中文机构名简称的自动生成研究

计峰　高沫　邱锡鹏　黄萱菁

复旦大学计算机科学技术学院 上海 200433

E-mail: {fengji, 0572237, xpqiu, xjhuang}@fudan.edu.cn

摘　要：本文提出了一个自动生成中文机构名简称的方法。我们的方法是将简称生成问题转化为等价的序列标注问题，并利用一阶条件随机场建立自动生成模型。不同于前人的工作，我们的方法没有使用分词信息。在高校及公司企业简称数据的实验中，F1 值分别取得了 86.18%和 75.89%。

关键词：简称生成，条件随机场

Research on Automatic Abbreviation Generation of Chinese Organization

Feng Ji, Mo Gao, Xipeng Qiu, Xuanjing Huang

School of Computer Science and Technology, Fudan University, Shanghai 200433

E-mail: {fengji, 0572237, xpqiu, xjhuang}@fudan.edu.cn

Abstract: In this paper, an automatic abbreviation generation method of Chinese organization has been proposed. In our method, we transformed the problem into an equivalent sequence tagging problem, and built up the automatic generation model through the first order conditional random field. Different from the works of others, our method didn't need the segmentation information. In our experiments in the high-school and corporation abbreviation datasets, F1 achieved 86.18% and 75.89% respectively.

Keywords: Abbreviation Generation, Conditional Random Field

1 前言

在现代汉语中，缩略语是一种被广泛使用的语言现象。据统计，在新闻文章中，约 20%的句子中含有缩略语。缩略语本质上是人们对于较长短语约定俗成的简洁表达，也常被称为简称。通常简称不能通过穷举的方式获得，因而在中文信息处理中一般都作为未登录词处理，更不能将简称与相应的全称联系在一起，从而影响了中文分词，共指消解，信息检索，机器翻译等相关系统的性能。

对于简称自动生成的研究已有一些相关工作。钟良伍[1]根据特定形式的全称，使用规则产生出若干候选简称，通过匹配的方法寻找出与简称数据库中相匹配的简称。这种方法依赖于规则，对于不符合特定形式的全称就无能为力了；同时该方法本质上只是寻找全称的一个匹配，并不是从全称中生成一个简称。Xu[2]认为简称识别是对全称所有可能候选简称的打分排序过程，并于 2006 年使用支持向量回归模型（SVR）建立了识别模型，但这种方法的效率较低，因而在他们的工作中将生成的候选简称的长度限定为不大于 4。Li[3]利用 Web 检索引擎，将全称与候选简称作为查询，并以共现次数对所有获选简称排序。该方法是在已知一个候选简称的情况下判断与全称属于相同类别的可能性，本质上并没有理解全称的具体意义，且与人们从全称生成简称的心理过程不一致。Chang[4]在 2004 年提出了一个 HMM 模型，认为全称中的每个词至少产生一个简称中的字。然而这个模型有很大的局限，因为其假设全称中不存在不产生字的词，而在实际中这种现象却大量存在，如图 1 中“中国/石油/化工/集团/公司”的简称“中石化”中没有出现“集团”或“公司”中的任何字。同时上述的方法都是建立在分词的基础上，因而依赖于分词。如果分词出现错误，那么简称就不能被正确识别出来。

本文中，我们认为全称是由单个字构成的序列，而简称本质上是通过从全称中提取出若干个字生成的，而不是对候选简称排序的过程。这也更加符合人们从全称中生成简称的过程。大部分的简称都是源自于对机构名实体的简写，因此本文主要关注于中文机构名简称的自动生成。本文的组织结构如下：第2部分将介绍自动生成简称的方法，第3部分描述了我们的实验方法以及对实验结果的分析，最后为全文总结和将来的工作。

2 中文机构名简称自动生成

中文机构名的简称主要有两个特点，其一为简称中的字都应出现在全称中，其二为简称中字出现的次序与全称中的次序保持一致。如图 1(a)中，“中国石油化工集团公司”的简称为“中石化”，“中”，“石”和“化”3个字都出现在全称中，并且保持了偏序关系。

图 1 机构名简称

这两个特点表明，机构名的简称可以通过序列标注模型的算法自动抽取得到。本节先引入简称的形式化定义，接着介绍一阶条件随机场模型，最后介绍在本文实验中使用到的特征。

2.1 简称的形式化定义

对于由 n 个字构成的机构名全称 $S = c_1 c_2 ... c_n$，存在另外一个由 m 个字构成的字符串 $A = a_1 a_2 ... a_m$，如果 A 中任何一个字 a_i 都是 S 中的字，并且 $m < n$，我们就称 A 为 S 的候选简称。对于候选简称 A，可以建立条件概率模型 $P(A|S)$ 计算其是 S 的简称的可能性，那么 S 的简称就应选择为条件概率 $P(A|S)$ 最大的候选简称 A^*，即

