每天都有大量的缩略语出现在报纸、媒体和网页上。一对多就是一个缩略语有可能有多个全称与之对应。

得到正确的全称对于了解文章意思和信息抽取方面有相当重要的作用[1, 2, 3]，为了解决英语缩略语的实时性和一对多，本文提出了基于网络的搜索方法。首先是基于Google的网络挖掘方法。共分为三个步骤：首先，搜索缩略语全称候选，从Google中搜索前200个摘要，从中提取全称候选。其次，根据我们规定的一些规则，过滤全称候选。最后，排序输出可能的全称。其次是基于Wikipedia的挖掘方法，根据Wikipedia的网页特点，通过人工的统计发现对于不同的缩略语一共可以返回四种不同网页，然后对这四种不同的网页做处理，挖掘全称。最后对这两种方法返回的缩略语的全称做比较。

以下本文从三个方面做分析，找出英文缩略语可能的全称。第二节介绍基于Google的挖掘方法，第三节是关于基于Wikipedia的挖掘方法，第四节对这两种不同方法返回的结果做比较分析。

2 基于Google的缩略语全称挖掘

基于搜索结果的全称抽取方法，共分为三个步骤：首先是基于首字母匹配的缩略语全称候选的提取；然后根据简单的词性标注过滤掉不可能的全称，并取前20个频度较高的候选。

本文在基于搜索结果的全称抽取方法[4]的基础上，提出了一种更完整的方法。基于Google的方法共分三为个步骤，首先，以缩略语做检索词，提取前200个摘要识别出缩略语全称候选，这部分与上述方法相同；然后，根据字母匹配规则、词性匹配规则过滤掉不可能的全称；最后，输出可能的全称。

2.1 缩略语全称候选的识别

首先以缩略语为检索词，检索英文网页提取前200个摘要。然后对返回的摘要进行处理，去掉所有非字母和数字的字符和链接。以EDI为例，对其中一个摘要进行处理得到如图1。然后对所得的摘要进一步处理，识别出全称候选。

EDI Education Development International **EDI** is a leading provider of vocational qualifications and online assessment solutions and an awarding body accredited by the UK regulatory authorities

图1 Google返回结果中包含全称摘要的一个实例

通过分析检索到的摘要，发现所有的摘要都包含缩略语，并且这些摘要只是一两句话或者二三十个词，因此本文把两个缩略语之间的长度作为搜索空间（如图2）。这样做考虑到了摘要中的每一个词，不会漏掉每一个可能是全称的全称候选。本文以缩略语为界把摘要分成若干个搜索空间。如图1中EDI的一个摘要可以按EDI分为两个搜索空间，即下划线部分。

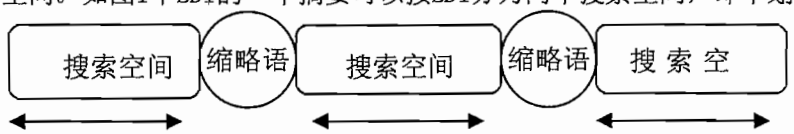

图2 搜索空间

本文对数字开头的缩略语的全称长度的确定与文献[5]中所提到的相同，如公式（1），|A|为缩略语长度，|D|为全称长度。对于只含字母的缩略语，通过对缩略语和其全称做分析后，我们规定所有的缩略语全称最小长度为2，最大长度为|A|×2。

$$\text{首字符为数字时}\begin{cases}\max.|D|=\min\{|A|+5,|A|\times 2\}\\ \min=2\end{cases} \tag{1}$$

本文利用首字母匹配法则进行全称抽取，如果搜索空间中某个词的第一个字母与缩略语的第一个字母匹配，根据规定的缩略语长度提取缩略语全称候选。以EDI的两个搜索空间为例，在第一个搜索空间中找到“Education”的首字母与缩略语的首字母‘E’匹配，然后根据缩略语全称的长度提取EDI的全称为“Education Development”和“Education Development International”。同时统计全称候选出现的次数，由于这两个全称是第一次出现则规定它们的频度为1。在第二个搜索空间中找不到与缩略语首字母匹配的单词，就不会返回全称候选。

分别对200个摘要做上述处理，就会得到一个全称候选集合。并根据全称的频度对所有全称候选进行降序排序。在这一步骤中本文只是对他们进行排序，所有的候选都会被用作下一步的处理。这样做得到会尽可能多的全称，提高了召回率。

