中文缩略语自动抽取初探*

崔世起 刘群 林守勋 孟遥 于浩 西野文人

中科院计算技术研究所 北京 100080

富士通研究开发中心有限公司 北京 100016

E-mail:sqcui@ict.ac.cn liuqun@ict.ac.cn sxlin@ict.ac.cn

摘 要：汉语中许多新生的词语都是短语的缩略形式。对缩略语的检测是未登录词识别的一部分，但用来作为训练语料的缩略语词典资源却很稀缺。本文提出一种在生语料中自动抽取中文缩略语的方法，首先获取候选缩略语集和源短语库，然后利用语言模型和对齐模型等特征进行候选缩略语和源短语的对齐，最后得到一部粗糙的缩略语词典。在实验中，在新词中进行缩略语提取的准确率达到了 51.4%，召回率达到了 81.7%。

关键词：缩略语，源短语，对齐模型

Research in Automatic Extraction of Chinese Abbreviations

CuiShiqi LiuQun LinShouxun MengYao YuHao Nishino Fumihito

Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080

Fujitsu Research & Development Center Co.,LTD, Beijing 100016

E-mail:sqcui@ict.ac.cn liuqun@ict.ac.cn sxlin@ict.ac.cn

Abstract: In Chinese, many new words are abbreviations of some phrases. The identification of abbreviations is part of the identification of unknown words, but we are short of the corpus of abbreviation to train the model. In this paper, we will present an approach to extract Chinese abbreviations automatically from unsegmented corpus. First we collect a candidate abbreviation set and a source phrase library, then make use of the features such as language model and alignment model to produce the pairs of abbreviation and source phrase, and get a rough abbreviation dictionary finally.

Keywords: abbreviation , source phrase, alignment model.

1 引言

缩略语是短语的缩写形式，汉语中的缩略语一般是把音节较长的词或者词组缩减为双音节词，这符合汉语词汇双音节化的主流趋势（《新词新语与新闻语言》）。例如“反对恐怖主义”

*本文承国家 863 项目(编号 2004AA114010 和 2003AA111010)以及中科院计算所和富士通研究开发中心合作项目资助。

被缩写为“反恐”，“足球彩票”被缩写为“足彩”。当然，缩略语不限于两个字，如“航空意外险”被缩写为“航意险”,但是三个或者更多字的缩略语数量较少。我们把被缩写的短语称为源短语，如，“足球彩票”是“足彩”的源短语，“足彩”是“足球彩票”的缩略语。

在进行新词检测的过程中，我们发现缩略语在新词中占据着较大的比重。把缩略语的识别反过来应用到新词检测的过程中，又可以提高新词检测的准确率。缩略语的研究主要分为三个工作（《A Preliminary Study on Probabilistic Models for Chinese Abbreviations》）。1.给出源短语，检测其缩略语。2.给出缩略语，检测其源短语。3.训练缩略语模型参数。在这三项工作中，都需要利用一部缩略语词典作为训练语料。但由于缩略语词典资源比较少，所以数据稀疏的问题比较严重。本文的工作是在无缩略语词典资源训练的情况下，利用生语料自动提取缩略语和源短语对，以此来生成一部缩略语词典。对此项工作，已有的研究比较少。

2 中文缩略语提取方法

本文设计的自扩展中文缩略语词典方法主要分以下几个步骤：

1. 获取较大的文本语料。该语料是未经过切分标注的生语料。由于我们进行的是无监督的学习，没有缩略语词典作为训练语料，所以生语料的规模要足够的大。我们通过从因特网采集大量的网页并进行网页正文的提取来产生实验所需要的庞大语料库。

2. 从语料中提取候选缩略语集。

3. 从语料中提取候选源短语库。

4. 进行缩略语和源短语的对齐。

2.1 提取候选缩略语集

在已知的中文词中，有部分是缩略语。我们可以拿一部通用词典作为候选缩略语集。这样做的一个弊端是可能词典中的词并不出现在文本语料中，而其源短语出现在文本语料中的机率会更小，导致无法检测到缩略语的源短语，或者将非缩略语误判为缩略语。因此以一部通用词典作为候选缩略语集会使得召回率和准确率都比较低。

对于缩略形式的新词，往往它的源短语也会被广泛的使用。例如，“非典”和“非典型肺炎”都会大量出现在媒体上。因此，我们假设缩略语和源短语在语料中共现。根据假设，我们可以从语料中提取新词作为缩略语的候选集。实验发现，新词中缩略语占了相当大的比重，我们可以通过对语料库进行检测得到新词（该项工作在笔者的另一篇论文中详细介绍），以此作为候选缩略语集。本文最后会在自动检测得到的新词构成的候选缩略语集上进行实验并给予分析。

