越南语文献中字母缩略语自动提取研究

张海云[1] 张超静[2] 毕玉德[3]

解放军外国语学院研究生系[1] 解放军外国语学院基础部[2] 解放军外国语学院亚非语系[3]洛阳 471003

E-mail: willians.student@sina.com; zhangchaojing@tom.com; biyude@gmail.com

摘 要：缩略语的使用顺应了语言的经济原则，但是也造成了越南语自然语言处理中的诸多障碍。本文在分析大量越南语字母缩略语使用特点的基础上，根据越南语字母缩略语词法、句法等特点，采用基于规则匹配的方法进行自动提取，最后生成一个缩略语及全称对应表。实验结果表明，本系统对已定义字母缩略语及其全称自动提取的准确率达到98.04%，召回率达到78.13%，F值达到86.96%。

关键词：越南语，字母缩略语，自动提取，规则

The Study of Automatic Extraction Abbreviations and their definitions in Vietnamese Literature

Zhang Haiyun[1] Zhang Chaojing[2] Bi Yude[3]

Department of Graduate Studies, PLA University of Foreign Languages[1], Department of Fundamental Studies, PLA University of Foreign Languages[2], Department of Asian and African Languages, PLA University of Foreign Languages[3], Luoyang 471003

E-mail: willians.student@sina.com; zhangchaojing@tom.com; biyude@gmail.com

Abstract: The use of abbreviations is compatible with the principle of economy in linguistics, but also leads to many obstacles in the natural language processing of Vietnamese. Based on the analysis of the usage features of a large number of Vietnamese abbreviations, this essay used a set of pattern-matching rules to automatically extract abbreviations and their definitions, and finally developed an abbreviation-definition pairing list according to the accidence and syntax features of Vietnamese abbreviations. The result of the experiment shows that the system achieved 98.04% precision, 78.13% recall and 86.96% F-measure for the defined abbreviations.

Keywords: Vietnamese, abbreviation, automatic extraction, rules

1. 引言

缩略语是各种语言中共有的现象。当今社会信息量迅速膨胀，为了提高表达效率，人们常常将一些结构复杂、语句较长的常用词或者短语进行缩减，形成“缩略语”。例如，将“Mặt trận Tổ quốc Việt Nam（越南祖国阵线）”缩写为“MTTQVN”，压缩了2/3的字母。缩略语的使用顺应了人们生活、工作和交往中的经济原则，体现出人类使用语言时力求经济、简便的自然心理趋势（李新，2004）。目前，在政治、经济、军事、医学等各领域中缩略语已经得到了广泛地使用。随着科学技术的发展，人际交流日益频繁，生活节奏加快，缩略语的使用呈现出快速增加的趋势。

然而，缩略语的广泛使用以及新缩略语的不断涌现形成了自然语言处理中未登录词的主要词源，导致了越南语自然语言处理在分词、词性标注、词义消歧、命名实体识别和指代消解等诸多问题上的严重障碍。同时，由于全称与缩略语在形式表现上的不同，对信息检索、关键词抽取、

机器翻译、问答系统等应用也会造成影响（Sun 等，2008）。如，以“Đảng Cộng sản Việt Nam（越南共产党）”作为检索条目，对仅包含该词两种缩略形式“ĐCSVN”或“DCSVN”的文本就可能产生漏检，反之亦然。可见，缩略语自动提取研究是自然语言处理中一项重要的基础性工作。

本文基于规则匹配的方法，利用越南语缩略语词法、句法等方面的特征对未经切分标注的越南语生语料进行自动提取，实现缩略语和全称之间的自动匹配，最后生成一部缩略语词典。

2. 缩略语自动提取的研究现状

从所掌握的资料来看，针对英语和汉语的缩略语自动识别和提取研究已经取得了一定的进展。英语主要针对生物医学领域中的缩略语自动提取进行了大量的研究，取得了较好的效果。生物医学知识日益增长，尤其是基因、细胞和药品等缩略语名称的大量使用给研究人员带来了许多挑战。借助缩略语自动提取系统，研究人员可以方便地动态识别生物医学文献中的缩略语。汉语缩略语自动提取研究已经进行了一定程度的实验和工程化，识别的准确率和召回率都达到了一定的高度，建立了高质量的缩略语知识库，但缩略语的自动识别还存在一些问题，离实用尚有距离（丁俊苗，2008）。而越南语缩略语的自动提取由于缩略语本身的复杂性和使用的随意性，目前尚无学者专门针对越南语缩略语的自动提取展开研究。

