基于双语平行语料的中文缩略语提取方法*

刘友强[1]，李 斌[1,2]，奚 宁[1]，陈家骏[1]

[1]南京大学 计算机软件新技术国家重点实验室，南京 210093

[2]南京师范大学 语言信息科技研究中心，南京 210097

E-mail: {liuyq,lib,xin,chenjj}@nlp.nju.edu.cn

摘 要：汉语缩略语在现代汉语中被广泛使用，其相关研究对于中文信息处理有着重要的意义。本文提出了一种从英汉平行语料库中自动提取汉语缩略语的方法。我们首先对双语语料进行词对齐训练，利用训练得到的词对齐信息抽取出候选中英文短语对。然后用 SVM 分类器提取出质量高的短语对。最后再从质量高的短语对集合中利用英文翻译及一些汉语缩略–全称对应规则提取出汉语缩略语及全称语对。实验结果表明，该方法提取出的缩略语具有较高的准确率，可以作为一种自动提取缩略语词典的有效方法。

关键词：缩略语；平行语料库；短语抽取；分类

A Bilingual-constrained Approach for Extracting Chinese Abbreviations

Liu You-qiang[1], Li Bin[1,2], Xi Ning[1], Chen Jia-jun[1]

[1]State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing 210093

[2]Research Center of Language and Informatics, Nanjing Normal University, Nanjing 210097

E-mail: {liuyq, lib, xin, chenjj}@nlp.nju.edu.cn

Abstract: Chinese abbreviations are widely used in modern Chinese texts, and the correlated research is important for Chinese information processing. In this paper, we propose an approach to extract Chinese abbreviations from Chinese-English parallel corpus. First we generate word alignments for the corpus, and extract Chinese-English phrase pairs consistent with the alignments. After that, we discriminate high quality phrase pairs from the bad ones by SVM Classifier. Then we extract Chinese abbreviation and full-form phrase pairs from the high quality group using their corresponding English translations and some rules. The experiments showed that our approach can extract abbreviations with high accuracy, and could be an effective way to extract Chinese abbreviation and full-form phrase pairs.

Keywords: abbreviation; parallel corpus; phrase extraction; classify

1 引言

缩略语是短语或词的全称的缩写形式，如“中国”简称“中”。由于其省时省力的效果，在自然语言中被广泛使用，是未登录词的主要来源之一。据研究，在一篇典型的中文新闻文章中，近20%的句子包含缩略语[1]。而未登录词对于中文的自动分词与词性标注等词法句法分析任务有很大的影响，这使得中文缩略语有较大的研究价值。

一般来说，现代中文缩略语的构成方式主要有四种：（1）语素方式：缩略语由原词语各部分语素构成。例：奥林匹克 运动——奥运。（2）中心词方式：缩略语由原词语核心的词构成。例：人造 地球 卫星——人造卫星。（3）混合方式：缩略语由语素和中心词构成方式混合使用而得。例：中央 电视台——中央台。（4）合并方式：缩略语由原词语中的并列词归纳而得。例：包退、包换、包修——三包。

从整体上看，缩略语研究可以分为缩略语的探测识别、简称-全称的对应（还原生成）两大类

* 该工作得到国家自然科学基金（61003112，61073119）、国家社会科学基金（10CYY021）、南京大学计算机软件新技术国家重点实验室（KFKT2011B03）的资助。

工作。在缩略语的探测识别方面，Zhu et al.[2]针对单字人名、地名简称，采取了基于分类器的预测模型；李斌等[3]对汉语单字国名采取了统计评分法进行识别。缩略语的自动识别研究工作主要集中于缩略语的“简称-全称”的还原、生成工作以及缩略语词典的自动构建。在还原、生成方面，Chang & Lai[1]将缩略语的生成和还原问题转化为隐马尔可夫模型（HMM）问题，使用缩略语词典进行训练。支流等[4]设计了一个基于模糊匹配的缩略语还原算法，从缩略语上下文和缩略语词典中获得备选的全称短语。在缩略语词典自动构建方面，崔世起等[5]针对未登录词，在生语料中使用重复串搜索技术和词性过滤获得候选缩略语集和全称短语库，再利用语言模型和对齐模型进行候选缩略语和全称短语的对齐，最后得到 148 对缩略-全称语对，准确率为 51.4%。武子英等[6]从词性标注语料中获得候选缩略语集和全称短语库后，利用上下文的相似度对缩略语和全称短语配对，从而获得缩略语词典，准确率达到74.1%。这两种方法都是在汉语单语文本上的工作，有两点不足：（1）缩略语的采集效率比较低。多重视“简称-全称”的对应，而作为对应前提的简称的自动识别则研究较少。（2）仅使用单语的缩略规则模板，导致准确率不是很高。

