基于用户查询日志和锚文字的汉语缩略语识别*

谢丽星[1] 孙茂松[1] 佟子健[2] 王灿辉[2]

[1]清华大学计算机科学与技术系 北京 100084 [2]搜狐互联网信息服务有限公司研发中心 北京 100084

E-mail: lavender087@gmail.com sunmaosong@gmail.com tongzijian@sohu-rd.com canhuiwang@sohu-rd.com

摘 要: 缩略语是自然语言的常见现象之一，其相关研究是中文信息处理领域的重要研究课题。本文针对缩略语的自动识别问题，采用用户查询日志和锚文字文件，运用 “同网站主题相关性”（即对应的 url 指向同一网站的查询词较为相关）的思想进行初步的缩略语、源短语对的抽取，然后采用一系列过滤规则，结合分词按照缩略语的形成方式进行分类，最后调用搜索引擎采用多策略来识别缩略语、源短语对。相比前人研究，我们的实验在规模和准确率上都有提升，其中用户查询日志的准确率为 68.33%，锚文字的准确率为 92.66%。

关键词: 用户查询日志，锚文字，缩略语的分类，搜索引擎

Identification of Chinese Abbreviations Using Query Log and Anchor Text

Xie Lixing[1] Sun Maosong[1] Tong Zijian[2] Wang Canhui[2]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, 100084

[2]Sohu Inc. R&D center, Beijing 100084

E-mail: lavender087@gmail.com sunmaosong@gmail.com tongzijian@sohu-rd.com canhuiwang@sohu-rd.com

Abstract: Abbreviation, a common phenomenon in natural language, has drawn a large body of research interest in the natural language processing community. In this paper, we propose a novel method introducing query log and anchor text to gain a new vision of identifying Chinese abbreviations. First, a heuristic algorithm is applied to extract suspect abbreviation-root pairs. Second, we use filtering rules concluded from actual observations and Chinese word segmentation techniques to classify the suspect pairs into candidate patterns. Finally, search engine is adopted to further validate the abbreviation-root pairs using multi-strategy methodology. Compared to previous studies, our experience shows improvements on scale and precision, with a result of 68.33% in query log and 92.66% in anchor text.

Keywords: query log, anchor text, classification of abbreviation, search engine

1 概述

自然语言的经济性原则导致了缩略语的出现，如“北京大学”简称“北大”。缩略语是未登录词的主要来源之一，应用广泛，据 Chang 和 Lai (2004) 的研究表明，新闻标题中大约有 20%的句子会使用缩略语。因此，缩略语的相关研究是自然语言处理的重要课题。它能提高自动分词和标注的准确率，还有助于提升机器翻译、问答系统、信息检索等系统的性能。

缩略语指语言中由固定说法经过压缩，省略或统括而形成的词语。各种语言的缩略语构成方式各有不同，英文缩略语主要是通过单词首字母缩写或者截断形成（例如 International Business Machines 简称为 IBM）。汉语缩略语的构成方式较为复杂，主要有四种：

（1）语素构成：只抽取原词语各部分语素来替代原词语。如：电视大学——电大；

*本项目承清华——搜狐搜索技术联合实验室项目的资助。

（2）中心词构成：抽取原词语中的核心成份。如：中国工农红军——红军

（3）混合法构成：语素、词和音节混合。如：广播体操——广播操；

（4）合并法构成：将原词语通过合并的手段简缩形成。如：包退、包换、包修——三包。

目前汉语缩略语词典版本较少，主要由专家根据个人知识编写，覆盖性较差，更新速度较慢，因此在大规模的web文本下自动识别缩略语的研究更显得迫切和重要。

汉语缩略语识别方面研究目前已取得一些进展。本文仅讨论缩略语的自动抽取，即自动提取缩略语、源短语对。在前人的研究中，缩略语自动抽取的大致流程如下：首先选定语料库（其中有的采用缩略语、源短语的平行语料库，有的采用经过人为分词和词性标注的语料库），然后提取候选缩略语、源短语集，再利用缩略语、源短语的长度等字对齐规则进行搜索匹配，或者采用机器学习的方法进行选择匹配对，最后输出正确的缩略语、源短语对。