2.2 提取候选源短语库

我们不可能设计一部可以覆盖所有短语的库，根据假设，源短语和缩略语来源于同一个语料库中，所以我们的源短语库需要从同一个语料库中提取。源短语之所以会被缩写，是因

为该短语使用频度较高，使用者本着简洁的原则才将其进行缩写，所以源短语必定在语料库中有着较高的出现频率。从词性上讲，绝大多数被缩略的源短语都是由实词构成的，所以我们只需统计由名词、动词和形容词构成的源短语。短语是由若干个词组成的，我们首先利用计算所的 ICTCLAS 对文本进行切词和词性标注，然后进行基于词的重复串搜索和基于词性的过滤（邹纲 2004），这样就获取到一个很大的源短语库。

2.3 缩略语和源短语的对齐

2.3.1 中文缩略语的构成规律

绝大多数情况下，把源短语切分为几个词，每个词用该词中的某字代替，就得到了缩略语。如“航空意外险”切分为“航空 意外 险”，然后缩写为“航意险”。在例外的情况下，“非典型肺炎”切分为“非 典型 肺炎”，但其缩略语却是“非典”。但由于例外的情况很少，为了降低问题的复杂性，我们在设计中给出了以下约束：

1. 缩略语中的每一个字，都唯一映射到源短语中的某一个词。
2. 源短语中的每一个词，都唯一映射到缩略语中的某一个字。
3. 不存在词序的重排。

形式化表示为：

$$Source = W_1 W_2 \cdots W_n, Abbreviation = C_1 C_2 \cdots C_n$$

$$W_i = C_{i1} C_{i2} \cdots C_{im}, \exists C_{ij} = C_t, (1 \le j \le m)$$

Source 表示源短语，*Abbreviation* 表示缩略语。不满足该约束的缩略语和源短语对，不在本文研究的范畴。

使用上述约束处理候选缩略语集和源短语库，可以得到大量的缩略语和源短语对，它们之间是多对多的关系。如“足彩”对应“足球彩票”和“足球博彩”，“初步裁定”对应“初裁”和“初定”。如图所示，每一个对应产生了一条连线，所以在候选缩略语集和源短语库之间存在大量的连线，其中有正确的连线表明了缩略关系，还有很多错误的连线。我们的目标是去掉错误的连线。

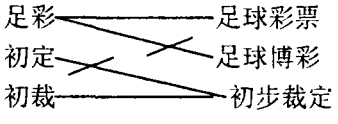

图 1

2.3.2 缩略语和源短语对的筛选

我们需要计算每一个候选缩略语和源短语对的概率 $P(Abbreviation, Source)$，来衡量该对有效的程度。根据约束条件，我们设计了 $P(Abbreviation, Source)$ 的计算公式。为了使得公式更加直观，我们使用 *Abbreviation* = “足彩”，*Source* = “足球彩票”来表示该公式。

P(足彩，足球彩票) = P_{lang}(足球彩票) × P_{align}(足彩|足球彩票)

其中，P_{lang} 是语言模型，P_{align} 是对齐模型。

在公式中，语言模型表达的是源短语合法的程度。源短语是自动抽取的，我们并不知道

它的合法性，所以公式中使用源短语的语言模型对其进行约束。短语的语言模型可以采用二元语法。

$$P_{lang}(\text{足球彩票}) = P(\text{足球}) \times P(\text{彩票}|\text{足球}) = \frac{N(\text{足球})}{N(total)} \times \frac{N(\text{足球彩票})}{N(\text{足球})} = \frac{N(\text{足球彩票})}{N(total)}$$

其中，N(足球彩票)表示短语“足球彩票”在语料中的频率，$N(total)$表示语料中的总词数。

对齐模型表示源短语和缩略语的对齐概率，即它们相对应的程度。一般来讲，对齐概率需要通过利用缩略语词典来计算。本文的研究目的是自扩展一个缩略语词典，所以无法直接利用词典计算对齐概率。

2.3.3 对齐模型的计算

剩下的一部分工作就是对齐模型的计算。在候选缩略语和源短语之间产生了很多连线，源短语是已经分好词的，这样源短语和缩略语的对齐概率就可以转换为词和字的对齐概率。

$$P_{align}(\text{足彩}|\text{足球彩票}) = P_{align}(\text{足}|\text{足球}) \times P_{align}(\text{彩}|\text{彩票})$$

根据给定的约束，我们可以在缩略语中的字和源短语中的词之间画出许多连线。如图所示：

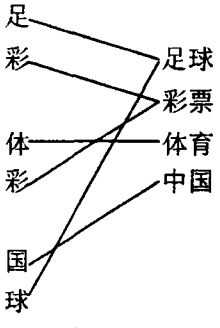

图 2

我们用 P_{align}(足|足球)表示“足球”缩写为“足”的概率，P_{align}(球|足球)表示“足球”缩写为“球”的概率。字和词之间的连线越多，表示该字和该词对齐的可能性越大，所以 P_{align}(足|足球)是“足”和“足球”之间的连线数的递增函数。

