旅游自动应答语义模型分析与实践①

王　彦[1,2], 左　春[1,2,3], 曾　炼[4]

[1](中国科学院软件研究所 软件工程技术研究开发中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
[3](中科软科技股份有限公司, 北京 100190)
[4](金童软件科技有限公司, 南京 210019)

摘　要: 针对常见问答系统采用的以词法分析为基础的浅层语义模型难以有效挖掘用户问句深层语义的问题, 本文立足于旅游问答应用领域, 采用组合范畴语法对旅游问句进行句法分析, 使用 Lambda 演算式表示问句语义, 以此构建旅游领域问句的语义模型, 以便于通过精确的问句语义快速查找应答结果. 研究首先进行旅游领域数据采集与语料标注的准备性工作, 并针对语料对旅游问句的句式句法进行分析; 然后采用基于概率的组合范畴语法的监督学习过程, 通过训练获得较为可靠的旅游问句语义词典; 最后根据语义词典及其他相关知识, 学习用户问句语义, 构建旅游自动应答语义分析系统, 着重于问句解析和相应的语义模型的构建. 通过在评测集上的验证, 这种语义解析方法在解析效果上有比较明确的提升.

关键词: 旅游问答系统; 组合范畴语法; lambda 演算; 语义模型; 监督学习

Analysis and Practice of Semantic Model in Tourism Auto-Answering System

WANG Yan[1,2], ZUO Chun[1,2,3], ZENG Lian[4]

[1](Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Sinosoft Co. Ltd., Beijing 100190, China)
[4](Jintongsoft Co. Ltd., Nanjing 210019, China)

Abstract: According to the weakness of shallow semantics models based on lexical analysis which are commonly used in QA system, shallow semantics models cannot accurately analyze the deep semantics of users' questions. This paper focuses on the tourism QA application field, adopts the combined category grammar (CCG) to parse the question sentences, and uses lambda calculus to express the question semantics, so that semantic models on tourism questions can be derived. And it's convenient to search answers according to such accurate semantic quickly. The research first carries out data acquisition and corpus tagging preparatory work，including the analysis of tourism question corpus both in sentence pattern and syntax. Then the supervised learning process based on a probabilistic CCG algorithm is used to train a reliable semantic dictionary. At last, an automated answering system is built according to the semantic dictionary and related knowledge, which is mainly about the question parsing and building of corresponding semantic models. The final result on evaluation dataset shows that the semantic analysis method has relatively clear improvement in analytical performance.

Key words: tourism QA system; combinatory categorial grammars; lambda calculus; semantic model; supervised learning

近年来, 随着信息技术的飞速发展, 人们存取、交换、查找信息的过程变得越来越快捷方便. 但是, 海量数据的无序增长给传统的信息检索技术带来了巨大的挑战, 如何快速获取简短、准确、关键的知识成为一个愈受关注的话题. 自动问答系统(QA)在这样的背景下逐渐成为研究和应用的热点, 与搜索引擎等传统信息检索方法相比, 它通过对自然语言处理技术的综合运用, 以较为准确、简练的自然语言回答用户提出的

① 收稿时间:2016-05-18;收到修改稿时间:2016-07-25　[doi:10.15888/j.cnki.csa.005640]

问题, 对用户的帮助更直接明显. 自动问答系统的性能直接地体现了计算机对自然语言处理的智能水平, 因而许多研究机构和公司都针对这一研究领域进行了比较深入的研究开发.

