基于例句语料库的现代汉语方位词用法自动识别研究*

买志玉[1,2]，赵丹[2]，昝红英[2]，张坤丽[2]

1. 中原工学院软件学院，河南 郑州 450007

2. 郑州大学信息工程学院，河南 郑州 450001

Email:maizhiyu@yahoo.com.cn

摘 要：本文根据目前已有的方位词研究成果以及已构建的现代汉语方位词用法词典和用法规则库，初步完成了对现代汉语方位词例句语料库的基于规则的用法自动识别，通过对自动识别结果的分析，调整和完善了现代汉语方位词用法词典和规则库，使基于例句语料库的现代汉语方位词用法自动识别准确率从56.41%提高到81.32%，为现代汉语方位词用法的机器识别打下一定的数据基础。

关键词：方位词；用法知识库；用法词典；用法规则库；语料库；自动识别

Research On The Automatic Recognition Of Contemporary Chinese Locative Words Usage In Corpus Of Example Sentences

Mai Zhiyu[1,2], Zhao Dan[2], Zan Hongying[2], Zhang Kunli[2]

1. College of Software, Zhongyuan University of Technology, Zhengzhou Henan 450007

2. College of Information Engineering, Zhengzhou University, Zhengzhou Henan 450001

Email:maizhiyu@yahoo.com.cn

Abstract: This paper discusses the automatic recognition of contemporary Chinese locative words usage in corpus of example sentences based on previous research results and the construction of usage dictionary and usage rule base for Chinese locative words. Through the adjustment of Chinese locative words usage dictionary and usage rule base based on the results of the automatic recognition, the recognition rate in corpus of example sentences is increased from 56.41% to 81.32%, which provides foundation for automatic recognition of contemporary Chinese locative words usage.

Keywords: Chinese locative words; knowledge base; usage dictionary; usage rule base; usage corpus; automatic recognition

1 相关研究

北京大学的俞士汶教授在《现代汉语广义虚词知识库的建设》[1]一文中将“广义虚词”的范围界定为副词、介词、连词、助词、语气词和方位词。郑州大学自然语言处理实验室自2006年起承担了现代汉语虚词用法知识库和现代汉语虚词用法自动识别的研究任务，目前已完成了现代汉语方位词用法知识库中现代汉语方位词用法词典和用法规则库。本文的研究是在此基础上，在例句语料库的范围内，完成了基于规则的现代汉语方位词用法自动识别。通过对自动识别结果的分析，调整和完善现代汉语方位词用法词典和用法规则库，建立“三位一体”的现代汉语方位词

*本文相关研究得到了国家自然基金项目（项目号 60970083）、北京大学计算语言学教育部重点实验室开放课题基金（项目号 KLCL-1004）和河南省科技创新人才杰出青年基金项目（项目号 104100510026）的支持。

用法知识库，从而为现代汉语方位词用法的机器识别打下一定的数据基础。关于现代汉语方位词用法知识库的相关研究以及现代汉语方位词用法形式化描述详见《面向机器识别的现代汉语方位词用法形式化描述研究》[2]。

2 基于例句语料库的现代汉语方位词用法自动识别

现代汉语方位词用法的自动识别可以在例句语料库和《人民日报》分词及词性标注语料库进行，本文的研究只涉及例句语料库。通过标注方位词的用法，可以全面考察语料库中出现的方位词的用法，根据方位词在真实语料中的用法特征，修正和调整现代汉语方位词用法词典和现代汉语方位词用法规则库，使之逐步完善。

2.1 基于规则的现代汉语方位词用法自动识别

基于规则的现代汉语方位词用法的自动识别是根据方位词用法规则库中的用法规则描述，判断出现代汉语方位词例句语料库每个出现的方位词的用法，并把该用法编码标注在该方位词的词性标注之后[3]。

例如，方位词“北”的一组用法规则描述如下：

$北

@<f_bei3_2>→R ^R→豆腐|味|货

@<f_bei3_1b>→L ^L→j|n

@<f_bei3_1c>→R ^R→n

@<f_bei3_1d>→M ^M→在|往|朝|向|自|从|由|p

@<f_bei3_1a>→

下面是方位词“北”在例句语料库中用法自动识别的正确样例：

{f_bei3_1d 北 f}自/p **北/f<f_bei3_1d>** 向/p 南/f 3/m 个/q 灯饰/n 造/v 景/Ng 巧夺天工/i 。

注：“{ }”中为正确结果，“<>”中为自动识别结果。

标注的结果表明，该语料中的方位词“北”的用法编码为<f_bei3_1d>。

基于规则的现代汉语方位词自动识别是使用郑州大学自然语言处理实验室开发的“基于规则的虚词用法自动标注系统”[3]实现的。该系统能够根据现代汉语方位词用法规则库为语料中每个出现的方位词标注上正确的用法编码 ID，并且对标注的结果进行统计，例如计算出识别的准确率，统计出方位词分词错误或其它错误。