总的来看，缩略语自动提取方法主要分为以下四类（Xu 等，2008）：

1、基于统计的方法。该方法依靠共现频率等统计特征对大型语料库进行统计分析，从而统计出可能的缩略语和全称对。其特点是实现方法较简单，通过统计量化的方式可以发现一些不规则的缩略情况，准确率较高。但是很难发现出现频率较低的缩略语，而且系统开销较大，耗费时间较长。

2、基于机器学习的方法。该方法由学习模块和生成模块两部分组成，首先通过学习模块训练缩略语的提取特征，然后生成模块利用这些特征从文本中提取符合条件的缩略语和全称对。但这需要耗费较多的时间进行学习，系统开销较大，复杂度较高。

3、基于规则的方法，也叫基于模板的方法。该方法根据缩略语的词形特点、缩略方式以及句法、语义等特征，总结出适合绝大多数缩略语及其全称的识别规则或模板，通过这些规则或模板识别出缩略语及其全称对。利用语言专家总结的自动提取知识进行缩略语识别准确率较高，系统开销小，运行速度快。但规则的建立主要依靠人工内省的方法，规则的完整性决定了系统的效果。

4、基于文本对齐的方法。该方法是建立在一个假设之上的：全称存在于缩略语的附近，并且按照缩略语中各字母的顺序包含缩略语中的所有或者几乎所有的字母。因此它总是试图找出缩略语及其全称之间的最佳匹配。这种方法实现方法非常简单，无需进行模板训练或总结提取规则，运行效率高，但目前基于文本对齐的方法尚无法提取出不规则的缩略语。

3. 现代越南语字母缩略语的定义及其构成方式

缩略语是语言符号的进一步符号化。对于缩略语的概念，不同学者有不同的提法和界定，各种语言的缩略语也都有其自身特征。由于目前尚无专门研究越南语缩略语的文献，而越南语同英语、法语等语言一样都采用拉丁字母文字，其缩略语的缩合方式在许多方面也借鉴了拉丁字母

词语的特点，因此我们主要参考英语缩略语的概念进行界定。

英国戴维·克里斯特尔编著的《现代语言学词典》将缩略语（abbreviation，也叫缩写，缩写词）分为四类：（一）首字母缩写词（initialism, alphabetisms），如 COD - Cash on Delivery（货到付款）；（二）缩略词（acronyms），如 NATO - North Atlantic Treaty Organization（北大西洋公约组织）；（三）截短词（clipped），如 ad - advertisement（广告）；（四）截搭词（blends），如 sitcom - situation comedy（情景喜剧）。本文所讨论的是“字母缩略语”，即由一个词组中各主要词的首字母、次字母、尾字母以及一些标点符号缩合而成的词，包括戴维·克里斯特尔所划分的首字母缩写词和缩略词。字母缩略语对应的完整形式我们称其为“全称”。对于 WTO、GDP 等外来缩略语，虽然它们在越南语中能够直接使用，并且按照越南语的语音有了新的读音方式，但它们作为国际通用缩略语并不反映越南语语言本身的特点，本文不认为它们是越南语字母缩略语。

我们对越南《人民军队报》2009 年 1-3 月份的文本内容中的字母缩略语进行了分析。从语言学的研究角度来看，越南语字母缩略语在词形上存在以下四种类型：

编号	词形分类	越南语缩略语（全称）	中文翻译
1	大写字母	QUTƯ (Quân uỷ Trung ương)	中央军委
2	大写字母+小写字母	TTg (Thủ tướng)	总理
3	大写字母+符号	QP-AN (quốc phòng - an ninh)	国防与安全
4	大写字母+标点	CTĐ, CTCT (công tác đảng, công tác chính trị)	党的工作，政治工作