由于前人研究所使用的语料库多是非真实环境，规模较小，时效性欠佳，有的还需要人工干预，实验结果准确率较低。针对这些问题，我们选用用户查询日志和锚文字作为实验数据，实现了一个缩略语自动抽取系统，并调用搜索引擎进行验证筛选。实验结果证明，我们的做法有效可行，相比于前人研究，从规模和准确率方面都得到了性能的提升。

2 相关工作

随着Web信息的日益膨胀，汉语缩略语在信息处理领域日益重要，一系列研究就此展开。

有些实验是采用缩略语、源短语的平行语料库作为实验数据的，如：Chang和Lai（2004）使用人工标注的源短语、缩略语的平行语料库作为训练数据（共1235对候选对），然后利用HMM来提取缩略语、源短语对。Chang和Teng(2006)，扩展了以上工作，采用ASWSC-2001作为实验数据（共94篇文档，大小为1,3248KB），提出了基于HMM的概率恢复模型(SCR)，用于将缩略语扩展为源短语，训练集准确率为62%，测试集准确率为50%。但他们仅考虑了单字缩略语与单个词源短语的对应关系，较日常接触的缩略语有一定差距。Lee（2005）列举了很多实例，来说明是源短语是如何构成缩略语的，并制定了一些规则将缩略语扩展成源短语，但没有给出定量结果。

有些实验摆脱了平行语料库，但需要人工干预，如：崔世起(2005)利用生语料，必要时加入人工干预，自动提取缩略语和源短语对。实验得到的候选缩略语为551个，源短语为28,6378个，准确率为51.4%。武子英(2006)利用上下文语义信息，基于余弦相似度自动抽取汉语缩略语。实验使用1998年《人民日报》的20万标注语料库，准确率为74.1%。尽管准确率较高，但是实验规模较小，且需要依赖新词发现和词性标注，初步提取候选集时还需加入人工干预，较为复杂。

有些实验不需要任何标注数据，如：Li(2008)根据缩略语与源短语的共现现象，使用字对齐规则进行自动提取（仅处理单一类型的缩略语），最终得到51K对候选对，准确率为51.3%。

针对上述方法选用的语料库时效性较差，规模较小，需要人工干预，且只解决单一类型（语素构成）的缩略语等缺陷，我们使用用户查询日志和锚文字，自动抽取候选缩略语、源短语对，然后对其进行分类，最后调用搜索引擎验证。我们的方法不需要任何标注数据和人工干预。

据我们所知，目前还没有人在大规模的用户日志和锚文字方面做过缩略语自动识别的研究。

3 算法设计

3.1 使用用户查询日志和锚文字的可行性分析

针对前人研究，要保证时效性，大规模，我们想到两类数据：锚文字与用户查询日志，它

们都包含单词与内容之间的语义关联。观察两类数据（图 1、2），发现：（北大、北京大学）是缩略语、源短语对，而用户日志中的（北大，北京大学研究生）是噪音。用户日志中的正确对的 url 的域名是相同的。由此推测，两类数据中可能包含很多缩略语、源短语对。

查询词	url
北大	lib.pku.edu.cn/portal/index.jsp
北京大学	www.pku.edu.cn
北京大学研究生	grs.pku.edu.cn

图1 用户日志部分数据

锚文字	北京大学；北大
锚文字个数	3；1

图2 Anchor text部分数据

我们最终选取搜狗实验室提供的用户查询日志（query log，大小 1.56G）和锚文字（anchor text，大小 43.1M）作为实验数据。本文中所用数据，如无特殊说明，均为搜狗实验室提供。

3.2 抽取候选缩略语、源短语对

实验数据中存在非中文字符，数字等噪音。因此要对实验数据进行预处理，过滤噪音词汇。

锚文字文件中，指向同一个网页的锚文字已被放在一起，无需处理；针对用户查询日志，我们提出了“同网站主题相关性”的概念，即对应的 url 指向同一网站的查询词可能较为相关。我们首先对 url 进行“域名的部分倒排”（如 pku.edu.cn 变为 cn.edu. pku，数字域名不做处理），然后对倒排后的 url 按照字母序排序。这样同一个网站对应的查询词都集中在某一个地方了。我们将这类查询词基于以下规则进行缩略语、源短语的初步提取：