2.2 过滤缩略语全称候选

由于缩略语是由某些名词或者短语缩略得到的，最后一个词不可能是介词、助动词、连词、情态动词或代词。因此我们对候选全称的尾词进行词性匹配，过滤掉最后一个词是这些词的候选。

字母匹配：本文把缩略语的每一个字母与全称候选短语中的每个词的第一个字母做匹配。如果匹配将其候选频度乘以2加权，然后排序做第一次输出。在实验中为了观察权重对全称正确率和召回率的影响程度，本文改变频度权重，把字母匹配的候选频度改为乘以20，再一次排序输出。在第一次加权后分析结果，发现已经有正确全称出现，但是第二次改变加权之后，几乎能找到所有的全称。可见字母匹配是影响全称正确率的主要因素。

排序输出：根据以上两个规则限制，输出所有匹配的全称候选，若没有匹配则输出频度排在前十的候选，以EDI为例它的全称集如图3。从图中我们可以看出几乎找到了所有的全称。

F	全称
300	Electronic Data Interchange
80	Electron Drift Instrument
40	Entrepreneurship Development Institute
40	Education Development International
40	Exterior Design Institute
20	Employment amp Disability Institute
20	Employment and Disability Institute
20	Executive Development International
20	Engineering Design and Innovation
20	Economic Development Initiatives
20	Economic Development Initiative
20	Early Development Instrument
20	Educational Design Institute
20	Extrusion Dies Industries

图3　基于Google搜索获得的缩略语EDI的全称

3　基于**Wikipedia**的缩略语全称挖掘

Wikipedia对于普通用户来说它是百科全书，而对于研究者来说它是一个非常有潜力的信息库[6]。本文统计出对于不同的缩略语，Wikipedia的检索系统会固定的返回四种网页。分别是无歧义网页(Disambiguation pages)、检索结果(search result)网页、可连接到无歧义网页的网页和单纯的文章(Articles)页。下面对不同的网页做不同的处理抽取缩略语的全称。

3.1　无歧义网页(Disambiguation pages)

对于一些常见的有多个意思的缩略语Wikipedia检索系统一般会直接链接到无歧义网页。无歧义网页简明扼要的显示了可能的缩略语全称。

一个缩略语可以对应多个全称，这些全称可以是任何方向和领域的，Wikipedia对不同专业领域的全称做了分类。如ACL，它的全称分为三部分：公司和机构名（Companies and organizations），运输业（Transportation），其它意思（Other meanings）。这种完整的分类方法是Google所不具备的，从Google中返回的全称是杂乱无序的。

由于无歧义网页就像目录一样列出了尽可能多的全称，几乎每个全称都是超链接。因此我们做如下处理挖掘全称，对于含有超链接的每一句话，首先返回这句话所包含的超链接部分。其次，本文根据首字母匹配规则过滤掉不可能的全称，同时标记被过滤掉的全称候选所在的行号。最后，根据第二步做的标记，再对标记的行做处理。首先根据句法匹配规则把标记的行用‘，’分为前后两部分。通过统计发现无歧义网页中缩略语的全称一般在每行中

‘，’的前面，因此我们对前半部分用首字母匹配规则进行处理确定全称，如果前半部分首字母匹配并且不是缩略语我们就可以把它作为全称候选输出；若没有找到全称候选再对后半部分做处理，由于返回的行只是很短的一句话并且一定会包含全称候选，我们根据字母匹配规则找到后半部分中与缩略语中每个字母匹配的短语，并将其作为全称候选。

以EDI为例我们得到的全称如下：

Efficient Drivetrains Inc.

Electronic Data Interchange

Extended Destination Index

Edinburgh Airport IATA

Entrepreneurship Development Institute

Economic Development Institute

E-Distribution Initiative

Electrodeionization

Edge Directed Interpolation

Extreme Deep Invader

E.D.I Mean

Education Development International EDI

从结果看出在无歧义网页中虽然返回了可能的全称，但是所包含的全称并一定都是正确的，比如得到的全称“Edinburgh Airport IATA”“Education Development International EDI”包含了错误信息，而“E.D.I. Mean”并不是全称。

3.2 检索结果(search result)网页

有的缩略语在Wikipedia中没有无歧义网页时，当输入这些缩略语时Wikipedia检索系统就会自动整合自己网站上出现过的包含所输入的缩略语的网页，像Google一样返回这些网页上的包含缩略语的一些摘要显示在检索结果网页上。