有别于传统信息检索系统中采用关键词组合的方式进行查询, 问答系统多是以自然语言的语句形式提出问题, 在旅游领域的语料上更是证实了这一点. 因此, 在构建自动应答系统时, 关键在于对问句采用不同方法进行分析处理, 以便精确获取问句语义, 从而快速查找对应答案. 随着问答系统获得广泛地研究, 许多不同种类的模型和各种处理的技术被不断提出. 按照问答系统三要素—问题、数据、答案的不同, 可以将问答系统分为限定领域和开放领域、结构化和非结构化、抽取式和产生式等不同类型[1], 针对不同类型所采取的处理技术也不尽相同. 目前, 国内外有一些比较成熟的问答系统. 麻省理工大学人工智能实验室开发的 Start 系统[2]基于 Web 应用, 能够向用户提供精确的信息; IBM 基于统计方法的问答系统 Watson[3], 提供了许多方面的问答服务. 国内许多大学和科研机构也对问答系统展开了深入的研究, 比如中科院计算所开发的基于知识问答系统的 HKI[4]支持自由提问, 能够提供各个领域的知识服务, 其他像清华大学和哈尔滨工业大学等也开展了很多富有成效的工作[5]. 但是, 应当注意到, 目前绝大多数问答系统(特别是中文问答系统)大多建立在基于词法的浅层语义分析上, 用户对这种仅使用简单的词袋模型所获取的检索结果并不满意, 所以问答系统需要开辟新的方法对问句语义进行理解[6]. 针对上述问题, 必须要采用一种能够进行深层语义理解的分析方法. 通过对研究所采用的旅游问答语料特点的分析, 我们采用了组合范畴语法(CCG)[7]来对问句进行分析建模. 组合范畴语法是爱丁堡大学的 Mark Steedman 等人提出的一种计算语言学的分析方法, 它基于形式语言学中句法和语义相对应的原则, 为计算机处理自然语言提供了句法和语义的透明接口. Artzi Yoav 等人用这种方法来处理机器指令语义[8]和地理问答语料[9], 通过快速的语义建模分析, 取得了良好的结果. 目前, 在中文应用上, 清华大学和微软亚洲研究院的一批学者构建了中文的 CCG 树库[10], 为使用这种方法进行中文语句句法分析奠定了基础.

本文引入组合范畴语法对旅游问句进行分析, 并构建问句的语义模型, 以此对问句深层语义进行挖掘. 研究的问答系统基于旅游问答领域, 是一种特定领域的问答系统, 这也是目前在工业界能够发挥实际作用的主要类型. 在实施过程中, 具体工作可以分为三步: (1)数据采集与语料标注的准备性工作, 包括问句的准备和问句语法和语义方面的标注; (2)采用基于概率的组合范畴语法的监督学习过程, 通过训练和验证过程获得较为可靠的语义词典, 作为问句解析建模的知识库; (3)构建自动应答系统, 着重于问句的语句解析和对应语义模型的构建, 并可以此作为服务模块供其他系统调用.

1 旅游自动应答语义系统概述

1.1 旅游自动应答系统架构

系统总体的架构图如图 1 所示.

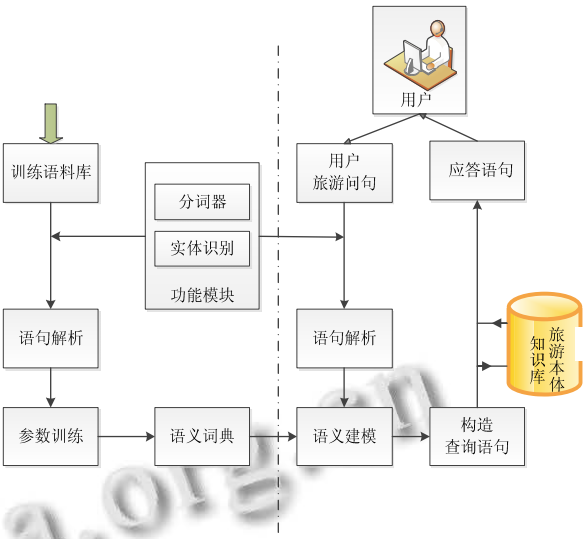

图 1 系统架构图

从系统架构图上可以发现, 整个系统包含两个主要的流程: 第一个流程是指图中左边部分, 它从旅游问句语料输入开始, 包括语句解析和参数训练等一系列过程, 主要目的是得到训练好的领域语义词典, 这是语义学习模块; 第二个流程是指图中右边部分, 由用户发起提问开始, 经过语义解析和建模, 学习并构造查询语句, 最后通过查询旅游领域本体知识库返回给用户对应的信息, 这是语义建模模块, 最主要目的是对用户问句构建语义模型.

