

童话故事概念关系的自动构建

吴敬桐, 陈玉泉

(上海交通大学计算机科学与工程系, 上海 200240)

摘 要: 提出一种主次模板方法, 构建面向童话故事的概念关系。对于主模板的构建, 引入部分句法分析的模式; 对于次模板的构建, 引入生物信息学中序列比对的方法, 同时在迭代过程中, 加入模板置信评价机制。对童话故事内容进行抽取概念对的实验, 结果表明, 采用主次模板方法具有较高的正确率。

关键词: 童话故事概念关系; 关系自动构建; 主次模板方法; 交互式动画

Automatic Construction of Fairy Tale Conceptual Relation

WU Jing-tong, CHEN Yu-quan

(Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

【Abstract】 This paper proposes a primary and secondary template method to construct the conceptual relation of fairy tale. To the primary template, it introduces a partial syntactic analysis model. To the secondary template, it introduces the bioinformatics sequence alignment method. At the same time in the iteration process, it introduces the template confidence evaluation mechanisms. It extracts the concept pair of the fairy tale, experimental results show that the primary and secondary template method has high precision rate.

【Key words】 fairy tale conceptual relation; automatic construction of relation; primary and secondary template method; interactive animation

DOI: 10.3969/j.issn.1000-3428.2011.15.0041

1 概述

在交互式高效动画研究中, 需要运用自然语言处理技术从童话故事文本中提取动画创作所需的情节信息, 用于辅助导演分镜头剧本创作以及交互式的动画创作。在提取童话故事情节信息时, 童话故事概念关系尤为重要。在分镜头剧本中的场景角色建模中, 需要自动构建童话故事中的概念关系。此类构建主要以 DIRPE^[1]方法为主, 但是单纯采用 DIRPE 方法构建模板, 在精度和覆盖范围上都无法满足需求。因此, 本文提出一种主次模板方法, 用于构建面向童话故事的概念关系。

2 背景介绍

交互式高效动画框架如图 1 所示。

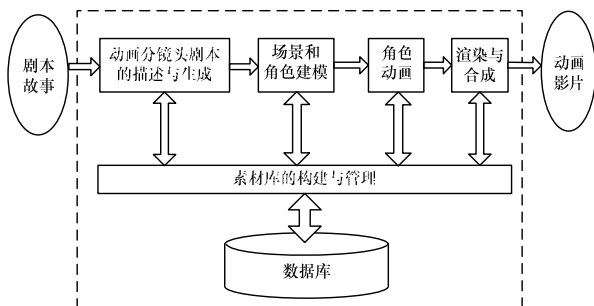


图 1 交互式高效动画框架

通过对剧本故事自然语言处理, 生成导演分镜头剧本, 再对分镜头剧本继续分析处理后, 进行场景和角色的建模, 以供后续的角色动画生成。其中在场景和角色建模中, 需要对故事集进行基础分析和场景分析。基础分析完成对故事集切分标注以及句法分析。场景分析抽取分镜头中的场景信息、角色信息、角色的动作信息, 以及角色与场景的关系信息。其中角色与场景之间的关系就需要通过概念关系的自动构建

来完成。

相对于文本分类中的特征抽取^[2], 概念关系的抽取有其特点。目前, 对于概念关系的构建主要有 2 种方式:

(1) 采用一个量化的特征来表征 2 个概念之间的关系。例如: 用于检索的依存模型中, 往往只是用强弱^[3]一个简单的分类来定义 2 个概念之间的关系, 或把强弱关系用一个量化的距离来表征, 但是并没有给出具体的概念关系。

(2) 由于存在概念关系构造笼统性的问题, 一些学者提出了基于明确概念关系的信息检索^[4], 即概念关系构建的另一种方式。如果单纯采用人工方式构建概念关系, 虽在精度和正确性上可以得到一定保证, 但其工作量相当大。

在概念关系自动建立的过程中, 文献[1]提出了从大量的数据集中抽取一个结构化的关系, 而这种关系又是用满足关系定义的元组来表示。DIRPE 的理论假设是模板和关系之间存在着对偶性: 通过大量训练数据反复迭代得到一个有很高正确率的模板集合, 来把文本中满足关系的元组抽取出来。文献[5]在 DIPRE 上改进了训练过程。文献[6]在抽取满足某种关系的概念对上也做了尝试。但是 DIRPE 对于模板的定义过于简单, 导致能够抽取的正确的概念对和概念模板的数量都偏低, 因此, 还是存在一定的改进空间。