中文缩略语的大量存在对汉-外统计机器翻译造成一定的影响。Li et al.[7]提出了一种获得中文缩略语英文翻译的方法。该方法首先识别英文语料中的实体，并翻译为中文短语，以此作为全称短语。然后，根据中文单语语料中短语的共现信息提取出缩略语，以英文实体为其翻译。该方法的目的是获得候选缩略语的英文翻译，因而对于缩略-全称语对的准确度要求不高。但这启示我们两种语言的翻译关系可以作为联系全称和缩略语的桥梁。

本文遵循从双语对译关系中挖掘全称-简称关系的思路，尝试找到一种准确率比较高的自动获取方法，以中文缩略语为研究对象，取得了不错的实验结果。我们首先从句对齐平行语料库中抽取出中英文短语对。然后根据短语对的一些特征训练出一个 SVM 分类器，将短语对根据对应的质量分为“好”与“不好”两类。从对应质量好的那一类短语对集合中，利用一些约束条件和英文翻译抽取中文缩略-全称语对。实验表明，该方法抽取出的缩略-全称语对有较高的准确度。

2 中文缩略语提取

从句对齐平行语料中提取中文缩略语的过程可分为三个部分：短语对抽取，短语对分类和缩略-全称语对的抽取。

2.1 短语对的抽取

这里短语对抽取采用基于短语的机器翻译[8]的短语对抽取方法，流程如图 1 所示。

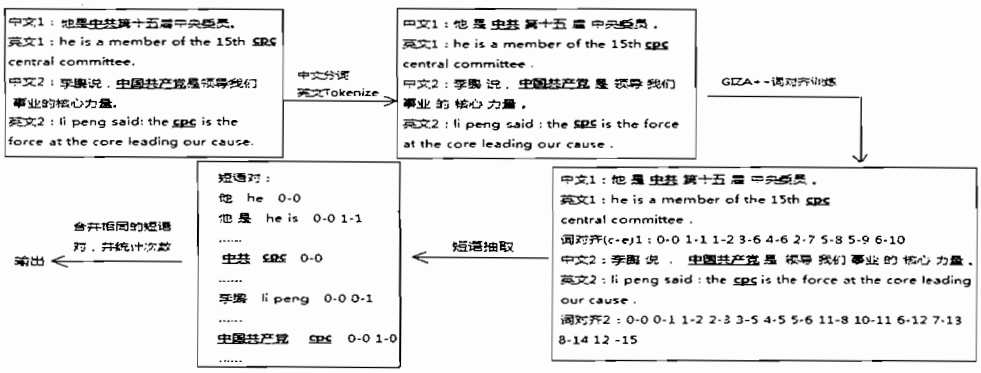

图 1 短语抽取流程图

抽取短语对的步骤：

（1）对平行语料中的中文分词，英文全部换成小写字母并将符号与词分隔开(Tokenize)。

（2）利用开源的词对齐训练工具 GIZA++[1]对平行语料进行词对齐训练。词对齐训练的目标是获得语句对中词的对应关系。如图 2 所示，连线的词之间存在对应关系。

（3）抽取与词对齐信息一致的中英文短语对。这里的短语不是指语言学上的短语。语言学意义上的短语要满足一定的语法结构。这里的短语是指由语句中连续的一个或多个词构成的语句的子串。比如从图 2 中的中文句子抽取的短语可以是“中共”、“中共 第十五”，而这些短语并不符合语言学上短语的意义。这里的短语对抽取可以使用开源的机器翻译系统 Moses[2]。

（4）合并相同的短语对，输出到文件。

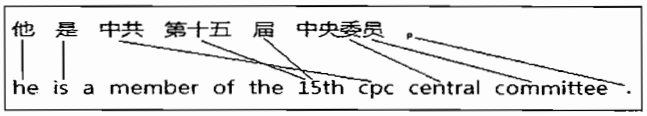

图 2 词对齐示例

从中英文句对 c_1^J 和 e_1^I 中抽取的与它们的词对齐信息 A 一致的短语对集 BP 可定义为[9]：

$$BP(c_1^J, e_1^I, A) = \{(c_j^{j+m}, e_i^{i+n}) \mid \forall (j', i') \in A : j <= j' <= j+m \leftrightarrow i <= i' <= i+n\}$$