对于这种网页的处理与对Google摘要处理的算法有相同之处，不同之处只不过不给加权值，只要匹配就输出。因为每个摘要只反映一个信息，每个摘要都是等价的、无重复的。

3.3 可链接到无歧义网页的网页

有些缩略语的全称是经常出现的，所以Wikipedia检索系统会直接链接到这些网页。但是这些缩略语也有自己的无歧义网页，要想得到尽可能多的全称，就需要通过网页中的链接映射到无歧义网页，再抽取全称。以AOL为检索词会返回如图4的网页，可链接到AOL的无歧义网页的“America Online”网页。

首先提取America Online网页的标题，作为其中的一个全称候选，然后根据网页中的链接（如图4）映射到无歧义网页。对无歧义网页处理的方法在上面已经提到。

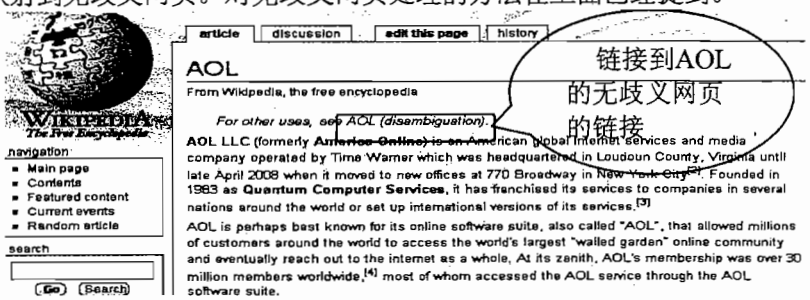

图4 以AOL为查询词在Wikipedia中返回的网页

3.4 文章（Articles）网页

有的缩略语在Wikipedia中只出现过一次，并且在句型is, known as，or等类似的句子中出现过。Wikipedia的检索系统就会直接链接到这些Articles当中。根据Wikipedia中Articles的格式，可以知道Articles的标题或者第一句中必含有它的全称。例如图5中缩略语SLZ返回的网页中，标题就是它的全称，第一句开头也是它的全称。

首先抽取文章网页中的第一段，然后根据句法结构匹配提取全称，最后输出全称。

句法匹配规则：在文章中缩略语和全称是成对出现的，而包含全称的句子有特殊的格式如含有:and 或 or等。我们总结句子结构如下：（1）全称 or 缩略语，（2）缩略语 or全称，（3）全称 is/are 缩略语，（4）全称 known as 缩略语，（5）全称(缩略语)（6）全称，缩略语。在这种情况下由于Wikipedia检索系统只返回文章网页，我们只能得到SLZ的唯一的全称为：“San Lorenzo High School”。

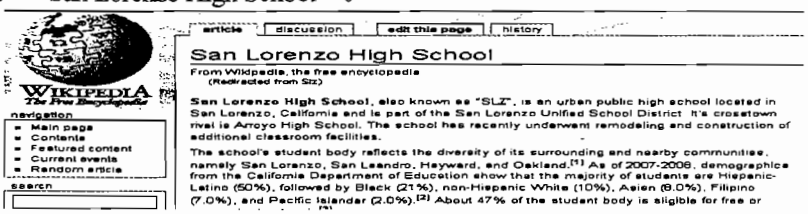

图5 以SLZ为查询词Wikipedia返回的Articles网页

4 实验结果分析

测试集：87个非常见缩略语，794个计算机专业缩略语。

本文所用的方法是在基于网络搜索结果[4]的基础上实现的一种正确率更高的方法，不同于基于混合文本的全称抽取，所以不需要找缩略语与全称的关系[7,8,9]。

基于搜索结果的方法[4]主要是要求返回所需要的一个正确的全称，表1中的94.5%是top-5中含有正确全称的概率，同时在他们的实验中并没有给出召回率。

在基于Google的方法中，有862个缩略语返回全称，其中有844个缩略语能返回正确全称。由表1可以看出基于Google的召回率为96.79%远远超过了基于Wikipedia和基于搜索结果的方法，这说明Google的召回率高于Wikipedia。其中一部分缩略语在Google中没有找到正确的全称，如“KUKB”，“QHRT”，“EJMG”等。同时还有一部分并没有返回全称，但并不是没有全称，当把它们人工输入Google后可以明显的找到全称，究其原因可能是Google限制了我们的查询。