3 主次模板方法

3.1 DIRPE 方法与主次模板方法的流程

DIRPE 方法的流程如图 2 所示。

基金项目: 国家“863”计划基金资助项目“交互式高效动画创作技术与系统”(2009AA01Z334); 国家自然科学基金资助项目“面向网页检索应用的汉语语义概念图表示方法研究”(60873135)

作者简介: 吴敬桐(1986—), 男, 硕士, 主研方向: 自然语言处理; 陈玉泉, 副教授

收稿日期: 2011-01-25 **E-mail:** wangzhenlingwu@163.com

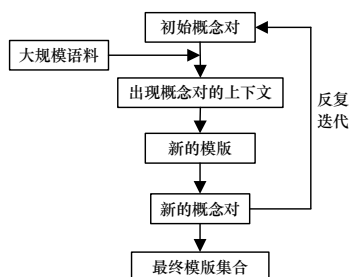


图2 DIRPE 方法的流程

主要步骤如下：(1)提供初始概念对。(2)从大规模语料中筛选含有概念对的候选句子。(3)通过两两候选句子的对比抽取公共部分形成新的模板。(4)通过新的模板从大规模语料中继续抽取新的概念对。(5)反复迭代训练生成最终模板集合。

主次模板方法的流程如图3所示。

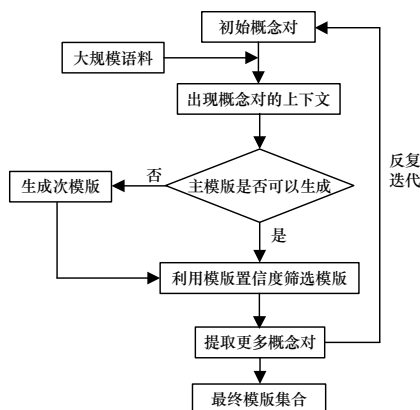


图3 主次模板方法的流程

主要步骤如下：(1)提供初始概念对。(2)从大规模语料中筛选含有概念对的候选句子。(3)采用主模板的形式处理候选句子。(4)在主模板失效的情况下，采用次模板的形式处理候选句子。(5)利用模板置信度评价机制筛选模板。(6)利用新模板抽取新的概念对。(7)反复迭代生成最终模板集合。

3.2 主模板的定义及生成

针对 DIRPE 模板定义的不足，所导致的抽取概念对和概念模板数量偏低的问题，本文提出了主次模板方法。对于主模板的构建，引入部分句法分析的模式，可用于保证单个模板的可理解性和正确率。对于次模板的构建，引入生物信息序列比对的方法，抽取含有相同关键词的句子中的公共部分。

通过对大量童话故事有特定关系的概念对的分析，发现由于童话故事特有的语言特点，往往概念对会在同一个分句中以一定特点的句法结构出现。例如：实验中所查找的“包含”关系中，有一个概念对为“农场”和“菜园”，在大规模童话故事集中，发现“农场”和“菜园”经常在这样一个短句中出现，“在农场的菜园里”。从中不难发现农场和菜园是在一个介词短语中出现，介词短语为：在……里。在后面继续的分析中，发现含有包含关系的概念对，经常会出现这样的介词短语结构。例如：“旅店”和“房间”，“小镇”和“商店”。因此，可以把句法结构的一定特征作为模板信息保存起来进行比对。又如：在实验中抽取“居住地(所属地)关系”中，发现在概念对出现的上下文短句中，出现的句法结构和前面的“包含”关系有明显的不同。比如：以“青蛙”和“池塘”概念对为例，出现的短句为“居住在池塘的青蛙”，可见其句法结构为：一个动词短语(居住在池塘)然后修饰一个名词(青蛙)，其中一个词在动词短语中，一个被动词短语修饰，

并且紧邻动词短语。从中不难发现也表现出一定的句法特征。

因此，主次模板方法在句法分析的基础上，利用在句法分析中出现的特征信息，总结概念对与特征信息的位置，定义了主模板结构：

<phrase, detail, firstIn, secondIn, order, confidence>

其中，主模板是由6部分组成；phrase指分句中出现的短语类型；detail指具体的短语结构；firstIn指判断概念对中第一个词是否在前面的短语结构中，如果在为1否则为0；secondIn指判断概念对中第2个词是否在前面的短语结构中，如果在为1否则为0；order表明词在概念对的前后顺序，第1个词在前order为1否则为0；confidence指模板的置信度。

