基于依存关系的问句理解与问句分类*)

林旭东[1,2] 彭 宏[1] 林丕源[2] 邓健爽[1]

(华南理工大学计算机科学与工程学院 广州 510640) (华南农业大学信息学院 广州 510642)[2]

摘 要 问句理解是问答系统的首要过程,问句分类是问句理解的主要组成部分,它在问答系统中具有非常重要的作用,因为问句类型有助于在文档中定位和抽取答案。问句分类的目标是基于预期的答案类型,准确地分类问句。本文提出依存关系规则与统计方法相结合,实现了基于依存关系的中文问句理解与问句分类机制。实验表明:支持向量机结合依存关系的特征抽取方法,获得了较高问句分类正确率。

关键词 问句分类,依存关系,依存关系树,命名实体识别

Question Interpretation and Question Classification Based on Dependency Relations

LIN Xu-Dong[1,2] PENG Hong[1] LIN Pi-Yuan[2] DENG Jian-Shuang[1]

(College of Computer Science & Engineering, South China University of Technology, Guangzhou 510640)

(College of Information, South China Agricultural University, Guangzhou 510642)[2]

Abstract Question interpretation is the first step of question answering system. Question classification is the main part of the question interpretation and it plays a crucial important role in the question answering system because categorizing a given question is beneficial to identify an answer in the documents. The goal of question classification is to accurately assign labels to question based on expected answer type. In this paper, we use dependency relation rules and statistical method to understand questions and classify questions. In this experiment, we perform the SVM algorithm and a dependency relationships feature extraction method to get high classification accuracy.

Keywords Question classification, Dependency relations, Dependency tree, Named entity recognition

1 引言

问题回答(简称 QA)系统为用户提供智能的人机界面,允许以自然语言提出问题,系统自动给出简短的正确答案。问题回答是机器学习、信息检索、信息抽取和自然语言处理等领域的研究热点[1]。问答系统的实现包括问句理解、信息检索和答案抽取等过程,问句理解中很重要的一个部分就是识别问句的类型,即问句分类。问句类型是定位和检验答案及制定抽取答案策略的关键因素[2]。

目前,问句分类的研究主要集中在两个方面:一是基于规则的方法,通过专家提取各种问句类型的疑问词与其它相关词组合的特征规则,通过规则来判定问句所属类型[3]。另一种方法是通过统计的方法来实现问句的分类,通过对真实的经过标注的问句语料进行统计学习,提取能表达各种问句类型的特征规则,建立学习模型,实现各种问句的类型识别[4]。

本文以《知网》为汉语常识性知识库,提出了以谓词核心驱动,结合问句案例文法规则,对问句进行命名实体识别、词义消歧、句法依存关系分析和语义依存关系分析,实现问句理解过程,抽取问句特征,生成特征向量。最后使用基于统计的机器学习方法支持向量机(SVM)实现问句分类。

本文重点介绍句法依存关系分析和语义依存关系分析,然后讨论问句理解和问句分类,最后给出实验结论。

2 句法依存关系分析

句法分析(parsing)是指在给定文法下来分析自然语言的层次结构,它是自然语言处理中的中心问题之一,并在自动问答、机器翻译、信息检索、信息抽取等领域中有重要的应用[5]。依存文法的句法结构的主要元素是语义依存关系,即句子中词对的二元关系,其中一个记为核心词(head),另一个记为依存词(dependent)。依存关系反映的是核心词和依存词之间语义上的依赖关系[6]。

最简单的句子是单句,仅包含主语和谓语,甚至主语有时也可以省略(如祈使句):

1) S->(SUB) VP

其中 S(Sentence)表示句子,SUB(subject)表示主语,VP(Verb Phrase)表示谓语,圆括号()表示括号内的成分可选(可以省略)。

单句的主语和谓语可进行如下扩展:

2) VP->Vi

3) VP->Vt OBJ

4) VP->Vtt OBJ OBJ

5) VP->(Vp) PP

6) SUB->NP

其中 OBJ(Object)表示宾语,PP(Predicative Phrase)表示表语,NP(Noun Phrase)表示名词成分;Vi 表示不及物动词,Vt 表示及物动词,Vtt 表示双宾及物动词,Vp 表示系动词。