在基于Wikipedia的方法中，由于Wikipedia并不完善召回率只有88.26%，Wikipedia的召回率低与Google。Wikipedia的正确率也不是很高，其原因是在Wikipedia中并不是所有的缩略语都会返回无歧义网页并且有规律的显示它们的可能的全称。

由于一个缩略语可对应多个全称，我们随机的从测试语料中抽取50个缩略语，对不同方法返回的结果进行人工分析。基于Wikipedia的方法返回的全称更准确，可以说返回的全称就是正确的，这是Wikipedia的网页结构化的结果，但是有的缩略语只返回一个全称或者返回了不正确的全称，这是由于此缩略语在Wikipedia中并没有专门的article网页或无歧义网页，同时这是影响基于Wikipedia方法的关键。基于Google的方法返回的结果不是很准确，一般常见的全称会排在前面，同时会返回全称的单数和复数形式，但是我们只要其中之一即可。通过分析同一个缩略语在基于Google的方法和基于Wikipedia的方法中返回的结果，我们发现前者的召回率和正确率比较高，但是后者对于每个有全称的缩略语返回的全称更准确。

$$召回率=\frac{返回包含正确全称的缩略语个数}{用做测试的缩略语个数}\times 100\% \quad (2)$$

$$正确率=\frac{返回包含正确全称的缩略语个数}{能返回全称的缩略语个数}\times 100\% \quad (3)$$

表1　用三种方法挖掘英文缩略语全称的实验结果

方法	召回率	正确率
基于搜索结果	—	94.5%
基于Google	96.79%	97.91%
基于Wikipedia	88.26%	93.01%

5　结论

本文分别基于Google和Wikipedia挖掘缩略语全称。实验结果显示基于Google的全称抽取效果优于基于Wikipedia的抽取。通过分析也看到Wikipedia的优越之处，它虽然召回率低，但是由于Wikipedia的网页特点对于单个缩略语它所返回的全称集比Google返回的质量更高。

在处理无歧义网页时，本文只用了网页上的文本，在将来的工作中，在处理文本的同时还可以考虑到超链接，通过超链接链接到网页进行处理，就会得到召回率更高的全称集。在以后Wikipedia不断完善的过程中，Wikipedia不仅会在全称的挖掘方面有突破，同时在查询翻译等方面也会有新的突破。

参　考　文　献

[1] Byrd, Roy, Yael Ravin, and John Prager. Lexical Assistance at the Information-Retrieval User Interface. Research Report, RC19484, IBM T.J. Watson Research Center, 1994.

[2] Kugimiya, Shuzo, Yoji Fukumochi, Ichiko Sata, Tokyuki Hirai, and Hitoshi Suzuki. Machine Translation apparatus having a process function for proper nouns with acronyms. US Patent 5161105, 1992.

[3] Maynard, Diana and Sophia Anaiadou. Term Extraction using a Similarity-based Approach. In Recent Advances in Computational Terminology, John Benjamins, 1999.

[4] Wen-Hsiang Lu, Jiun-Hung Lin, and Yao-Sheng Chang. Improving Translation of Queries with Infrequent Unknown Abbreviations and Proper Names Computational Linguistics and Chinese Language Processing. The Association for Computational. 2008. 7-9.

[5] Park, Y. and R. J. Byrd, Hybrid text mining for finding abbreviations and their definitions. In Proc. of Conference on Empirical Methods in Natural Language Processing. 2001. 2-4.

[6] Olena Medelyan, Catherine Legg. David Milne and Ian H. Witten Mining Meaning from Wikipedia. Department of Computer Science, The University of Waikato Private Bag 3105 Hamilton, New Zealand. 2008. 1-17.

[7] Kleinberg, Jon. Authoritative sources in a hyperlinked environment, In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, May 1997.

[8] Larson, R. Bibliometrics of the World-Wide Web: An exploratory analysis of the intellectual structure of cyberspace, Technical Report, School of Information Management and Systems, University of California, Berkeley, 1966. http://sherlock.sims.berkeley.edu/ docs/asis96/asis96.html.

[9] Sundaresan, Neel and Jeonghee Yi. Mining the Web for Relations, In The Ninth International World Wide Web Conference, 2000. http://www9.org/w9cdrom/363/363.html.