在具体的系统实现过程中, 引入了分词器、实体识别模块以及查询语句构造模块等功能作为辅助, 因为这些不是本系统的核心功能, 所以采用了开源的实

现.

本系统可以也可以作为服务模块, 提供语句解析功能, 为更完善和综合的问答机器人项目提供支持.

1.2 旅游问句句法和语义表示

1.2.1 问句句法表示

在组合范畴语法中, 所有的语句成分都被赋予一个范畴(category). 某些范畴是基本范畴(例如 NP, S 等), 这些范畴被称为参数; 而另外一些范畴则是复杂范畴, 是由其他范畴运算而来, 这些范畴本质上是一种函数, 其记号将会表明运算的方向和最终结果[11].

复杂范畴的标注类型为 α/(\)β, 其中 α 和 β 是基本范畴, α 表示运算的结果范畴, β 表示运算中的参数, /(\)代表运算的方向: /代表参数必须出现在函数右侧, 而\代表参数必须出现在函数左侧.

因此, 范畴可以递归地定义如下:

(1) 和 NP 是范畴;

(2) 果 A 和 B 是范畴, 那么 A/B, A\B 也是范畴.

在句法分析过程中, 问句根据各语句所属范畴可以生成相应的句法分析树. 例如, 对问句“北京有哪些景点”进行范畴标注之后, 可以得到句法解析树如下图 2 所示.

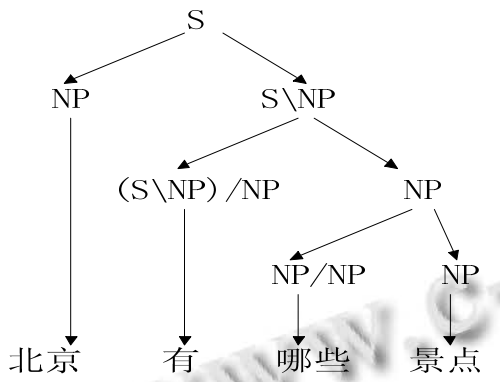

图 2　“北京有哪些景点”的句法解析树

1.2.2 问句语义表示

在使用组合范畴语法对问句进行建模时, 仅使用上节所述的范畴来标注语句成分是不够的, 它很明显缺乏语义的表示, 为了更好的表达问句语义, 我们引入 Lambda(λ)表达式对问句语义进行标注. λ 表达式是 Church 等人于 40 年代引入, 用于无歧义地构建函数表达式的方法, 它属于一阶逻辑, 表达能力强但逻辑系统并不复杂, 很适合进行运算[12].

问句语义模型中正是使用了这样的抽象对句子的语义进行标注, 与简单的 λ 表达式不同, 语义模型中使用的是类型化的 λ 表达式, 该表达式中的参数都具有特定的类型(type).

类型的定义也可以递归地定义如下:

(1) 和 t 是类型

(2) 果 a 和 b 是类型, 那么<a,b>也是类型

其中 e 是 entity 的简写, 代表实体; t 是 truth 的简写, 代表真值. <a,b>代表了某个函数, 该函数的参数类型为 a, 值类型为 b(可以类比数学中函数的定义域和值域).

在语义模型中, 简单范畴将对应一个简单类型(e 或者 t), 例如单个名词(包括专有名词)指代的大多是现实中存在的实体, 因此大多数 NP 可以对应为 e; 对于复杂范畴, 通常需要复杂类型来对应. 例如语句“我喜欢旅游”中的“喜欢”需要两个实体类型的参数, 最后获得的是一个真值(即“A 喜欢 B”这个命题是否为真), 因此其语义表达可以近似为:

$$\lambda x.\lambda y.like(x, y) \quad (1)$$

其中 x 和 y 是两个参数, 式(1)表示的语义即是 x 喜欢 y.