其它句子成分,如宾语、表语、定语、状语、补语和名词成分也可进行如下扩展:

7) VP->Vi (COMP)

8) VP->Vt (COMP) OBJ

9) VP->Vtt (COMP) OBJ OBJ

10) COMP->(得) ADJ

*)基金项目:广东省科技攻关项目(A10202001),广州市科技攻关项目(2004Z2-D0091)。林旭东 博士研究生,主要研究方向:自然语言处理、Wed 文本挖掘、信息检索;彭 宏 教授,博士生导师,主要研究方向:数据挖掘、智能计算、生物信息处理;邓健爽 博士研究生,主要研究方向:人工智能、网络智能搜索、数据挖掘。

11) OBJ->NP
12) PP->ADJ
13) NP->(ADJ) NP
14) NP->Np
15) NP->Pron
16) NP->Num (Mw)
17) ADJ->(ADV) ADJ
18) ADJ->Adj (的)
19) ADJ->NP (的)
20) ADV->(ADV) ADV
21) ADV->Adv (地)
22) ADV->(Pre) OBJ

其中 ADJ(Adjective)表示定语或形容词成分，ADV(Adverbial)表示状语，COMP(Complement)表示补语；Np(Noun Phrase)表示名词，Pron(Pronoun)表示代词，Num(Numeral)表示数词，Mw(Measure word)表示量词，Adj(Adjective)表示形容词，Adv(Adverb)表示副词，Pre(Preposition)表示介词。

各种句子成分中，使用最复杂又灵活的是状语。它既可以位于谓词之前(句中)，也可位于句首和句尾：

23) S->(ADV) (SUB) (ADV) VP (ADV)

中文还有两个特殊的句式：'把'字句和'被'字句(这两种句式还可扩充定语、状语和补语等句子成分)：

24) S->SUB 把 OBJ Vt
25) S->SUB 把 OBJ Vtt OBJ
26) S->SUB 被 Vi
27) S->SUB 被 OBJ Vt
28) S->SUB 被 OBJ Vtt OBJ

句子各成分还可以通过并列连词进行连接：

29) SUB->SUB (Conj1) SUB
30) VP->VP (Conj1) VP
31) OBJ->OBJ (Conj1) OBJ
32) ADJ->ADJ (Conj1) ADJ
33) ADV->ADV (Conj1) ADV
34) COMP-> COMP (Conj1) COMP
35) NP->NP (Conj1) NP
36) PP->PP (Conj1) PP

其中 Conj1(Conjunction)表示并列连词。

前面的文法规则所能生成的都是单个的句子，即单句。自然语言中还有大量的复合句。从句法上说，复句是由具有一定逻辑关系的两个或两个以上的单句组成的复合句子：

37) S->S(Conj1)S
38) S->(Conj2)S (Conj2)S
39) SUB->(Conj2)S
40) OBJ->(Conj2)S
41) PP->(Conj2)S
42) ADV->(Conj2)S

其中 Conj2(Conjunction)表示非并列连词。

3 语义依存关系分析

为了描述事件类概念的必要角色框架，《知网》定义了 76 种动态的角色和属性。我们确定的依存关系标注集共包含 82 个依存关系，其中参考《知网》和清华大学语义依存关系的汉语语料库[7,8]，定义了 55 个语义依存关系；参考哈工大汉语依存树库，定义了 27 个句法依存关系。具体参看表 1 句法和语义依存关系。

表 1 句法和语义依存关系

句法依存关系	谓词主语	主题主语	从句主语
	谓词宾语	介词宾语	间接宾语
	从句宾语	定语	状语
	补语	重叠结构	方位结构
	地点结构	时间结构	数量结构
	同位结构	并列结构	关联结构
	兼语结构	语气结构	'的'字结构
	'地'字结构	'得'字结构	'把'字结构
	'被'字结构	一般分句	关联分句语
语义依存关系	关系主体	所有者	存现体
	经验者	施事	描写体
	占有物	受事	触及部件
	内容	类指	部分
	整体	结果	目标
	代价	结果事件	事件过程
	接续	伴随	描述
	限定	程度	方式
	频率	范围	评论
	起始时间	终止时间	时段
	时距	处所	状态
	相伴体	参照体	比较内容
	比较量	手段	工具
	材料	来源	原因
	目的	根据	来源
	原因	目的	根据
	方向	条件	让步
	递进	并列	除了
	核心成分		