其中，c_j^{j+m} 表示由句子 c_1^J 中的第 j 个词开始到第 $j+m$ 个词结束构成的短语，e_i^{i+n} 的含义与此类似。集合中的条件含义可理解为：中文短语 c_j^{j+m} 中的词只与英文短语 e_i^{i+n} 中的词有对应关系，e_i^{i+n} 中的词也只与 c_j^{j+m} 中的词有对应关系。如图 2 中，“第十五 届”与“15th”是一个与对齐一致的短语对，而“第十五”与“15th”则与对齐不一致。

抽取短语对的过程中，为了提高效率，我们排除了那些不太可能作为一个缩略语或缩略语英文翻译的短语。排除条件有：（1）中文或英文短语中含有标点符号；（2）中文短语的边界词为“了”、“是”、“个”等一些不太可能作为缩略语或其全称边界的词；（3）英文短语边界词为介词，或词尾为“the”等一些不太可能作为缩略语或全称的英文翻译边界的词。

2.2 基于 SVM 分类器的短语对分类方法

由于语料库中的噪声以及训练出来的词对齐不可能完全正确，使得相当多的一部分中英文短语对事实上并不对应。这些并不对应的短语对会影响到后面缩略语提取的准确度和效率。因此，我们采用四个特征来衡量中英文短语对的对应质量。并据此训练出一个基于 SVM[10]（支持向量机）的分类器，将短语对根据对应质量分为“好”与“不好”两类。

对于中-英短语对 C-E，其中 C=$c_1c_2\ldots c_n$，E=$e_1e_2\ldots e_m$，采用的四个特征为：

（1）C 翻译为 E 的短语翻译概率,采取极大似然估计。

$\phi(E \mid C) = \dfrac{count(C,E)}{\sum_{E'} count(C,E')}$，其中 $count(C,E')$ 为短语 C 与 E' 对应的次数。

（2）词汇化翻译概率，C 中的词翻译为 E 中的词的概率平均值。

$\varphi(E \mid C) = \underset{A}{Max}\{\varphi(E,A \mid C)\} = \underset{A}{Max}\{\dfrac{1}{m}\sum_{j=1}^{j=m}\dfrac{1}{|\{i \mid (i,j) \in A\}|}\sum_{\forall (i,j)\in A} w(e_j \mid c_i)\}$，其中 A 为训练得到的 C-E 中词的对应关系，由于训练过程中对于相同的 C，E 可能有不同的对应关系，我们这里采用值最大的 $\varphi(E,A \mid C)$ 作为 $\varphi(E \mid C)$ 的值。其中，$w(e_j \mid c_i)$ 为根据语料词对齐信息得到的词翻译概率，采用极大似然估计。

（3）$\phi(C \mid E)$，即 E 翻译为 C 的短语翻译概率。

（4）$\varphi(C \mid E)$，即 E 中词翻译到 C 中词的概率平均值。

[1] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

[2] http://www.statmt.org/moses

2.3 缩略语抽取算法

经上一节分类后得到比较可靠的中-英短语对，接下来的任务就是从这些短语对中提取出候选缩略-全称语对。算法分为两部分：第一部分（2.3.1）抽取出一个缩略-全称语对的候选集，第二部分（2.3.2）对这个候选集进行过滤，获得一个准确度较高的缩略语词典。

2.3.1 匹配约束

我们将中文短语按字长度进行分组，长度不超过 5 的短语被认为是候选的缩略语。一对中文短语对 C_1-C_2（C_1 为缩略语，C_2 为全称语）被选为一对候选缩略-全称短语对，当且仅当：（1）C_1 中的字都在 C_2 中出现；（2）C_1 和 C_2 存在相同的英文翻译。

2.3.2 噪音过滤

为提高缩略-全称短语对的准确性，要对其中的噪音进行过滤。我们对抽取出的候选缩略语对进行了词性标注，使用的工具为 ICTCLAS[1]。我们将候选缩略语的词性限于名词(n)、动词(v)、形容词(a)、区位词(b)及数词(m)。经过观察，我们发现抽取出的候选缩略-全称语对的一些特性。主要分为以下几类：