容易发现, 语义模型中的句法和语义标注存在着对应关系, 仍然以“我喜欢旅游”为例, 对其加入语义表示之后, 其句法和语义的同构性如表 1 所示.

表 1　“我喜欢旅游”语义生成表

形式	语法标注	语义标注	备注
喜欢	S\NP/NP	λx.λy.like(x, y)	初始状态
喜欢 旅游	S\NP	λy.like(y, 旅游)	令 x=旅游代入
我 喜欢 旅游	S	like(我, 旅游)	令 y=我代入

2 问句语义模型的构建

为了在基于旅游领域的问答系统中获取问句的深层语义, 问答系统将问句的语义解析分为四个步骤, 即问句整理、语句标注、语义学习和语义建模, 下面将从这四个方面进行阐述, 另外在最后也将对查询语句构造进行简要的讨论.

2.1 问句整理

本次研究的问句集来自于合作方携程公司所提供的旅游领域的真实用户语料, 原语料问句共有 11358 条. 通过筛选剔除, 排除以下情形语句:

(1) 明显语法错误或不完整;

(2) 长问句(超 30 字), 此种问句可以引入小句切分, 故不在此考虑;

(3) 他对语句处理有严重干扰的情况, 如包含特殊字符等.

最后, 获得实际可用的语句 8754 句.

由于所有语句都来自旅游领域, 通过对其分析, 并参考哈尔滨工业大学文勖等人提出的中文问句分类体系[13], 将这批问句按照提问意图分为食、宿、行、景四个大的种类. “食”是指有关美食、餐饮等方面的问题, 如“香港有什么好吃的”; “宿”是指住宿、休憩方面的问题, 如“帕劳当地有希尔顿酒店可以预订吗”; “行”是指关于交通、旅游线路方面的问题, 如“卡帕莱的行程攻略”; “景”是指关于风景、门票等方面的问题, 如“欢乐谷的门票多少钱”. 为了做进一步的标注, 从各种类语句里面按照该种类数量比例随机选取了共 600 条问句作为标注集.

问句分类整理后, 可以着手分词和实体识别的工作. 本次研究采用的分词器是以哈工大社会计算与信息检索研究中心研发的 “语言技术平台(LTP)” 为基础的“语言云”平台, 它为用户提供高效精准的中文自然语言处理云服务, 对问句分词结果如下图 3 所示. 对于实体名词的识别, 一方面可以借助于分词器完成一部分功能; 另外也可以采用一些算法对实体进行识别, 并增量式的添加名词实体, 从而准确的获取实体.

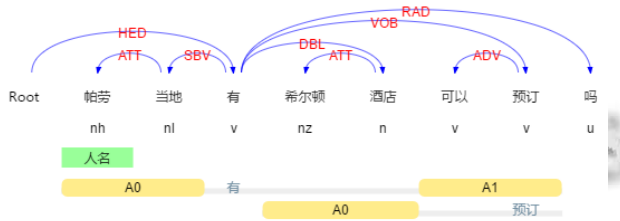

图 3 问句分词结果

2.2 语句标注

在研究工作中, 为了语义建模分析所需要进行的标注主要包括整句标注(只需要标注句子和对应的λ表达式)和词汇标注(语法信息和词汇信息, 标注格式为: 词形:-词法信息:语义信息), 再借助于清华 CCG 树库, 即可进行下一步的训练.

不论是句法还是语义的标注工作, 都要依靠对问句语法语义进行深入的分析, 才能为接下来的语义建模做好准备. 通过对标注集的分析, 可以发现每一种类的问句下面都有一些常见句式, 它们包含了典型的问句句法结构特点. 深入研究这一类典型问句的句式特点, 可以快速规范语句的标注流程. 下面以“哪些河流流经湖北”为例, 详细讲解一般的语句标注流程:

(1) 整个句子参考其语义给出语义标注. 例如, 对于句子“哪些 河流 流经 湖北”, 它希望找到流经湖北的且属于河流类型的实体, 因此其整句的语义标注为:

(lambda $0:e (and:<t*,t> (river:<r,t> $0) (loc:<lo,<lo,t>> $0 湖北:p))) (2)

在式(2)中, 引入了若干 Lambda 函数进行语义表示, 其中 loc:<lo,<lo,t>>表示“位于”, river:<r,t>表示“河流”, and:<t*,t>是取交集逻辑, lambda 是参数引导符(与λ符号相同), $0 就是一个参数.