2.2 自动识别过程的改进工作

（1）关于时间词的形式化表示

方位词经常与表示时间的词连用，表示时间。表示时间的词常见的有时间词和量词，例如“四月里”、“三年里”，用法规则表示为“^L→t|q”。但是还有很多表示时间的名词，并没有包含在用法规则中。例如“这个星期里几乎天天下雨”中，“星期”是名词，与“里”连用表示时间。如果把名词加入用法规则，则用法规则的覆盖面过大，因为名词与方位词连用也可表示处所或范围，例如“教室里”、“人群里”。但是如果不把这些表示时间的名词加入到用法规则中，则该用

法规则的识别率会非常低。因此，将表示时间的名词以列举的方式加入到用法规则，用法规则调整为“^L→t|q|时间|小时|日|月|星期|季度|年|年头”。以“里”为例，识别正确率由 55.56%提高到 88.89%。

这里采用列举的方式没有完全包含所有可能出现的表示时间的名词，这就留待以后遇到具体的例句再进行补充。

（2）调整现代汉语方位词用法词典中某些方位词的用法

方位词用法词典中方位词的用法描述以吕叔湘《现代汉语八百词》[4]为主，同时参考了张斌《现代汉语虚词词典》[5]、俞士汶、朱学峰等《现代汉语语法信息词典详解》（第二版）[6]、《现代汉语词典》（第 5 版）[7]、《人民日报》分词及词性标注语料等资料。当几种参考资料对方位词用法描述有冲突时，必须对方位词的用法进行调整。

例如，方位词用法词典中方位词“上”的其中一种用法为“名词+上，指物体的顶部或表面”，例如“房子上”、“桌子上”，这是根据《现代汉语八百词》的描述来定义的，而在《人民日报》分词及词性标注语料中把这种用法归为方位词“上{shang}”（轻声）的用法，经过讨论决定，把这种用法从方位词“上”的用法调整到“上{shang}”的用法中。

（3）调整现代汉语方位词用法词典中的例句。

现代汉语方位词用法词典所举的例句中，有些方位词已经和其它的字组合成词，不能作为方位词单独识别出来，这样的例句必须被删除。例如，“路旁都种上了树”，分词软件标注该例句的结果为“路旁/s 都/d 种/v 上/v 了/u 树/n”，这说明“路旁”已经独立成词，该例句已不能担当方位词“旁”的例句，必须将其删除。

原先的例句语料库中共有 1025 条例句，经过对每个方位词补充大量的例句后，语料库中的例句总数达到了 2452。语料库规模的扩大，使得在考察方位词用法规则时能够更全面，更系统，所得到的自动标注结果更可靠。

（4）在对规则进行考察的过程中，不断调整每个方位词所对应的一组规则的排序问题。

每个方位词的多种用法对应了一组规则，选择特征描述操作性较强的特征作为高优先级进行排序[8]。一组规则的排序问题会极大地影响到自动识别的准确率，这实际上是自动识别过程中重点要完成的工作。

例如，方位词“北”共有 2 个义项 5 种用法，第一个义项表示“四个主要方向之一，清晨面对太阳时左手的一边”，其第三种用法为“北+名词”，用法编码为<f_bei3_1c>。第二个义项表示“北部地区，在我国通常指黄河流域及其以北的地区”，其用法为“北+名词，名词仅限例句中列举的名词”，用法编码为<f_bei3_2>。方位词“北”上述两种用法规则及排序原先为：

@<f_bei3_1c>→R ^R→n

@<f_bei3_2>→R ^R→豆腐|味|货

经过考察，发现用法<f_bei3_2>的特征描述操作性比用法<f_bei3_1c>要强得多，所以将两种用法规则的排列顺序进行了前后对调，同理，对方位词“北”其它用法规则的排序也进行了调整，调整后，除去分词错误，方位词“北”的识别率由原来的 31.25%提高到了 100%。

（5）对常用方位词的用法进行了细分。

原先的现代汉语方位词用法词典中一般是一个义项下的所有用法都写在同一条用法规则中，这样做使得自动识别的准确率较高。但是考虑到现代汉语方位词用法知识库对自然语言处理的应用，这种用法规则描述方式并不能提供足够的有价值的信息，所以对于常用的方位词，还是对其

用法进行了细分。

例如方位词“北”，其第一个义项为“四个主要方向之一，清晨面对太阳时左手的一边。”，其用法编码为“f_bei3_1”，用法描述为“单用。|名词+~。|~+名词。|介词+~。”(|表示并列)。仿照《现代汉语八百词》的描述风格对“北”的用法进行细分，描述如下：

f_bei3_1a 单用。

f_bei3_1b 名词+~。

f_bei3_1c ~+名词。

f_bei3_1d 介词+~。

对应的用法规则也由 1 条调整为 4 条。这样做的好处会在今后自然语言处理的深层次的应用中逐渐体现其价值。

通过对以上 5 个问题的改进工作，基于规则的现代汉语方位词的自动识别准确率已经由原先的 56.41%提高到 81.32%，准确率有了明显的提高。见表 1。

表 1　针对例句语料库改进前后的自动识别结果对比

	改进前	改进后
测试例句总数	1025	2452
含有分词错误的例句数	245	16
去掉分词错误和没有规则的例句总数	780	2436
测试正确例句数	440	1981
方位词自动识别准确率	56.41%	81.32%