表 1

越南语字母缩略语在缩略方式上存在以下五种模式：

编号	缩略方式	越南语缩略语（全称）	中文翻译
1	首字母缩合	BĐCL (bộ đội chủ lực)	主力部队
2	首字母+次字母缩合	ThS (thạc sĩ)	硕士
3	首字母+尾字母缩合	TTg (Thủ tướng)	总理
4	含有符号的首字母缩合	QP-AN (quốc phòng - an ninh)	国防与安全
5	含有介词的缩合（介词可用其他符号代替甚至省略）	QP-KT (quốc phòng với kinh tế) BGD&ĐT (Bộ Giáo dục và Đào tạo) DQTV (dân quân, tự vệ)	国防与经济 教育培训部 民兵，自卫

表 2

另外，越南语字母缩略语还存在字母变异的问题。为了书写的方便，越南人在缩写时根据越南语自身的特点，有可能把字母“Ư”缩写为“U”或“W”，把字母“Đ”缩写为“Đ”或“D”，把“PH”缩写为“P”或“F”。例如，“Quân uỷ Trung ương（中央军委）”有可能缩写为“QUTƯ”，也有可能缩写为“QUTW”；“Đảng Cộng sản Việt Nam（越南共产党）”有可能缩写为“ĐCSVN”，也有可能缩写为“DCSVN”；“Phó chủ nhiệm（副主任）”有可能缩写为“PCN”，也有可能缩写为“FCN”。因此，在制定提取规则是需要充分考虑这些变异情况。

4. 基于规则匹配的越南语字母缩略语自动提取方法

许多文献都规定了缩略语使用的方法，字母缩略语在第一次使用时常常需要按照“全称（缩

略语)”或“缩略语（全称）”的格式进行定义。而且，人们在进行词语缩略时都需要遵循一定的缩合习惯，比如首字母缩合、首字母+次字母缩合、首字母+尾字母缩合等等。基于以上分析，我们认为基于规则匹配的方法是切实可行的。我们设计的越南语缩略语自动提取系统主要流程如下：

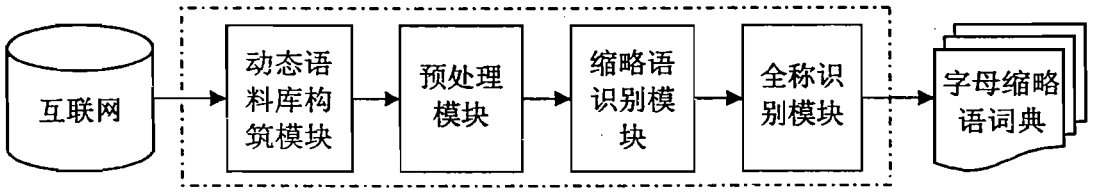

图1

4.1 动态语料库构筑模块

动态语料库构筑模块的主要任务是获取较大规模的文本语料。由于我们的目标是要在没有缩略语词表的情况下，自动提取缩略语及其全称对，构建一部越南语字母缩略语词典，所以需要利用足够大规模的生语料进行提取。显然通过手工的方法下载分析成千上万的网页是不可能的，因此通过动态语料库构筑模块自动获取较大规模的文本语料就显得尤为重要。

从互联网上直接下载的语料是半结构化的，往往结构复杂，含有大量 HTML 语言标记，语料信息湮没在芜杂的网页标记中。我们主要通过 HTML 标记的起止锚点来提取有效信息。经过分析，每个网站都有自己的一套HTML 标记的起止锚点模板，比如：用“<content>”和“</content>”分别表示正文的开始和结束。利用这些起止标记锚点，我们可以解析出大量网页的语料信息。

