汉语空间关系中射体识别问题的研究与分析

赵纪元，李晗静，赵铁军

（哈尔滨工业大学语言语音教育部-微软重点实验室，哈尔滨 150001）

摘 要： 本文利用了在汉语的空间关系表达中射体的概念，结合语料和统计数据详细分析了射体的语法、语义、结构特点以及特殊用法。在此基础上提出了基于 Winnow 算法的射体识别策略，并结合射体的语言特点，给出了一套较为完整的特征方案。实验结果显示，该方法封闭测试 F 测度可达 63.16%以上。同时，在对相同特征描述的基础上，进行了一组基于规则的射体识别实验，其结果与分类算法结果相差不大。可见，在特征不变的情况下，采用不同的方法进行射体识别，性能相差并不大。因此，进一步提升识别效果的关键是要不断寻找更有效的特征。

关键词： 射体识别；空间关系；Winnow 分类

Research and Analysis on Trajectory Recognition in Chinese Spatial Expression

Zhao Jiyuan, Li Hanjing, Zhao tiejun

(Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract: This paper presents the concept of trajectory in Chinese spatial expression, and analyzes its syntax, semantic and special usage, on the base of the corpus and statistic data. A Winnow algorithm was proposed, whose features were designed according to the character of trajectory. Experimental results indicated that the F-Score in close-test could reach 63.16%. At the same time, we designed a group of rules on the base of the same features above, and did experiments with them. The results get little difference from the Winnow algorithm. Thereby, we can conclude that when the same features are used, the performance of different methods is similar. So we should focus on finding better features in the next step to improve the performance.

Keywords: trajectory recognition; spatial relation; Winnow

1 引言

语言是人类进行信息交流与沟通的重要途径之一。当面对自然语言对现实物理空间的描述时，人类一定会在脑海中形成相应的空间场景，以此来帮助交流。人脑是如何轻松地完成这样的转换呢？这个问题的回答非常有助于理解人类认知机制。如果利用计算机模型来理解、模拟和仿真这种转换，那么将具有深远的意义。

解决这一问题的首要步骤，就是要深入理解自然语言对人类认知空间的表达方式，以及从中抽取出有关于空间关系的知识[1]。提到空间关系的表达，就必然想到描述的主体和参照物。在研究中，我们对空间关系的主体给出定义，即“射体”。充分分析和识别自然语言中的射体信息，对于完整的理解空间关系，以及分析人们的认知心理有着极其重要的意义。

基金资助：国家自然科学基金项目（60575041、60373101）

作者简介：赵纪元（1982-），女，黑龙江，硕士，jyzhao@mtlab.hit.edu.cn

本文正是从以上目的出发，陈述了汉语空间关系理解中的关键问题——射体识别。文章首先从射体的概念出发，结合语料和统计数据，全面分析了射体的语法，语义，特殊语言现象，以及它在空间关系表达中的作用。同时，给出了在汉语语料中手工标注射体的规则。

在以上研究的基础上，把射体识别问题抽象成为一个二值分类问题，提出了基于 Winnow 算法的射体识别策略。射体识别的难度在于如何使语义、甚至语用可计算。传统的方法通常需要人工构建知识库，不便于扩展。而 Winnow 算法具有特征综合能力强、效率较高等优点[2]。

分类问题中，特征的选择极其重要，选择区分度好的特征对于取得满意的识别效果有着决定性的意义。本文结合射体的语言特点，提出了一整套比较全面的特征构建方案，并以此为基础进行了射体识别实验。最后，详细分析了实验的结果，并与基于规则的识别方法相比较，提出了下一步特征修订的方案。

2 射体及相关概念

先来看一组句子：

S1：杯子在桌子上。

S2：墙上有一幅画。

S3：房子前面有一棵树和一口井。

以上的每句话都表达了一种空间关系。例如 S1 所给的场景中涉及两个物体“杯子”和“桌子”，方位词“上”和介词“在”说明了杯子与桌子的空间位置关系。同时，容易发现叙述者的观察焦点集中在“杯子”上，即陈述的主体是“杯子”，而“桌子”只是用来表明“杯子”位置的一个参照物。同理，S2 说明了“画”和“墙”的关系，焦点物是“画”，参照物是“墙”。S3 和前两个句子稍有不同，它的场景中包括三个物体“房子”“树”和“井”，这时焦点物就有两个“树”和“井”，参照物是“房子”。这样的句子在汉语中频繁出现，它们体现了描述空间关系的四个要素——焦点物、参照物、方位词和介词。