主模板是从大规模童话故事集中，包含种子概念对的上下文获得，步骤如下：

(1)从大规模语料中查找含有种子概念对内容的短句，并且概念对的2个词之间的距离不超过5，2个词之间的距离定义为2个词之间间隔的字的数量。例如以“在农场的菜园里”这个短句为例，概念对的2个词“农场”和“菜园”之间的距离为1。

(2)对于含有概念对并且符合要求的短句做句法分析。

(3)对于分析的结果填充主模板。

例如：以包含关系为例：概念对为农场和菜园，出现的短句为“在农场的菜园里”。模板的形式为：

<介词短语，在……里，1, 1, 1, 0>

其中，phrase表示介词短语；detail为“在……里”；firstIn为1；secondIn为1；order为1；Confidence为0(置信度默认是0，后面会评价置信度的值)。

3.3 单纯采用主模板的问题

从上面主模板定义的过程中看出，主模板对于结构的依赖性相对较大，而且文中概念对之间必须存在相邻的句法结构，或者概念对存在于一个句法结构中。好处是可以在很大程度上限制模板的精度，但是由于主模板对句法结构要求严格，会导致候选模板的数量下降，从而导致单纯采用主模板形式抽取概念关系的覆盖率下降。因此，对于分句中含有概念对，但不符合主模板结构的候选分句，在一定程度上也可以挖掘模板信息，达到有效抽取的目的。在主模板的基础上定义了次模板，在主模板失效的条件下，可以采用次模板的形式进行抽取，从而保证抽取概念关系的覆盖率。

3.4 次模板的定义及生成

次模板以一个五元组来表示，具体形式为：

<before, middle, after, order, confidence>

其中，before为包含概念关系的句子中第1个种子词前面的内容；middle为包含概念关系的句子中2个种子词之间的内容；after为包含概念关系的句子中第2个种子词后面的内容；order最后表明词在概念对的前后顺序，正序关系为1否则为0；Confidence为模板的置信度。

以“包含”关系为例：(城镇，街道)作为一个包含关系的种子。其中的一个句子为：“位于西西王国西边的一个城镇中包含很多错落有致的街道”，模板形式为：

<位于西西王国西边的一个，中包含很多错落有致的，NULL, 1, 0>

以大规模童话故事集为基础，次模板的形成基本步骤如下：

(1)对任意2个句子模板中的内容进行词的切分。

(2)运用生物信息中序列比对的方法对任意2个句子的before, middle, after3个部分的内容进行匹配找出公共的

内容。

(3)采用规则筛选，去除噪声信息。

(4)以模板的形式保存 before, middle, after3 个部分公共内容。

这种生成模板的方法和通常模板生成方法有很大的不同。在通常模板生成方法中，模板的组成部分是多个上下文的最长公共字符串，每一个种子的所有上下文只生成一个由向量表示的模板，与这个模板的匹配需要人为定义一个经验阈值，这使得最终得到的模板召回率和正确率很难同时保证。

生物信息序列比对的方法为：在词串序列中任意位置插入空格，使得 2 个序列长度相同，形成词与词之间或词与空格之间的对应。然后通过不同的对应方式选出最匹配的结果。

由于任意 2 个句子都会生成一个模板，必然会导致模板对单个种子的拟合程度很高，而且单个种子的模板数量也会很大。比如在包含关系中只是(城镇, 街道)种子获得的模板就有上万个。为了解决这个问题，可以通过模板有效性规则来消除一些无用的模板。