解析得到的数据还需要进行数据清洗和格式化。抽取的信息中往往含有大量的网页特效标记和超链接信息标记，而不仅仅是简单地抽取起止标记中间的内容。有的网页中为了网页效果将句子或段落打断分行显示，因此需要特别处理来去除标记和空格，连接分开了的段落，进行语料信息的格式化，从而生成实验所需的庞大语料库。

4.2 预处理模块

预处理模块的主要任务是对语料库中的语料进行初步的处理。首先将输入的文本自动去除仅包含数字或者是数字加上百分号（%）、连字号（-）的括号，然后将文本切分为句子，提取出包含括号的句子。如果一个句子中含有多个括号，则以右括号“)”为标记，将句子切分为若干个子句。例如，将句子“Ngày 8/6/2007 Ban Biên tập Tạp chí Quốc phòng toàn dân (QPTD) đã phối hợp với Bộ Tư lệnh Quân chủng Hải quân (QCHQ) tổ chức cuộc trao đổi ý kiến.”切分为“Ngày 8/6/2007 Ban Biên tập Tạp chí Quốc phòng toàn dân (QPTD)”、“đã phối hợp với Bộ Tư lệnh Quân chủng Hải quân (QCHQ)”和“tổ chức cuộc trao đổi ý kiến.”三个子句。

4.3 缩略语识别模块

我们主要从两个方面进行越南语字母缩略语的识别：

1）字母大小写。通常情况下，越南语词语只有作为句首或专有名词时，才会将词语首字母大写。而越南语字母缩略语在进行缩合时，通常是将各词形的首字母改为大写形式后缩合到一起。

因此一个越南语缩略语词形中通常包含两个及以上大写字母。根据这一规则，在进行文本检索时，除Ⅲ、Ⅳ、Ⅻ等罗马字母外，如果一个词形中含两个或两个以上大写字母则判定为字母缩略语。如："QUTƯ（中央军委）"、"TTg（总理）"、"ThS（硕士）"、"QP-AN（国防与安全）"、"BGD&ĐT（教育培训部）"等等，每个词形中都包括两个或两个以上的大写字母。

2）元音与辅音的搭配情况。越南语词形类似于汉语拼音，由"首辅音+韵+声调"三部分组成，其结构图如下：

<table>
<tr><td colspan="4">声　调</td></tr>
<tr><td rowspan="2">首辅音</td><td colspan="3">韵</td></tr>
<tr><td>介音</td><td>主要元音</td><td>韵尾</td></tr>
</table>

表3

例如，词形"nguyễn（阮）"的首辅音为"ng"，韵为"uyễn"，其中"u"为介音，"yê"为主要元音，"n"为韵尾，"ê"上的"~"是声调。主要元音是一个词形中不可缺少的成分，除了"a"、"o"、"ô"等少数几个元音可以单独成词使用以外，绝大多数情况下主要元音需要与首辅音和韵尾搭配构成词形，如，"đường（路）"中主要元音"ươ"需要与首辅音"đ"和韵尾"ng"搭配构成词形。因此，如果一个词形中出现了"首辅音+首辅音"的情况，则可判定为缩略语。如"BĐCL（主力部队）"中"B"、"Đ"、"C"、"L"均为首辅音，不可能是一个常规的越南语单词。

4.4 全称识别模块

我们主要利用表2中所列出的缩略语构词模式来设计缩略语对应全称的识别规则。具体步骤如下：

1. 字母转换。由于越南语缩略时存在字母变异问题，对于包含变异字母的缩略语需要在匹配前将经过变异的字母还原。例如：将字母"W"还原为字母"Ư"，将字母"F"还原为"PH"。而字母"D"可能是由字母"Đ"缩略时变异产生，也可能是字母"D"直接缩略而成，因此会转换出这两种可能情况的不同缩略语。

2. 将转换后的缩略语各大写字母和标点符号放入数组 S[n]中。例如，缩略语"TCT"对应的数组元素依次为 S[1]="T"，S[2]="C"，S[3]="T"。