我们沿用 Langacker 提出的认知语法，把焦点物定义为“射体”（trajectory），简写为 TR[3]。从以上的例子不难看出，射体是空间构造关系的主体，是叙述者要陈述的对象，也是人们观察的焦点所在。焦点物在语义上有以下特点：1）面积或体积较小；2）空间上不固定；3）时间上不持久；4）结构较简单；5）对说话双方来说它的位置或方向是不知道或不能确定的[4]。从语法上讲，充当射体的词语一般为名词或代词。

参照物为空间关系中焦点物的方位确定提供参考，在研究中我们称其为“界标”（Landmark），简写为 LM。界标可以是一个预先存在的物体，或物体的集合[5]。它具有以下特点：1）面积或体积较大；2）空间上相对固定；3）时间上相对持久；4）结构较复杂；5）对说话双方来说他的位置或方向是已知的或可以确定的[4]。充当界标的词语一般为名词。

我们把具有方位词、介词和界标这三个要素的，表示空间关系的短语或词串，从原文中提取出来，称作“空间表达式”（spatial expression），简写为 SE，以一个元组的形式列出，即 SE=（介词，方位词，LM），其中的某一项可以为空[3]。一般一个空间表达式只对应一个界标，但对应的射体数却可以是多个。

3 射体和空间表达式的标注

由于目前国际上没有针对汉语空间关系的语料库，因此我们对中文版《伊索寓言》10 卷 432 篇进行标注。从中随机选择 325 篇作为训练语料，其余的作为测试语料。以下研究和实验都是以这个语料库为基础的。为了研究语料中的空间关系，还要在句子切分、词性标注和指代消解的基础上对空间表达式、界标和射体进行标注。其中，射体的标注规则具体如下：

(1)射体的标记为“[TRi]”，标注在充当射体的词语后面，i 是该射体对应的空间表达式编号；

(2)只标注表示客观物理空间关系的空间表达式和射体，具有隐喻意义的不标注；

(3)如果一个空间表达式对应多个射体，则这些射体标号相同，并在文中依次标出；

(4)在为指代消解的对象标注 TR 时，应标在方括号内，紧跟在作射体的词语之后，如果有多个射体，依次标出；

(5)如果一个空间表达式的同一个射体的名称在文中多次出现，只标注与该空间表达式距离最近的射体；

(6)尽量标在名词或代词的后面，遇到形如“兄妹/nc[TR1][TR2] 俩/m”“渔夫/nc[TR1][TR2] 们/k”这样有后接成份的名词时，应把射体标在相应的名词之后。

4 射体与空间表达式的关系

4.1 射体和空间表达式的数量对应关系

由前面的例子可知，射体和空间表达式的对应关系不一定是 1：1 的，如 S1、S2 的空间表达式对应一个射体，S3 的空间表达式对应 2 个射体。在对语料统计的过程中，我们发现射体和空间表达式的对应关系包括 0：1，1：1 和 n：1（n>1）三种，结果如下：

表 1. 射体和空间表达式数量对应关系

	1:1		2:1		3:1		射体总数
	射体数	占比例	射体数	占比例	射体数	占比例	
训练语料	539	87.64%	58	9.43%	18	2.93%	615
测试语料	203	90.63%	18	8.04%	3	1.34%	224

1：1 和 n：1 的情况比较好理解，从语义上讲就是同一个物体作为一个或多个物体的方位参考。但为什么会出现 0 射体的情况呢？经研究发现，射体和空间表达式 0：1 时，射体多为替代或省略的情况，如汉语中的“的字结构”等。

4.2 射体和空间表达式的位置关系

射体和空间表达式不但在对应关系上有规律，它们在上下文中的相对位置分布也很有规律，多数情况下，射体和它所对应的空间表达式分布在同一个子句内。具体分布如下：

表 2. 射体与空间表达式的位置关系

	同一子句		同一句子不同子句		同一段内不同句子		不同段内		射体总数
	射体数	占比例	射体数	占比例	射体数	占比例	射体数	占比例	
训练语料	439	71.38%	158	25.69%	18	2.93%	0	0%	615
测试语料	159	70.98%	63	28.13%	2	0.89%	0	0%	224