(2) 句子的每个词进行范畴标注. 标注过程中可以先参考短语结构语法的标注方式对句子层级进行分析, 然后对于非基本范畴的词, 通过范畴演算, 赋予其正确的复杂范畴, 这种针对词语赋予范畴的过程可以参考清华中文树库里对应词语的范畴进行. 上句的短语结构语法标注如下:

哪些 河流 流经 湖北

Pron NP VP NP

NP VP

S

上述范畴中的 Pron 和 VP 是 CCG 中没有出现的基本范畴, 因此需要对这些范畴通过范畴演算, 赋予由其他基本范畴演算获得的非基本范畴如下:

哪些 河流 流经 湖北

Pron(NP\NP) NP VP(S\NP/NP) NP

NP VP(S\NP)

S

如上所示, “哪些河流”的范畴是 NP, 因此“流经湖北”共同被赋予范畴 S\NP; “哪些河流”的范畴是 NP, “河流”也是 NP, 因此“哪些”的范畴为 NP\NP; 同样地, “流经”的范畴是(S\NP)/NP, 它和右侧的 NP 共同组成 S\NP 范畴.

(3) 句子的每个词进行语义标注. 首先对基本范畴的词赋予其语义, 再对非基本范畴的词进行标注, 标注过程是参数化的过程, 实际上是函数赋值操作的逆过程.

最后, 语句完整的标注形式如下所示:

语句(Sentence): 哪些 河流 流经 湖北

句法(Syntax): S

语义(Semantic): (lambda $0:e (and:<t*,t> (river:<r,t> $0) (loc:<lo,<lo,t>> $0 湖北:p)))

2.3 语义学习

由于每条训练样本是一个句子 x_i 及其语义形式 z_i, 既没有包含句子的CCG语法结构 y_i, 使得能够知道 x_i 是如何映射到 z_i 的, 也没有包含映射所需的词条, 句子的语法结构是作为概率模型中的一个隐含变量. 为此, 文献[14]提出了一种期望最大化的算法对词典进行自动生成, 并进行模型的参数估计. 本文采用这种算法来进行旅游问句语义的学习.

算法 1. 基于 PCCG 语义解析的学习算法

输入: 训练集 $E=\{(x_i, z_i): i=1,\mathrm{L}, n\}$ 初始词典 Λ_0

输出: 训练好的词典 Λ; 参数 θ

```
Begin
  初始化 θ ← θ(0)
  for t ← from 1 to T do
    for i ← from 1 to n do
      λ ← Λ0 ∪ GenLex(xi, zi)
      π ← Parse(xi, zi, λ, θ(t−1))
      λi ← {l | l是用于生成π的词条}
    Λt ← Λ0 ∪ U_{i=1}^{n} λ
    //参数估计
    θ(0) ← Estimate(Λt, E, θ(t−1))
  return 词典 Λ(T) 和参数 θ(T)
End
```

算法中, Parse 函数针对输入参数, 输出能够生成 z 的概率最高的语法结构, 如果有多个, 则返回一个解析集合. Estimate 函数进行参数估计, 使用的是条件概率的方法来进行判断. 该学习算法的核心在于 GenLex 函数, 它输入句子及其语义形式, 启发式的生成可能的词条.

最后, 算法生成的部分语义词典如图 4 所示.