（1）缩略语的字长比源短语短，且缩略语中的每个字均在源短语中出现；

（2）缩略语字长不超过 7，且不含有 1 个以上二字词

（3）源短语/缩略语的长度比值在[1.33,4]之间，且二者最长公共子串的字长不超过 4

初步提取完后，我们研究了地名在缩略语中的影响，发现：（市二中，沈阳市二中），该候选对中地名是噪音；（北大，北京大学）是正确候选对，因此，我们需用地名词表选择性过滤。

此外，人名也会对实验结果造成影响，需要过滤。

3.3 依据缩略语简缩方式分类

初步抽取完候选对，我们发现候选对中的缩略语的形成方式较多（图 3），准确率也不一样。比如语素构成的缩略语准确率较高（图 4），子串方式较差（图 5），因此需要进行分类。

北大	北京大学
车管所	车辆管理所
北科大	北京科大
北科大	北京科技大学
高检	最高人民检察院
城建	北京城建
档案	档案馆

图3 抽取出的部分候选对

北大	北京大学
车管所	车辆管理所
北科大	北京科大
北科大	北京科技大学
高检	最高人民检察院

图4 部分候选对中正确部分

档案	档案馆
城建	北京城建

图5 部分候选对中错误部分

我们采用 ICTCLASS 分词词表，针对概述中提到的缩略语构成方式，主要选取以下三类：

(1)singleChar 类:语素构成，源短语中每词对应缩略语中的一字，如：北京大学——北大

(2)mixtureNoMissing 类：混合法构成，但是没有缺少任何词，如：广播体操——广播操

(3)missingMiddle 类：覆盖首尾，中间缺词，剩余词每词一字，如：中华人民共和国——中国

注：这 3 类覆盖度较广，对齐较好，准确率也较高，故本文暂时只处理这 3 类。前人研究至多只处理 1、2 两类（语素构成和混合法构成），因此我们比前人处理的类型要多。

3.4 调用搜索引擎进行验证

分类后，我们发现候选对中既有正确的也有错误的。正确的如：（北大，北京大学），错误的如（婚后，离婚以后）。这种错误通常是由于语义不一致造成的。因此尽管对齐较好，仍需筛选。基于时效性的考虑，我们采用搜索引擎作为验证工具，主要从两方面考虑：内容方面，缩略语和源短语语义相同，因此检索它们得到的结果大致一样，表现为摘要标题的 url 链接相似；共现现象方面，二者存在共现现象，如标题中使用缩略语，正文中使用源短语等，表现为搜索引擎得到的结果中，二者可能出现在同一段摘要中。

综上所述，共提出以下两大类方法（以下方法均选取搜索结果的前 20 项摘要）：

（1）基于 url 的相似度比较

分别检索缩略语和源短语，比较 url 第一级，如相同，相似计数加 1

（2）基于内容中缩略语与源短语共现方式的启发式搜索（两种方式）

（a）只将缩略语作为关键词送入搜索引擎查询，如“北大”

统计每项摘要中缩略语与源短语同现次数、缩略语出现次数、源短语出现次数，如下：

$$P(full \mid abbre) = \frac{\text{Count[abbre,full]+0.5f*fullCount}}{\text{abbreCount}} * \frac{20}{size}$$

(b)将缩略语和源短语同时作为关键词送入搜索引擎查询，如“北大+北京大学”

统计每项摘要中缩略语与源短语同现次数（Count[abbre, full]），如下计算权重：

$$P(abbre, full) = \text{Count[abbre,full]} * \frac{20}{size}$$

4 实验结果及相关分析

4.1 候选对检索前后结果

由于候选对对齐方式较好，为了尽可能多地保留可能正确的候选对，此处我们选取 url 相似计数大于 0，启发式搜索权重大于 0.0，给出两个数据集的相关统计，如表 1、 2、3。

数据来源	分类结果	url 比较结果	仅检索缩略语	同时检索缩略语源短语
Query log	7412 对	4794 对	1190 对	**5354 对**
Anchor text	507 对	**457 对**	410 对	410 对