4.3 射体的特殊用法

由表 2 可知，通常情况下，射体距离它所对应的空间表达式较近，但也有少数比例的射体距离空间表达式较远，例如不在同一句子里，这是由射体的特殊用法所造成的，主要包括两方面：射体的省略和空间焦点的转移。

一、射体的省略

射体的省略是指：该射体为文中一直在叙述的主体，它在前文中已经被提过，但在接下来的句子（或子句）里没有对该射体进行重复的说明，就出现了新的空间表达式，以至于在新的空间表达式和前一空间表达式之间找不到射体。根据射体充当的语法成份的不同，把射体省略分为两种情况：主语省略和宾语省略。

对语料中，射体不在相邻两个空间表达式之间的情况进行研究，发现训练语料中射体省略的情况占了 87.5%，这其中全部是主语省略。测试语料中射体省略占了 86.67%，其中 80%为主语省略，6.67%为宾语省略。可见，射体省略的绝大多数为主语省略。

二、空间焦点的转移

空间焦点的转移是指：在行文的过程中，人们关注的空间焦点从一个事物转换到了另一个事物，也可以说是在对主体的叙述中插入对其他物体的描述。直观上，类似电影中镜头的来回切换。分析发现，射体不在相邻两个空间表达式之间时，在训练语料中有 12.5%是空间焦点转移的情况，测试语料中则有 13.33%。

5 基于 Winnow 算法的射体识别

在对射体充分理解的基础上，我们把射体识别问题抽象成为一个二值分类问题，设计了基于 Winnow 算法的

射体识别实验。本文选择 SNoW(Sparse Network of linear separators)作为分类器，它是基于 Winnow 算法的线性分类器，文献 6 提出 SNoW 体系结构。

5.1 实验的处理对象

首先给出以下定义：

(1)SE：等待与当前射体样本匹配的空间表达式，也叫待匹配空间表达式。

(2)SEB：前一相邻句子中，距离 SE 最近的空间表达式。

(3)SEN：后一相邻句子中，距离 SE 最近的空间表达式。

本实验的处理对象为：文章中 SEB 与 SE 之间以及 SE 与 SEN 之间的所有词语。

5.2 特征空间的构造

使用基于特征向量的机器学习算法，最重要的过程是特征的选取。我们结合射体本身的语言特点，提出如下的特征空间：

(1)词性特征：从经过词性标注的文本中，把待处理词语对应的词性抽取出来，作为特征。

(2)动词特征：将待处理词语与其所在的句子中每个动词的距离作为一类特征。该距离记为 Lverb，分为以下四种情况：该动词与待处理词语在同一个子句；在同一句子，不同子句；在同一段落，不同句子；在不同段落。此外，上述动词的词性也为一类特征。

(3)“把被”特征：考查待处理词语所在子句中是否出现“把”字（或“被”字），分别用一个特征表示。该词语与其所在子句中“把”字（或“被”字）的距离为一类特征。

(4)语义特征：Hownet[7]是表示中文语义关系权威且有效的工具，这里利用 Hownet2005，统计小样本，得到上位义原集合 *sym*{物质，群体，部分}。集合 *sym* 中的每个元素就是一个语义特征。

(5)空间表达式特征：分别考查待处理词语与 SEB，SE 和 SEN 之间的距离，并将其作为一类特征。以上每类距离又分为：该空间表达式与待处理词语在同一个子句；在同一句子，不同子句；在同一段落，不同句子；在不同段落，四种情况。

5.3 实验结果及评价

首先确定 Winnow 分类器的参数，α、β、和阈值 θ。基于枚举法，在训练语料上，使用包含 5.2 节阐述的所有类别特征的特征空间，取 F 测度值的最大值对应的参数值，α 是 1.7，β 是 0.5，θ 是 2。

在最优参数下，使用 SNoW 在训练语料和测试语料上分别进行测试，统计了实验的整体识别结果，并把射体按照表格 1 和表格 2 的分类，统计了每类射体识别的召回率。具体如下：