规则 1 before, middle, after 中最多只能有一个为 NULL。

规则 2 before 最左边的内容不能是<ANY_STRING>。

规则 3 before, middle, after 中不包含句末标点符号。

通过以上 3 个规则的筛选，得到本次的模板候选。

对于“包含”关系中“小镇”和“街道”概念对的一个次模板的生成实例如图 4 所示。

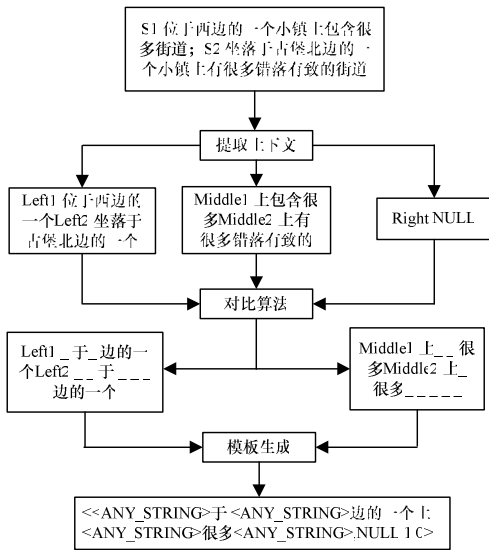


图 4 次模板生成实例

3.5 模板置信评价机制

定义一个合理的模板置信度对于模板在下一轮的抽取是至关重要的，这既影响下一轮新的概念对的生成，也影响新的模板的生成，所以对于模板置信度的评价是采取概念关系模板的关键步骤。很直观地可以想到用模板可以抽取正确概念对的数量来评估模板的置信度，但是还应该考虑到模板的覆盖能力，以及对于新的概念对产生的贡献能力。因此，定义如下置信度公式：

$$Confidence(P) = ConR(P) \times ((1 - \alpha)1b(P_r) + \alpha 1b(P_{new}))$$

$$ConR(P) = \frac{P_r}{P_r + P_w}$$

其中， P_r 表示该模板抽取的正确的概念对的数量； P_w 表示抽取的错误的概念对的数量； $ConR(P)$ 是用来衡量模板的正确率的； P_{new} 表示抽取的正确概念对中新的概念对的数量，用

来衡量模板的覆盖能力，以及对于新的正确概念对的贡献能力。采用这种对于模板置信度的评价机制，既考虑到了模板的正确率，同时也考虑到了模板的覆盖能力和对于新的概念对产生的贡献能力。

4 实验与系统评价

这部分将给出采用主次模板方法抽取概念对的实验结果，并与经典的 DIPRE 方法的性能作一下对比。实验采用的训练数据是从网上收集的大量童话故事文集，由于网上大部分是 html 格式，因此首先对文集作一下预处理，也就是去掉一系列的标签，获取童话故事内容。用分别以“包含关系”，“居住地(所属地)关系”，“上部关系”为例，在大规模童话故事集中进行试验。初始种子如表 1 所示。

表 1 初始种子

包含关系	居住地(所属地)关系	上部关系
城镇 街道	动物 巢穴	荷叶 水面
小镇 商店	青蛙 池塘	烟囱 房屋
农场 菜园	番茄 菜园	花草 地面
旅店 房间	鸟 树木	人 大街
森林 树木	人 房子	狼 田野

在本次实验中，对主次模板方法和 DIPRE 系统中的方法作了比较。由于纯文本的内容，因此在实现 DIPRE 时没有加入概念对的 URL 属性。本文对 2 个系统各自进行 3 轮训练，实验结果如表 2 所示。

表 2 概念对与模板的数量的对比

迭代次数	主次模板方法		DIPRE 方法	
	概念对个数	模板个数	概念对个数	模板个数
第 1 次	336	1 136	336	964
第 2 次	1 304	3 905	1 021	3 071
第 3 次	3 548	10 578	3 013	9 027
总数	5 188	15 619	4 370	13 062

在第 1 轮中，2 种方法都使用同样的 336 个初始种子概念对。后 2 次在各自系统的上一轮生成的概念对中，挑选所有正确的作为下一轮的模板获取。从实验结果可以看出，主次模板方法比 DIPRE 在相同文本集上提取了更多的模板和概念对。

考察主次模板方法和 DIPRE 方法抽取的模板在大量童话故事核心名词之间关系建立的覆盖性和正确性，也就是模板被应用于核心名词关系建立中所能得到的概念关系的总数和正确概念关系的数量。本文将 2 种方法 3 次迭代训练得到的模板集运用于测试文本，结果如表 3 所示。就包含关系而言，主次模板方法的正确率达到了 82.13%，采用 DIPRE 方法获得的正确率是 70.45%。虽然其他 2 种关系的识别效果不如包含关系，但是与 DIPRE 相比效果还是要好一些。整个实验也表明采用主次模板方法提取的模板集合比 DIPRE 有更高的覆盖能力和正确率。

表 3 2 种方法的概念对个数与正确率对比

关系	主次模板方法		DIPRE	
	概念对个数	正确率/(%)	概念对个数	正确率/(%)
包含关系	5 188	82.13	4 370	70.45
居住地(所属地)关系	4 894	80.34	4 233	69.56
上部关系	5 463	81.47	4 526	67.19

5 结束语

本文采用主次模板方法构建童话故事中的概念关系。在主模板的构建过程中，采用了部分句法分析的模式，从而可

(下转第 148 页)