2.4 语义建模

得到用户语义词典之后, 针对用户新提出的领域问题, 可以结合句法分析和语义词典问句语义的合成, 即语义建模的过程. 语义建模的过程由组合范畴语法的运算规则控制, 而在这一过程中, Lambda 函数所表示的语义也能够同步参与运算, 它采用的是一种动态规划的思想. 文献[15]提出了一种上下文无关的算法, 能够在多项式时间内完成语义结合, 本文在语义建模的过程中采用了这种方法.

```
北京 :- NP : 北京 : c
寺庙 :- NP : (lambda $0:e (temple:<te, t> $0))
哪些 :- NP/NP : (lambda $0:<e, t> $0)
香格里拉 :- NP : 香格里拉 : c
行程攻略 :- S\NP : (lambda $0:e (strategy:<e, t> $0))
好玩的 :- S\NP : (lambda $0:e (interesting:<e, t> $0))
热闹 :- S\NP : (lambda $0:e (interesting:<e, t> $0))
卡帕莱 :- NP : 卡帕莱 : c
餐饮 :- S\NP : (lambda $0:e (food:<e, t> $0))
隐贤山庄 :- NP : 隐贤山庄 : sc
游玩指南 :- S\NP : (lambda $0:e (guidance:<<e, t>, t> $0))
三亚 :- NP : 三亚 : c
天气 :- S\NP : (lambda $0:e (weather:<e, t> $0))
```

图 4 旅游自动应答系统语义词典(部分)

算法 2. CKY 算法

输入: 自然语言句子 $\mathrm{word}=(w_0, w_1, \mathrm{K}, w_n)$

输出: 语义形式 z

```
Begin
  //从词典中找到每个单词的相应的词条,初始化chart
  tag()
  for j ← from 0 to length(words) − 1 do
    for i ← from j downto 1 do
      for k ← from i to j−1 do
        //应用二元规则,如向前应用、向后组合等
        result ← applyBinaryRules(char[i,k], char[k+1,j])
        add result to chart[i,j]
        //应用一元规则,如类型提升、类型改变等
        result ← applyUnaryRules(chart[i−1, j])
        add result to chart[i,j]
  //返回CCG语法结构根节点范畴的概率最高的语义形式
  return getSemanticForm(chart[0,length(words)−1])
End
```

算法中, applyBinaryRules 函数表示对该语句部分应用组合范畴语法的二元规则, 如前向/后向的应用规则、组合规则等; applyUnaryRules 函数表示对该语句部分应用组合范畴语法的一元规则, 如类型提升、类型改变规则等; getSemanticForm 函数用于获取生成结果中范畴为 S 类型的语句概率最高的语义形式, 这样我们就可以获得问句最终的语义模型.

以前述“北京有哪些旅游景点”这个问句为例, 其语义模型的生成过程如图 5 所示. 由图可以发现, 根据词语的范畴和组合范畴语法的运算规则, 在语义模

型生成的过程中, 多次运用了后向函数应用规则, 依次生成了语句段“哪些 景点”、“有 哪些 景点”以及最终问句“北京 有 哪些 景点”的语义形式.

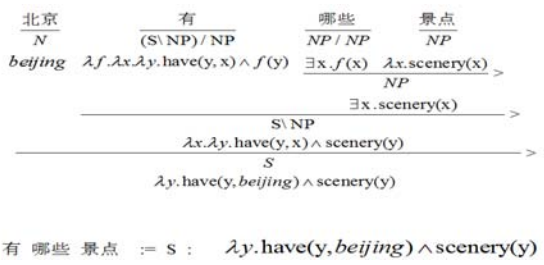

图 5 “北京有哪些景点”的语义生成树

2.5 查询语句构造

文献[16]提出了一种将语义三元组转换成 SPARQL 语句, 然后在本体知识库中进行查询的方法, 本文即采用这种方法将已经构建的语义分析模型进行类似转换.

该查询语句的构造依据在本体中构造 OntoTriple 与三元组中 Subject 和 Object 相对应的思想: Subject、Predicate、Object 构成的三元组<Subject, Predicate, Object>是 Ontology 中的最基本元素, 也对应着 SPARQL 查询的三元组方式<S, P, O>, SPARQL 查询就是通过已知 S、O、P 中的一个或两个已知的元素来确定其他的元素的值. 在 SPARQL 中, S 是本体的 Individual; P 指本体中的属性(DataProperty 和 ObjectProperty); O 是某个 Individual 对应的属性值. 当 P 是 DataProperty 时, O 是一个 Literal, 当 P 是 DataProperty 时, O 则是一个 Individual. 它们所确定的类型示例如表 2 所示.