（1）候选缩略语为单字的情况。此时的抽取出的候选缩略-全称语对可以分为以下几类：

1. 人名、地名等专有名词的缩略。这是单字缩略最常见的情况。如“阿/b-阿根廷/nsf”，“董/nr1-董建华/nr”。这一类缩略-全称语对准确性比较高。
2. 候选缩略语与候选全称有相同的意义，但不是缩略语对。如“园/ng-公园/n”。
3. 噪音。这类语对并不是缩略-全称的关系，是由于词对齐信息不完全正确导致的错误。如“他/rr-表示/v 他/rr”。这类候选缩略语和全称语的词的个数和词性往往不相同。

因此对于候选缩略语为单字的语对，我们根据词性标注的结果选取第一类，也即选取缩略或者全称词性标注为人名(nr)、地名(ns)、机构团体名(nt)及其他专名(nz)的候选语对。

（2）候选缩略语字长为 2，3，4，5 的情况。此时，采用语素构成的候选缩略语正确率很高，而采用中心词构成的候选缩略语正确率较低，是大部分噪音的来源。针对这个特点，我们选取的候选缩略语对分为以下几类：

1. 候选缩略语和全称语为单个词且被均标注为人名(nr)、地名(ns)、机构团体名(nt)及其他专名(nz)。如“国家计委/nt - 国家发展计划委员会/nt”。这里对于专名的处理要求比(1)中严格是因为专有名词的字长较长时更有可能与一些长的短语产生对应关系，尽管这些短语不是它的全称。比如“非洲/nsf-非洲/nsf 国家/n”。同样地，长的专有名词在上下文中也经常被简称为短的非专有名词，然而，这种缩略形式并没有被固定下来。比如“军委/n-中央军事委员会/nt”。
2. 语素构成方式。这类候选缩略-全称语对的准确率较高。根据候选全称语的词长，我们再将之分为两类。候选全称语的词长大于 1 时，我们直接将之选取到缩略语词典中。如“海基会/n-海峡/n 交流/vn 基金会/n”。候选全称语词长为 1 时，此时我们的选取条件是：候选缩略语不是候选全称语的子字符串。如 “中科院/n-中国科学院/nt”。这样做主要是为了排除主要的词重叠的候选缩略-全称语对，这类短语对意义相近，但不是缩略-全称关系。如“人大/n 常委会/n-全国人大常委会/nt”。
3. 混合构成方式。以混合方式构成的候选缩略语中，有很大一部分是由字长较短的缩略语和其他词组合成的短语。例如，“中国/ns 社科院/n-中国/ns 社会/n 科学院/n”由“社科院/n-社会/n 科学院/n”与“中国/ns”组合产生。这一类的候选缩略语对于缩略语词典没

[1] http://www.ictclas.org/

有太多意义。因此我们只选择候选缩略语为单个词的候选缩略-全称语对，如“藏族/nz-藏/b 民族/n”，从而过滤掉由字长较短的缩略语和其他词组合成的候选缩略语。

经过这些规则筛选后，得到的缩略语词典的准确率会得到很大的提高。但是，这些规则也会排除掉一部分真正的缩略语，使得召回率有所下降。

3 实验

3.1 实验过程

（1）语料预处理：实验数据来自于中英机器翻译的平行语料 LDC2003E14[1]，我们从中随机选取了 20 万句对。对中文语料进行分词，英文语料全部换成小写字母并将符号与词分隔开。分词工具采用 Stanford Chinese Segmenter[2]。

（2）词对齐训练：将预处理后的语料用开源软件 GIZA++训练得到词对齐信息。

（3）抽取短语对：按照 2.1 中的方法抽取短语对，最终得到 114446 个短语对。抽取短语对的过程中同时计算 2.2 中提出的衡量中英文短语对对应质量的四个特征。

（4）SVM 短语对分类：从中英文短语对集合中选取 186 条短语对，根据中英文短语是否对应，手工标注为“+1”(是)、“-1”(否)两类。为获得高召回率，我们放松了对应标准。以标注后的数据为训练集，得到一个 SVM 分类器。从短语对集合中随机挑选出 30 条短语对(正负数据各一半)用于测试，结果如表 1。正确率为 65.2%、召回率为 100%，F 值为 78.9%。用 SVM 分类器对短语对分类后得到结果为正的短语对 91884 句，占总短语对数的 80.28%。

表 1　中英文短语对对应质量分类器的测试结果

	正	负
预测为正	15	8
预测为负	0	7

（5）匹配约束：选出（4）中分类后标注为“+1”的中英文短语对。其中的中文短语经过匹配约束（2.3.1）后得到候选缩略-全称语 12639 对。根据候选缩略语的字长统计情况如表 2。