$$A^* = \arg \max_A P(A|S)$$

由于候选简称 A 是由全称 S 中的字构成的，并且保持字序，因此我们可以构造一个与 S 同样长度的序列 E，等价地表示候选简称 A。构造的方法如下，序列 E 中每个位置的取值可以是 I 或者 O，其中 I 表示 S 中对应位置的字出现在候选简称 A 中，O 表示不出现在 A 中。例如图 1 (a)中，简称中出现的字“中”、“石”和“化”对应位置处被标为 I，而其他不出现在简称中的字对应的位置被标为 O。可以看出这种构造是等价的，候选简称 A 可以构造出唯一的序列 E，同样地序列 E 也可以还原出唯一的候选简称 A。因此原先的问题就转化为序列标注问题

$$E^* = \arg \max_E P(E|S)$$

2.2 一阶条件随机场模型

我们已经将原问题等价地转化为序列标注的问题，而序列标注问题已有大量的研究。本文中采用了一阶条件随机场模型[5]。

条件随机场是一个建立在无向图上的概率模型。假设给定一个观察序列的随机变量 $\boldsymbol{X} = (x_1, x_2, ..., x_n)$，以及一个同等长度的标记序列的随机向量 $\boldsymbol{Y} = (y_1, y_2, ..., y_n)$，那么一阶条件随机场刻画了这样一个条件分布

$$p(\boldsymbol{Y}\mid\boldsymbol{X})=\frac{1}{Z(\boldsymbol{X})}\exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(y_i,y_{i-1},\boldsymbol{X})\right\}$$

其中归一化因子 $Z(\boldsymbol{X})=\sum_{\mathbf{y}\in\boldsymbol{Y}}\exp\left\{\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(y_i,y_{i-1},\boldsymbol{X})\right\}$，$f_k(y_i,y_{i-1},\boldsymbol{X})$ 为特征函数。

可以看出，条件随机场的特征函数以整个观察序列 $\boldsymbol{X}$ 以及标记 y_i 和 y_{i-1} 为自变量，因而可以充分利用上下文信息作为特征，并且能够使用非独立的特征，从而增强了条件随机场模型的表达能力。相比较于条件最大熵模型，由于特征函数中引入了 y_{i-1}，使得标记序列 $\boldsymbol{Y}$ 可以全局归一化，因而克服了“标签偏置”的问题。但也大大增加了模型的复杂度。研究表明，在相同的序列标注任务中，条件随机场的性能都要好于最大熵模型。

2.3 特征选择

影响全称中的字是否被抽取出来的因素有很多，表 1 列出了本文使用的特征。我们主要考察了 3 大类的特征，分别为 n 元上下文特征，地名词典特征以及环境变异度特征**错误！未找到引用源。**，其中 n 元上下文特征在实验中作为基本特征。

表 1　文中实验中使用的特征

特征名称	特征	说明
n 元上下文特征	c_i	1 元特征，$i=-2,-1,0,1,2$
	c_ic_{i+1}	2 元特征，$i=-2,-1,0,1$
	$c_{-1}c_1$	2 元特征
	$c_{-1}c_0c_1$	3 元特征
地名词典特征	$dict(c_{-1}c_0)$	$c_{-1}c_0$ 是否出现地名词典中
	$dict(c_0c_1)$	c_0c_1 是否出现地名词典中
	$dict(c_{-2}c_{-1}c_0)$	$c_{-2}c_{-1}c_0$ 是否出现地名词典中
环境变异度特征	$av_n(c_0)$	c_0 字的 av 值，$n=1,2$

地名词典特征是个二值特征，是通过察看当前字与连续相邻的 n 个字构成的短语是否出现在地名词典中。地名词典收集了全国各大省市及其主要城市，共 113 个。本文中 n 选取为 1 和 2。如在“上海东方电视台”中，如果当前字为“海”时，与前 1 个字构成“上海”并出现在词典中，那么该特征的值为 1；而与后一个字构成的“海东”却不是地名，特征的值就为 0。

环境变异度特征反映了字的边界特征。Zhao[6]从未标注语料中统计出这个特征，并将其应用于分词系统中，对于未登录词的识别具有很好的作用。本文中由于我们的方法并没有使用分词信息，因此环境变异度可以直接从我们的语料上统计获得。

字符串 s 的环境变异度定义为

$$AV(s) = \min\{L_{av}(s),\ R_{av}(s)\}$$

其中 $L_{av}(s)$ 和 $R_{av}(s)$ 分别表示与 s 左邻接和右邻接的不同字的数量。该值反映了字符串 s 与其上下文环境之间的独立性，当 s 中的第一个字或者最后一个字正好是一个词的边界时，那么该值较大；而当 s 仅是一个词中的一部分，那么该值较小。因此该特征一定程度反映了 s 成为边界的可能性。

根据字符串 s 可以定义字 c 的环境变异度为在窗口大小为 n 的字符串 s 的环境变异度的最大值，其中字符串 s 必须包含字 c，即

$$av_n(c) = \max\{AV(s) \mid |s| = n,\ c \in s\}$$

3 实验结果与分析

3.1 实验数据收集

由于目前为止还没有正式公开的中文机构名简写的语料，所以我们利用互联网以及两条简单的规则，人工收集了中国高校的全称名录以及公司企业的全称名录。共获得了 525 所高校的全称简称对，以及 1411 家公司企业的全称简称对。我们随机选取了 80%的数据作为训练，20%的数据作为测试。