表 1 singleChar 类型的相关统计

数据来源	分类结果	url 比较结果	仅检索缩略语	同时检索缩略语源短语
queryLog	4300 对	**3652 对**	1406 对	2508 对
Anchor text	341 对	**310 对**	226 对	203 对

表 2 mixtureNoMissing 类型的相关统计

数据来源	分类结果	url 比较结果	仅检索缩略语	同时检索缩略语源短语
queryLog	1, 2734 对	**6399 对**	551 对	5328 对
Anchor text	302 对	**259 对**	114 对	134 对

表 3 missingMiddle 类型的相关统计

结果分析：

“验证”前，对于用户日志，共得到 2,4446 对候选对，对于锚文字，共得到 1150 对候选对，前者大约为后者的 20 倍；“验证后”，对于用户日志，最多保留(url 查询结果) 1,4845 对，锚文字是 1026 对，前者是后者的 15 倍。

4.2 分词对于分类结果的影响分析

这里我们简单探讨了一下分词对于分类结果的影响，总结出两点：

（1）词表有限，对有些词识别不出来，因此分为单字词，从而导致分类错误

如（沧月，沧海月明），每词一字，本该划为 singleChar 类，但由于词表中有“沧海”没有“月明”，源短语的分词结果为沧海|月|明，缺少“明”，被误判为 missingStartOrEnd 类。

（2）由于采用前向最大匹配分词算法，会造成切分错误，从而导致分类错误

如（丽江地图，丽江旅游地图），缺失“旅游”一词，本属于 missingMiddle 类，但由于是前向最大分词，源短语的分词结果为丽|江|旅游地|图，被误判为每词一字，为 singleChar 类。

4.3 评测结果

这里我们给出两个数据集上的准确率统计（图 6，7）。对于 query log，针对每种方式每种类型，随机选取 200 对，统计准确率；对于 Anchor text，统计所有候选对的准确率。