定义下面的公式：

$$P_{align}(\text{c}|\text{w}) = \frac{N(c,w)}{N(w)} \times 100\%$$，$N(c,w)$表示字 c 和词 w 之间的连线数目，$N(w)$表示与词 w 连线的总数目。

图 2 中，假设候选缩略语集是{足彩，体彩，国球}，源短语库是{足球彩票，体育彩票，中国足球}，那么 P_{align}(足|足球)=50%，P_{align}(球|足球)=50%，P_{align}(彩|彩票)=100%，P_{align}(票|彩票)=0。

至此，$P(Abbreviation, Source)$的值可以通过短语的语言概率、词和字的对齐概率进行

计算。对每个候选缩略语的多个候选源短语，计算 $P(Abbreviation, Source)$ 的值，取得分最高的作为结果。然后设定阈值，对产生的缩略语和源短语对进行筛选。

3. 实验结果

我们在实验中选用的候选缩略语集包含在 50 万张网页中自动检测的 551 个新词，这些新词未经过人工筛选，其中包括 93 个缩略语。短语库包含在相同的语料上自动提取的 286378 个短语。利用缩略语和短语的对齐方法，我们最后检测到缩略语和源短语 148 对，其中正确的有 76 对。

候选缩略语	候选源短语	准确率	召回率	F 值
551	286378	51.4%	81.7%	63.1%

表 1

因为新词中缩略语占了较大的比重，而且一般情况下，新词的使用还没有完全稳定下来，新词和其源短语会同时出现在语料中，并且共现度非常高。比如“伊战”和“伊拉克战争”，“医保”和“医疗保障”等都可以正确的召回。由于完全是在无监督训练的情况下进行缩略语的检测，所以我们认为以上结果还是不错的。

另外，已知词中的缩略语比重较低，而且已知词的源短语已经很少在语料中出现，所以如果以已知词做为候选缩略语集，检测的准确率和召回率会很低。当然，如果不追求较高的性能，本文的方法仍然适用。我们用该方法在已知词中找到了不少缩略语、源短语对，如“诚信”和“诚实守信”、“城建”和“城市建设”、“电汇”和“电子汇款”、“政绩”和“政府绩效”、“预支”和“预先支付”等。

4. 结论

针对缩略语词典资源稀少的现状，本文提出了一种自扩展缩略语词典的方法，通过自动在 Internet 上采集网页，利用 ICTCLAS 进行切词，使用重复串搜索技术和词性过滤方法获取候选缩略语集和候选源短语库，然后使用短语的语言模型以及词和字的对齐模型进行缩略语和源短语对的筛选，只需要很少的人工干预，就可以得到一部粗糙的缩略语词典。本文只对缩略语提取做了初步的研究，在新词中进行缩略语的提取取得了不错的效果，对已知词的缩略语提取，需要做进一步的研究。

致谢 笔者在撰写本文的过程中，得到了刘洋、熊德意的帮助，在此表示感谢。

参 考 文 献

[1] Jing-Shin Chang, Yu-Tso Lai.A Preliminary Study on Probabilistic Models for Chinese Abbreviations. Proceedings of the Third SIGHAN Workshop on Chinese Language Learning. ACL-2004. Barcelona, Spain.

2004.pp. 9-16.

[2] Andi Wu, Zixin Jiang. Statistically-Enhanced New Word Identification in a Rule-Based Chinese System. Proceedings of the Second Chinese Language Processing Workshop. Hong Kong. 2000.

[3] Huang, Chu-Ren, Wei-Mei Hong, and Keh-Jiann Chen. "Suoxie: An information based lexical rule of abbreviation." In Proceedings of the Second Pacific Asia Conference on Formal and Computational Linguistics II, 1994, Japan, pages 49–52.

[4] 邹纲，刘洋，刘群 et al.面向 Internet 的中文新词语检测.中文信息学报，2004，18(6)：1~9

[5] Richard O.Duda, Peter E.Hart, David G.Stork. Pattern Classification Second Edition. ISBN: 0-471-05669-3. October 2000.

[6] Chang, Jing-Shin and Keh-Yih Su, 1997. "An Unsupervised Iterative Method for Chinese New Lexicon Extraction", International Journal of Computational Linguistics and Chinese Language Processing (CLCLP), 2(2): 97-148.

[7] 郝全梅,劳臣.新词新语与新闻语言.新闻战线.2004,7.