表 3. 射体识别整体结果

	F（%）	精确率（%）	召回率（%）
训练语料	63.16	59.48	67.32
测试语料	59.34	55.43	63.84

表 4. 各类射体的召回率

	射体类型	召回率（%）	射体类型	召回率（%）
训练语料	1:1	70.32	同一子句	85.19
	2:1	50	同句不同子句	26.58
	3:1	33.33	不同句	0
测试语料	1:1	64.04	同一子句	82.39
	2:1	61.11	同句不同子句	19.05
	3:1	66.66	不同句	0

同时，我们利用 5.3 节提到的特征空间，设计了基于规则的射体识别实验。用规则的方法描述了射体的五大类特征，并定义了识别的优先顺序[2]。实验结果如下：

表 5. 基于规则的射体识别结果

	F（%）	精确率（%）	召回率（%）
训练语料	60.03	60.23	59.84
测试语料	61.54	62.38	60.71

可见，基于规则的方法与线性分类的方法结果相差并不大，这主要是由于他们使用了相同的特征。因此，提升射体识别问题效果的关键在于不断增加有效的特征，逐步完善特征空间。为此，我们仔细考察了 SNoW 生成的分类网络中，各类特征的权重，以及它们在分类中所起的作用。得出以下结论：

(1)词性特征普遍权重较高。其中名词类词性对正例的识别很有效。其余的词性对反例的识别很有效，例如，形容词、副词、动词和助词。

(2)空间表达式特征对分类也很有效。其中，射体与其对应的空间表达式之间的距离特征是正例识别中权重最高的特征，对分类极其重要。我们曾作过实验，发现去掉空间表达式特征后，实验的结果下降很明显，F 测度仅有 19.35%，这也验证了以上结论。

(3)动词特征和“把被”特征，由于它们都是考察目标词和其他词的距离关系，所以在分类的作用上体现出共同的特点。即距离数较小的特征使用频率较高，权重也较高，距离数较大的特征作用很小。这是由于对距离的取值范围定义过大，从而造成了特征空间的数据稀疏。

(4)语义特征中有一部分对正例识别有一定作用，但不明显。这是由于我们只用了 Hownet 中的最基础的语义，并没有深入展开并充分利用 Hownet 的语义空间。

可以发现，射体识别的特征空间还可以进一步扩充。在对射体充分理解的基础上，结合以上分析结果初步提出以下方案：

(1)该词语是否在前文充当过射体，考查该词语上次出现的位置与当前位置之间的距离；

(2)该词语与空间关系中介词、方位词以及界标的位置关系；

(3)修改动词特征距离范围；

(4)修改“把被”特征距离范围；

(5)充分利用 Hownet，给出全面的语义特征。

6 结论

本文利用了汉语空间关系表达中射体的概念，结合语料详细分析了射体的结构、语义和语用。并在此基础上提出了基于 Winnow 线性分类的射体识别策略，给出了一套较为完整的特征空间，考察了各类射体的识别结果。最后，把分类方法与规则方法的实验结果进行了比较，发现在特征不变的情况下，识别效果相差并不大。但基于分类的方法有自动化程度高，便于统计结果等优点，所以在对算法的效率要求较高的情况下，应尽量考虑这种策略。

另外，特征的选择对提高射体识别问题的效果有着重要的意义，所以，我们下一步工作将集中精力寻找更好的特征，构造更丰富的特征向量，不断改善射体识别的性能。

[注]本文的研究和撰写过程中，得到了韩延海、叶利军、李理、李世奇等人的大力协助，特此致谢！

参考文献：

[1] 方经民.汉语空间方位参照的认知结构.世界汉语教学,1999，4 期:32-33

[2] 李晓静,李生,赵铁军.基于自然语言理解的空间内实体自动摆放的研究. 电子与信息技术，2006

[3] 李晓静,李生,赵铁军.汉语中方位参考点恢复研究. 计算机研究与发展，2006 年，43 卷

[4] 刘宁生.汉语怎样表达物体的空间关系.中国语文.1994.3:170-171

[5] Sharon Rose Clay,Jane Wilhelms. Put:Language-Based Interactive Manipulation of Objects.IEEE,1996:31-39

[6] Littlestone. N. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine Learning. 1988.2:285–318

[7] 董振东,董强. 知网. http://www.keenage.com.1999