表 2　候选缩略-全称语对统计

字长	1	2	3	4	5
候选集数目	2292	5772	1705	2333	533

（6）噪音过滤：对经(5)得到的候选缩略-全称语对采用 ICTCLAS 进行词性标注。对得到的带有词性信息的候选缩略-全称语对进行噪音过滤(2.3.2)。最终得到缩略-全称语 710 对。

3.2 实验结果和分析

经过 3.1 中的实验步骤，我们得到最终的缩略-全称语词典。表 3 显示的是按缩略语字长和组合方式给出的统计结果。结果显示提取的缩略语以 2 字长的居多，占到总数的 64%。字长为 4 和 5 的缩略语比较少。这一方面是我们提取过程中的偏向，另一方面是字长为 2 的缩略语在自然语言中分布确实很多。在缩略语构成方面，我们的方法偏向于语素构成方式，占总数的 71.83%。混合方式占 16.05%，而中心词构成方式产生的缩略语主要来源于专有名词，因而数量不多。另外，对于合并方式我们的算法没有考虑，原因是这类缩略语和全称短语的英文翻译往往不相同。

[1] http://projects.ldc.upenn.edu/TIDES/mt2003.html

[2] http://nlp.sttanford.edu/software/segmenter.shtml

表3 提取出的缩略语统计表

缩略语	语素构成方式	中心词构成方式	混合方式	总
字长1	143 (阿-阿根廷)	14 (江-江/主席)	0	157
字长2	348 (国安-国家/安全)	61 (清华-清华/大学)	46 (藏族-藏/民族)	455
字长3	19 (经贸委-经济/贸易/委员会)	5 (黑龙江-黑龙江/省)	42 (军事学-军事/科学)	66
字长4,5	0	6 (民主革命-民主/主义/革命)	26(地空/导弹-地对空/导弹)	32
总	510	86	114	710

关于缩略语对的质量，我们采用抽样检测的办法，随机抽取了100条缩略-全称语对考察。统计的结果如表4所示，准确率达到91%。其中错误的例子一部分是由分词和词性标注错误以及短语词对齐不精确造成，如“韩国/nsf - 韩三国/nr”及“美/b - 韩美/nr”。另外一部分则属于使用统计方法不可避免的误差，如“我军/n - 我国/n 军用/b”。

表4 缩略语词典的抽样检测

缩略语	正确	错误	正确率
字长为1：22	18	4	81.8%
字长为2：64	61	3	95.3%
字长为3, 4, 5：14	12	2	85.7%
总：100	91	9	91%

4 结论及未来工作

本文提出了一种从双语平行语料中提取缩略语词典的方法。与其他方法相比，我们利用了语言之间的翻译关系，获得较为可靠的候选集。需要的人工标注量很小，最终的缩略语词典正确率比较高。实验中，我们的噪音过滤方法使得一些好的缩略语被过滤掉。在今后的研究中我们将探寻更好的解决方法，比如用更多的信息，如短语的上下文特征[6][11]，来过滤候选集。

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] Xiaodan Zhu, Mu Li, Jianfeng Gao et al. Single Character Chinese Named Entity Recognition[C].In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, ACL, 2003.

[3] 李斌, 力芳. 中文单字国名简称的自动识别. 计算机工程与应用 2006, 42(28).

[4] 支流, 朱学锋, 段慧明等. 中文缩略语还原技术初探[C]. 全国第八届计算语言学联合学术会议(JSCL-2005).

[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] 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, 28(16).

[7] Zhifei Li and David Yarowsky. 2008. Unsupervised Translation Induction for Chinese Abbreviations using Monolingual Corpora. Proceedings of ACL 2008, pages 425-433.

[8] Philipp Koehn, Franz Joseph Och, and Daniel Marcu. Statistical Phrase-Based Translation. In Proceedings of HLT/NAACL. 2003.

[9] F.J.Och, C.Tillmann, H.Ney. 1999. Improved alignment models for statistical machine translation. In Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora 1999, pages 20-28.

[10] V.Vapnik and C.Cortes. 1995. Support vector networks. Machine Learning, 20, 273-293.

[11] Boxing Chen, George Foster, Roland Kuhn. 2010. Bilingual Sense Similarity for Statistical Machine Translation. In Proc. of the 48th Annual Meeting of the Association for Computational Linguistics, pages 834-843.