表 2 Type S 和 Type O 类型示例

Type O	Type S	类型
Individual	Literal	查询 S 中值为 O 的属性
Individual	ObjectProperty	查询 S 中对象属性 O 的值
Individual	Individual	查询 S 和 O 实例的关系
Individual	DataProperty	查询 S 中数据属性 O 的值
Class	Individual	查询 O 在本体中所属类型

由前面分析可以看出, 针对问句构建的语义模型中, 单个用于表示语义的 Lambda 函数就等价于一个三元组, 因此整个问句即可转换成一组级联的 SPARQL 语句. 根据表 2, 根据 TypeS 和 TypeO 可以确定查询类型, 然后可由 OntoQuery 转换成 SPARQL 语句.

之所以采用上述这种方法进行查询转换, 是因为它提供了一种从语义表示到数据集查询的桥梁, 能够生成一种较为简单的查询语言, 快速查找问句答案, 极大节省了工作量.

3 实验及结果分析

3.1 语义学习结果及分析

在使用 PCCG 算法对整个有效问句集进行句法解析之后, 词汇与范畴之间的对应关系如表 3 所示.

表 3 词汇范畴对应表

词汇类型	数量(个)
包含 1 个范畴的词汇	20281
包含 2 个范畴的词汇	5053
包含 3-5 个范畴的词汇	3603
包含 6-10 个范畴的词汇	938
包含 10 个以上范畴的词汇	390
总词汇	30247

由表 3 可以看出, 不同词汇类型之间的分布符合语言学中存在的 Zipf 定律[17], 也在一定程度上说明了问句集构建的合理性. 表中只包含一个范畴的词汇项占到整体词汇的 67%以上, 是因为在旅游问句集中, 大部分词汇都是以景点名为代表的实体名, 它们在问句中的用法单一, 所以一般只对应一个 NP 范畴. 而包含多个范畴的词汇大部分是动词之类承担主要语义的词汇.

3.2 语义建模性能结果及分析

本节通过实验对旅游问句语义模型转化成 SPARQL 查询语句, 并对旅游问答本体知识库进行查询和自动应答, 最后将应答结果与标准问答集进行比较, 以此来判断语义解析结果的有效性.

如前所述, 为了测试问句语义解析模型在测试集中的性能, 实验从问句集中提取 600 条典型句子, 这些句子在该领域中构成了旅游领域常见的问法知识. 然后将这一批被标注的句子组成训练集, 在问句集中另选 200 句作为测试集. 实验首先在训练集上训练得到语义词典, 然后在测试集上进行语义建模解析, 将解析结果转成查询语句在本体知识库中查找对应的回答, 最后将其与构造的标准问答集进行比对. 语义解析的性能指标包括准确率(precision)、召回率(recall)、和 F1 度量(F1-measure). 其中准确率是所有包含应答结果的测试语句中正确应答的句子的比例, 召回率是

实验中正确应答的句子在测试集包含正确应答语句中的比例, 而 F1 度量是准确率和召回率的调和平均值. 为了作为参照, 另外选取了两种常用的问答系统应答结果搜索的方法: 向量空间模型方法和 QA 配对模型方法[18], 实验结果如表 4 所示.

表 4 语义解析的性能分析

方法	准确率	召回率	F1 值
向量空间模型	0.54	0.71	0.61
QA 配对模型	0.68	0.63	0.65
语义解析模型	0.75	0.68	0.72

从实验结果可以看出, 语义解析模型在自动应答时的准确率相比其他两种方法有了明显的提升, 在召回率方面仅次于向量空间模型, 并且拥有最高的F1 值. 这是因为, 语义解析模型不仅停留在词法层面, 而且深入到句法层面上对问句语义进行分析, 所以能够获得较高的准确率. 但是, 向量空间模型使用简单的词汇相似度计算, 有一种尽可能多匹配的趋势, 所以虽然在召回率上表现比较好, 但准确率却因而受到影响, 也拥有最差的 F1 值. 综上, 基于组合范畴语法的旅游自动应答语义模型在对问句的语义分析上具有比较明显的优势.