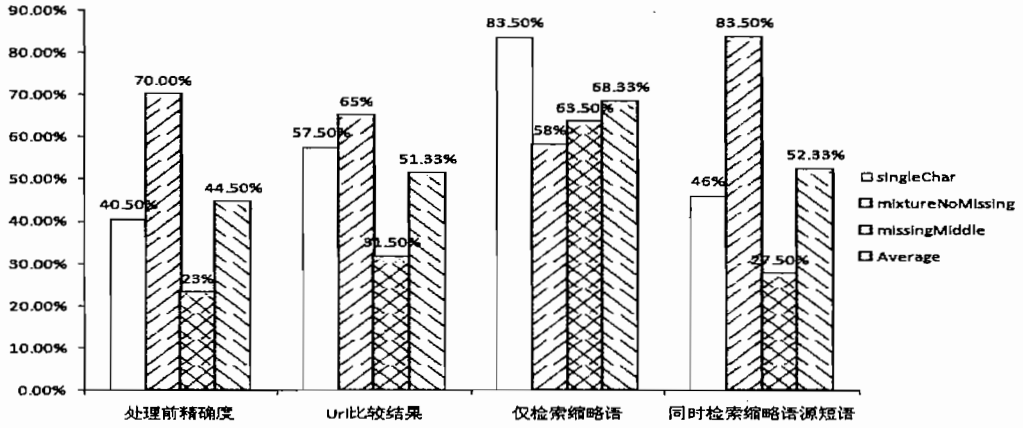

图 6 query log 的准确率统计

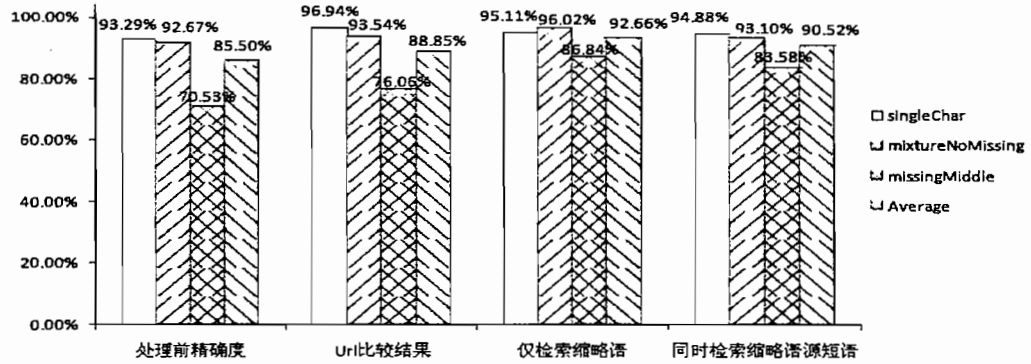

图 7 anchor text 的准确率统计

结果分析：

(1)对于用户查询日志和锚文字结果的比较：

从提取出的候选对来看，用户日志的数量要远大于锚文字文件（15倍左右）；

从平均准确率来看，锚文字（85%以上）要远高于用户日志（不超过69%），有两点解释：

(a)用户日志存在较多人为错误和缺失，人名及主题词的大量存在也严重影响准确率；

(b)锚文字文件多半摘自网页正文，可信度高，错误少，垃圾少，人名和主题词也较少。

(2)三种类型的缩略语准确率比较：

(a)用户日志上：mixtureNomissing类最好，singleChar类其次，missingMiddle类最差。

(b)锚文字上：singleChar类最好，mixtureNomissing类其次，missingMiddle类最差。

(3)对于调用搜索引擎的三种方法的比较：

采用搜索引擎验证分类结果会丢失一些匹配对，但能提高准确率。这在两个数据集上都体现了一致性。但由于阈值大小及抽样数目的选取，会对准确率造成影响，从而造成极个别的反差现象（query log中的mixtureNoMissing类结果），此处不去深究。就这三种方法的效果来看：

(a)从保留下来的缩略语源短语对数来看：

url比较策略 > 同时检索缩略语和源短语 > 仅检索缩略语；

(b)从准确率来看，这三种方法一般情况下都比求证前的精确度要高，其中：

仅检索缩略语 > 同时检索缩略语和源短语 > url比较策略。

5 结论及未来工作

本文针对目前缩略语研究中需要使用平行语料库等现状，采用时效性较好的大规模用户查询日志和锚文字语料库，自动抽取汉语缩略语、源短语对，并对其进行类别划分，调用搜索引擎采用多策略来进行验证，与前人研究相比（准确率通常只在50%左右），大大提高了精确度（图6、7中显示query log最好为68.33%，锚文字最好为92.66%）。后续，仍有一些问题有待我们研究：

（1）在更大规模的用户查询日志和锚文字上进行研究；

（2）考虑加入常用词过滤、引入点击次数信息来解决缩略语“一对多”现象（正确的可以保留）；

（3）减轻分词对于错误类别的影响，争取解决更多类型的缩略语的自动提取。

参 考 文 献

[1] Jing-Shin Chang and Yu-Tso Lai. 2004. A preliminary study on probabilistic models for Chinese abbreviations. *In Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing, pages 9-16.*

[2] H.W.D Lee. 2005. A study of automatic expansion of Chinese abbreviations. MA Thesis, The University of Hong Kong.

[3] Jing-Shin Chang and Wei-Lun Teng. 2006. Mining Atomic Chinese Abbreviation Pairs: A Probabilistic Model for Single Character Word Recovery. *In Proceedings of the 5rd SIGHAN Workshop on Chinese Language Processing, pages 17-24.*

[4] Zhifei Li and David Yarowsky.2008. Unsupervised Translation Induction for Chinese Abbreviations using Monolingual Corpora. *In Proceedings of ACL-08: HLT, pages 425–433.*

[5]崔世起, 刘群, 林守勋等. 中文缩略语自动抽取初探[C]. 全国第八届计算语言学联合学术会议(*JSCL-2005*).

[6]支流, 段慧明, 朱学锋等. 中文缩略语知识库建设[C]. 第三届学生计算语言学研讨会论文集, 2006.

[7]武子英, 郑家恒. 现代汉语缩略语自动识别的方法研究[J]. 计算机工程与设计, *2007, (16).*