2.3 自动识别过程存在的问题

目前对于现代汉语方位词用法的自动识别还存在以下的问题：

（1）有些方位词具有实指和虚指两个义项，人来理解时非常容易区分，但是由于用法规则基本一致，自动识别时无法区分。

例如方位词“背后”有两个义项：后面（实指）；后面（虚指），有引申义。所对应的例句为：

- 背后的吊脚楼远远地伸向河面<r>
- 这样，背后的动因就相当清楚了：如果没有竞争，谁也不会把可能得到的那一份利润拱手让人。<r>

两个例句中的“背后”后面紧跟的都是名词，用法上是一样的，只是意思上分别表示实指和虚指，通过现有的用法规则描述是无法将两个义项分开的。通过分析发现，“背后”的实指和虚指实际上是通过后面所跟是具体物体名词还是抽象名词来决定的，所以考虑将抽象名词列举出来，放入用法规则中。当然抽象名词不可能一下子列举完整，可以通过不断积累不断补充，使得自动识别的准确率提高。

（2）方位词有两种用法描述为“名词+方位词”和“方位词+名词”，所对应的用法规则为“L→n”和“R→n”，但是在例句中，方位词前面或后面并没有紧跟名词，而是被一些修饰名词的修饰语、连词或者其它方位词把方位词和名词间隔开，这样就造成该规则无法被识别。对这种用法的规则描述要进行进一步的考察确定。

（3）有些方位词用法区分明显，但是总结成用法规则不好区分，例如方位词前面可以加表示处所的名词，也可以加表示范围的名词，这两种用法是不同的，但是总结成用法规则都一样，

即“L→n”，在自动识别过程中出现识别错误的几率很高。可以考虑用基于统计的方法提高自动识别的准确率。

（4）有些方位词有多个义项，但是每个义项所对应的用法是完全一致的，对于这样的情况是最难解决的。要想对这种情况进行准确的形式化描述，还需要对语法分析进行更深层的挖掘，寻找出适当的规则描述，使机器更准确地识别出复杂语句中方位词的不同用法。同时也可以考虑在自动识别过程中使用基于统计的方法来提高识别率。

3 进一步工作展望

下一步的工作主要是：（1）借鉴基于例句语料库的现代汉语方位词用法自动识别的经验，针对1998年1月和2000年1月～6月《人民日报》分词及词性标注语料进行基于规则的现代汉语方位词用法自动识别。《人民日报》分词及词性标注语料的自动识别结果没有正确依据可参照，不能通过自动识别软件来统计自动识别的准确率，必须依靠人工逐句校对来判断每个方位词的用法标注是否正确。为避免人工对方位词用法理解的主观性和片面性，校对工作由多人分别进行，根据汇总意见，解决自动识别过程中发现的问题，不断调整和完善现代汉语方位词用法词典以及用法规则库。（2）根据校对后的《人民日报》分词及词性标注语料中方位词用法的自动识别结果，利用隐马尔科夫模型（Hidden Markov Model, HMM）、支持向量机（Support Vector Machine, SVM）、最大熵（Maximum Entropy, ME）以及条件随机场（Conditional Random Fields, CRF）等机器学习的统计模型进行学习，实现基于统计的现代汉语方位词用法自动识别。

参 考 文 献

[1] 俞士汶,朱学锋,刘云.现代汉语广义虚词知识库的建设[J].汉语语言与计算学报, 2003,1:89~98.

[2] 赵丹,张坤丽,昝红英,等.面向机器识别的现代汉语方位词用法形式化描述研究[C].第十一届汉语词汇语义学研讨会（CLSW2010）论文集.苏州:苏州大学,2010.

[3] 袁应成,昝红英,张坤丽,等.基于规则的虚词用法自动标注算法设计与系统实现[C]，第十一届汉语词汇语义学研讨会（CLSW2010）论文集.苏州:苏州大学,2010.

[4] 吕叔湘.现代汉语八百词[M].北京:商务印书馆,1980

[5] 张斌.现代汉语虚词词典[M].北京:商务印书馆,2003.

[6] 俞士汶,朱学锋,王惠,等.现代汉语语法信息词典详解(第二版) [M].北京:清华大学出版社,2003.

[7] 中国社会科学院语言研究所词典编辑室编.现代汉语词典（第5版）[M].北京：商务印书馆,2005.

[8] 昝红英,朱学锋. 面向自然语言处理的汉语虚词研究与广义虚词知识库构建[J]. 当代语言学,2009(2):124-135