4 结语

为了解决现有方法对问答系统中问句深层语义理解不够准确的问题, 本文引入了组合范畴语法对问句进行分析, 并在此基础上提出构建问句的句法和语义模型, 通过语义学习和模型构建, 进行问句的语义理解. 在研究过程中, 不仅积累了丰富的领域相关的语言知识和对应的语料语库, 更是构建了一个切实可行的旅游问句语义模型分析系统, 对于相关问句的语义解析工作有着极大的帮助. 在接下来的工作中, 最主要的是从语义泛化、特殊句式理解和查询优化等几个方面加强系统的性能, 进一步提高其实用性. 语义泛化和特殊句式理解主要从问句种类入手, 一方面要更加完备旅游领域问句标注种类, 减少未知句式干扰; 另一方面要针对典型特殊句型(如省略句、倒装句)着重分析, 提高解析准确度, 这在文献[13]和文献[18]有初步的讨论. 查询优化主要应对大数据集中快速查找的问题, 针对查询转换后的 SPARQL 语句进行优化, 这方面可采用算法较多, 如文献[19]提出的 HMSST 算法等, 都可以获得较好的效果.

参考文献

1 毛先领,李晓明.问答系统研究综述.计算机科学与探索, 2012,6(3):193–207.

2 Katz B, Lin JJ, Felshin S. The START multimedia information system: Current technology and future directions. Proc. of the International Workshop on Multimedia Information Systems. 2002.

3 Ferrucci D, Brown E, Chu-Carroll J, et al. Building watson: An overview of the deep QA project. AI magazine, 2010, 31(3): 59–79.

4 冯东辉.NKI 知识界面:实现人和知识的对话.中国科学院计算技术研究所,2001.

5 秦兵,刘挺,王洋,等.基于常问问题集的中文问答系统研究.哈尔滨工业大学学报,2003,35(10):1179–1182.

6 白硕.自然语言处理与人工智能.中国计算机学会通讯, 2015.

7 Steedman M, Baldridge J. Combinatory categorial grammar. Encyclopedia of Language & Linguistics, 2006: 610–621.

8 Artzi Y, Zettlemoyer L. Weakly supervised learning of semantic parsers for mapping instructions to actions. Trans. of the Association for Computational Linguistics, 2013, 49–62.

9 Artzi Y, Zettlemoyer L. UW SPF: The University of Washington semantic parsing framework. Computer Science, 2013.

10 宋彦,黄昌宁,揭春雨.中文 CCG 树库的构建.中文信息学报,2012,26(3):3–8.

11 李可胜,邹崇理.基于句法和语义对应的汉语 CCG 研究.浙江大学学报:人文社会科学版,2013(6):132–140.

12 Steedman M. The syntactic process. Computational Linguistics, 2015, 27(1): 146–148.

13 文勖,张宇,刘挺,等.基于句法结构分析的中文问题分类.中文信息学报,2006,20(2):33–39.

14 Clark S, Curran JR. Wide-coverage efficient statistical parsing with CCG and log-linear models. Computational Linguistics, 2007, 33(4): 493–552.

15 Cocke J. Programming languages and their compilers: Preliminary notes. Courant Institute of Mathematical Sciences, New York University, 1969.

16 莫桂烽.基于本体的人物图片检索系统的研究[学位论文].北京:中国科学院大学,2015.

17 许文霞.齐普夫定律与中文词频分布机理.情报科学,1986,(1):29–36.

18 张亮.面向开放域的中文问答系统问句处理相关技术研究[博士学位论文].南京:南京理工大学,2006.

19 董书晓,汪璟玢.HMSST:一种高效的 SPARQL 查询优化算法.计算机科学,2014,41(S2):323–326.