3.2 实验结果及分析

我们使用工具CRF++[1]训练一阶条件随机场模型。基准系统为只使用了 n 元上下文特征训练的模型。在实验中我们设置高斯先验为 1。评价指标使用了被普遍采用的精度(P)，召回率(R)和F值(F)。实验结果如表 2。

表 2　公司企业与高校数据上的实验结果

	公司企业			高校		
	P	R	F1	P	R	F1
base	0.7558	0.7395	0.7476	0.8358	0.8796	0.8571
base+dict	0.7596	0.7440	0.7517	0.8402	0.8765	0.8580
base+av	0.7573	0.7418	0.7495	0.8401	0.8796	0.8594
base+dict+av	0.7659	0.7520	0.7589	0.8309	0.8951	0.8618

从上述实验结果中可以看到，基准系统的性能要低于其他所有增加特征的系统，最好的结果为使用了全部特征的系统，最大的增幅达到 1.1%。高校上的实验结果普遍要好于公司企业的实验结果。我们认为造成这种情况可能是因为公司企业数据的来源不太一致。我们收集的公司企业数据包括了上海和深圳两大交易所的上市公司以及中国 500 强企业。而上市公司在上海和深圳交易所的简称命名方法不太一致，上海证交所一般采用 4 个字的简称，而深圳证交所更多见的是 3 个字的简称。另外一方面，公司企业的名称中可能包含多个表示企业的短语，如“实业”，“股

[1] http://crfpp.sourceforge.net

份”，“公司”等，而这些短语通常都可以作为简称的一部分。但是如果这些短语同时出现，分类器将很难决定应该将哪个短语放入简称中。如在我们的实验中，“重庆市迪马实业股份有限公司”的实际简称为“迪马股份”，而我们的标注结果是“迪马实业”。“迪马实业”虽然不是正确的简称，但也是能够被接受的简称。实际上，我们的实验结果中有很多类似这样的结果，如“河南神火煤电股份有限公司”的正确简称为“神火股份”，而标注的结果是“神火煤电”。

比较实验的结果，我们发现对于公司企业的简称，地名词典的作用较大。地名词典的特征在基准系统的基础上提高了0.4%多，而词的环境变异度只提高了不到0.2%，而在高校的简称抽取实验中又正好相反。造成这种现象的原因可能是由于两种机构名的简称方式存在很大不同。公司企业通常可以选取连续的若干个字就能作为其不含歧义的简称，如“中国银行股份有限公司”通常只需开始的“中国银行”4个字作为其简称，因而通常全称中的词会完整地出现在简称中。而在高校的简称中，这种现象较少出现，如“华东师范大学”虽然也可以简称为“华东师大”，但是人们更愿意接受“华师大”，因为后者更加简洁并也不带有歧义，因此高校的简称多是从全称中的词挑选一个字代表该词放入简称中。在标记序列上表现为公司企业的简称序列通常含有一段连续的I，而高校的简称序列I通常是以跳跃若干个 的方式出现。这种现象使得完整的地名信息相对于公司企业的简称更加重要，而环境变异度的信息对于高校数据相对更加重要。

4 总结

中文机构名的简称是日常生活中常见的现象，本文认为中文机构名简称是通过从全称中选取若干个字形成的，而不是在一堆候选简称中通过排序挑选出来的。基于这样的认识，我们采用条件随机场的序列标注模型完成了对中文机构名简称自动抽取的实验。不同于前人提出的方法，我们的方法不需要分词。在高校与公司企业数据的实验中，分别达到了75.89%和86.18%的F1值。

但是我们也可以看到，我们的数据量并不是很大，因此接下来我们会继续收集数据，不仅是中文机构名，而且包括更多类型的实体，用于更多的实验。同时目前我们使用到的特征并不是很多。我们都知道中文是一种讲究韵律节奏的语言，人们在生成简称的过程中也相当注意韵律节奏的问题，因此可以想象到韵律节奏的特征将对于简称的自动抽取应该有较大的帮助。在接下来的工作中，我们考虑加入此类特征，并希望能够提高性能。

参 考 文 献

[1] 钟良伍, 郑方. 基于中文机构名简称的检索方法研究. 中文信息学报, 2007, 21(1):38~42.

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

[3] Zhifei Li, David Yarowsky. Unsupervised Translation Induction for Chinese Abbreviations using Monolingual Corpora. Proceeding of ACL-08:HLT, 2008:425~433.

[4] Jing-Shin Chang, Tu-Tso Lai. A Preliminary Study on Probabilistic Models for Chinese Abbreviations. Proceedings of the Third SIGHAN Workshop on Chinese Language Learning, 2004:9~16.

[5] Charles Sutton, Andrew McCallum. An Introduction to Conditional Random Fields for Relational Learning. Book chapter in Introduction to Statistical Relational Learning. MIT Press. 2006.

[6] Hai Zhao. Character-Level Dependencies in Chinese: Usefulness and Learning. Proceedings of the 12th Conference of EACL, 2009:879~887.