3. 将数组 S[n]中的元素依次同匹配空间中各词形的首字母和标点符号相匹配，检索出缩略语对应的全称。由于缩略语中各大写字母通常都与全称各主要词形的首字母相对应，可以利用该特点制定匹配规则，检索出正确的全称。具体来看，分为两种情况：

1）"缩略语（全称）"模式。其匹配空间为左括号到右括号之间的词形（如图2）。

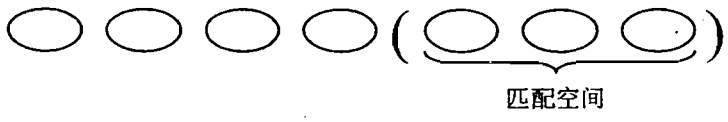

图2

2）"全称（缩略语）"模式。经过对越南《人民军队报》2009年1-3月份的文本内容中字母缩略语的分析，全称中词形和标点符号的总数不超过字母缩略语中字符数量的2倍。若字母缩略语中大写字母的数量为n，则其匹配空间为左括号之前2n个词形（如图3）。

图 3

我们根据缩略语的模式及其构成方式制订了匹配模板（见表 4）。S_n 表示缩略语中的大写字母，Q_n 表示全称中词形的首字母。

序号	缩略语	全称	示例
1	$S_1S_2S_3S_4$	$Q_1Q_2Q_3Q_4$	BĐCL (bộ đội chủ lực)
2	$S_1S_2S_3S_4$	Q_1Q_2, Q_3Q_4	TSVM (trong sạch, vững mạnh)
3	$S_1S_2S_3S_4$	Q_1Q_2 - Q_3Q_4	QPAN (quốc phòng - an ninh)
4	S_1S_2 - S_3S_4	Q_1Q_2 - Q_3Q_4	QP-AN (quốc phòng - an ninh)
5	S_1S_2 - S_3S_4	$Q_1Q_2Q_3Q_4$	KT-KT (kiến thức kinh tế)
6	S_1S_2 - S_3S_4	Q_1Q_2, Q_3Q_4	CNH-HĐH (công nghiệp hoá, hiện đại hoá)
7	S_1S_2 - S_3S_4	Q_1Q_2 với Q_3Q_4	QP-KT (quốc phòng với kinh tế)
8	S_1S_2, S_3S_4	Q_1Q_2, Q_3Q_4	QP, QS (quốc phòng, quân sự)
9	S_1S_2 / S_3S_4	$Q_1Q_2Q_3Q_4$	TCC/T (tác chiến cấp tỉnh)
10	S_1S_2 / S_3S_4	Q_1Q_2 của Q_3Q_4	SĐKHTQ/Đ (sơ đồ kế hoạch tiến quân của địch)
11	S_1S_2&S_3S_4	Q_1Q_2 và Q_3Q_4	BGD&ĐT (Bộ Giáo dục và Đào tạo)

表 4

具体匹配方法如在子句“Huyện đã tổ chức quán triệt sâu sắc các quan điểm, đường lối quân sự, quốc phòng (QS-QP)”中，缩略语“QS-QP”对应的数组元素依次为 S[1]=“Q”，S[2]=“S”，S[3]=“-”，S[4]=“Q”，S[5]=“P”，其匹配空间为“quan điểm, đường lối quân sự, quốc phòng”。按照匹配模板，从匹配空间的左侧开始，首先对 S[1]对应元素“Q”进行匹配，可以匹配上词形“quan”，但是接下来的词形“điểm”并不能匹配上 S[2]对应的元素“S”，匹配失败。将匹配空间缩小为“điểm, đường lối quân sự, quốc phòng”，重新开始匹配，检索到词形“quân”匹配上 S[1]对应的数组元素“Q”，词形“sự”匹配上 S[2]对应的元素“S”，标点符号“，”根据模板 6 匹配上 S[3]对应的元素“-”，词形“quốc”匹配上 S[4]对应的元素“Q”，词形“phòng”匹配上 S[5]对应的数组元素“P”，最后提取出正确的字母缩略语及其全称对“QS-QP| quân sự, quốc phòng”。

5. 实验及评价

5.1 实验结果

我们以越南科技类杂志《科学活动》（Hoạt Động Khoa Học）2009 年 7~9 期电子版文献语料作为实验数据，该语料包括 63 篇文献，共计 151515 个越南语词形。我们将准确率、召回率和 F 值作为系统的测试指标，以领域专家的意见作为参考标准。我们邀请了三位具有硕士以上学位的越南语教师分别对语料中的字母缩略语及其全称进行人工提取，得到一份包含字母缩略语、全称和文献编号的统计表。分析三位教师的反馈信息，实验语料中共有 64 个缩略语和全称对。利用

规则匹配的方法，我们最后提取到缩略语和全称对有 51 个，其中正确的有 50 个：

语料中缩略语及全称对个数	提取出的缩略语及全称对个数	提取出的正确缩略语及全称对个数	准确率	召回率	F 值
64	51	50	98.04%	78.13%	86.96%

表 5

5.2 实验结果分析

从测试结果来看，缩略语及全称对提取的准确率较高，达到了 98.04%。出现的一个错误在于提取的结果不是越南语缩略语而是英文缩略语（Sucrose Phosphate Synthase – SPS，蔗糖磷酸化酶），虽然它也符合匹配规则，但是不属于越南语缩略语。而系统的召回率为 78.13%，还有待提高。出现漏检的原因主要有以下三种情况：

（1）越南语特定表达的特殊格式。如“Nghị định số 119/1999/NĐ-CP của Chính phủ”中，“NĐ-CP（政府决定）”中“NĐ”的全称“Nghị định”位于缩略语的前面，而“CP”的全称“Chính phủ”位于缩略语的后面，因此缩略语的全称需要重新进行整合。

（2）缩略语与全称都作为注释放在括号中。如...(sinh viên, học viên cao học - HVCH, nghiên cứu sinh - NCS và...)中，“HVCH（học viên cao học 硕士研究生）”、“NCS（nghiên cứu sinh，博士研究生）”的缩略语和全称都是作为注释形式放在括号中，而不是采取一般的全称（缩略语）或缩略语（全称）模式。

（3）全称本身包含缩略语。如“Nghiên cứu SHTT (NCSHTT)”中，“SHTT”本身就是“Sở hữu trí tuệ（知识产权）”的缩略语。

6. 结论与展望

针对越南语缩略语对自然语言处理的障碍，以及缺乏越南语缩略语词典的现状，本文提出一种基于规则匹配的方法，通过自动在互联网上采集网页构筑动态语料库，利用越南语缩略语词法、句法等方面的特征对未经切分标注的越南语生语料进行自动提取，实现缩略语和全称之间的自动匹配，最后生成一部缩略语词典。该方法是对越南语字母缩略语及其全称自动提取的一个尝试，实验结果表明，我们的方法是有效的。未来我们将在考察更多越南语语料的基础上，根据越南语自身的特点，制定更加全面和完善的提取规则，进一步提高自动提取的效果。

参 考 文 献

[1] Xu, Yun, Wang, Zhihao, Zhao, Yuzhong & Xue, Yu. A New Alignment Algorithm to Identify Definitions Corresponding to Abbreviations in Biomedical Text [C] // *Proceedings of the First International Workshop on Knowledge Discovery and Data Mining*. Washington, DC: IEEE Computer Society, 2008: 118-124.

[2] Sun, Xu, Wang, Houfeng & Wang, Bo. Predicting Chinese Abbreviations from Definitions: An Empirical Learning Approach Using Support Vector Regression [J]. *Journal of Computer Science & Technology*, 2008，23 (4): 602-611.

[3] 戴维·克里斯特尔. 现代语言学词典[M]. 沈家煊，译. 北京：商务印书馆，2007.

[4] 丁俊苗. 现代汉语缩略语自动识别研究的现状与展望[J]. 渭南师范学院学报，2008，23（6）：39-43.

[5] 李新. 最新实用英汉缩略语速查手册[M]. 北京：华龄出版社，2004.