4 问句理解

问题回答系统允许用户以自然语言形式的问句提问，系统需要理解问句的语义。问句理解需要完成中文分词、问句分析、问句分类等。其中问句分类是问句理解的核心，因为问句类型的确定对答案抽取具有重要指示作用。问句理解的过程包括词法分析(分词和词性标注)、命名实体识别、词义消歧、句法依存关系分析和语义依存关系分析等。

中文分词使用中科院的中文分词处理软件工具包 2.0 版。它不但可以完成中文分词功能，还可以给出词性标注，分词正确率高达 97.58%(973 专家组评测)，未登录词识别查全率均高于 90%。

如："谁是新中国的主要缔造者和领导人?"，分词和词性标注结果为：

"谁/r 是/v 新/a 中国/ns 的/u 主要/b 缔造者/n 和/c 领导人/n ？/w"

进行句法依存关系分析，得出句法依存关系树：

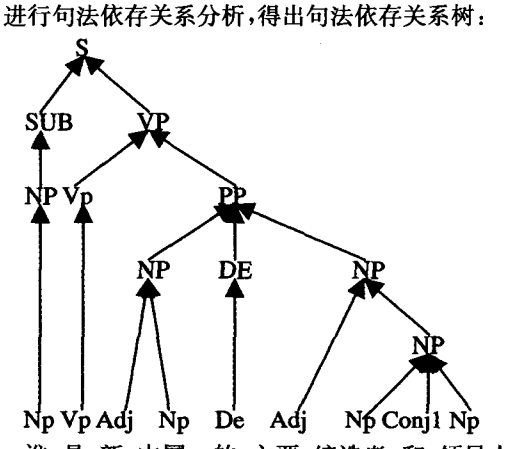

图 1 句法依存关系树

之后进行语义依存关系分析，得出语义依存关系树：

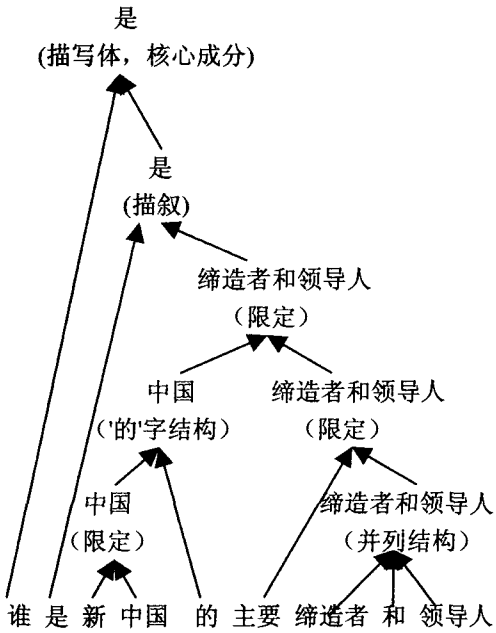

图 2　语义依存关系树

5　问句分类

目前，中文问答系统没有一个公认的中文问题库标准。我们使用哈工大信息检索研究室提供的问题库，包含 6 个大类，65 个子类，共 4394 个问句。

基于中文分词结果进行问句分析，结合《知网》进行问句的信息成分抽取和语义依存关系分析。首先确定问句谓词，它是句子各成分的分界点；其次提取疑问词；最后确定问句中心和约束条件。

例如：分析问句“谁是美国总统夫人?”

谓词：是；问句中心：夫人

疑问词：谁；约束条件：美国 总统

通过句法依存关系分析和语义依存关系分析，结合命名实体识别与词义消歧，可以更好地理解问句，提高问句的分类正确率。这种方法抽取的问句特征，比特征词抽取更好地理解问句，主要有以下三个优点：

(1)能表达长距离语法与语义依赖关系：

例如：第一个在太空行走的美国人是谁

分词和词性标注结果为：

第一/m　个/q　在/p　太空/s　行走/v　的/u　美国/n　人/n　是/v　谁/r　?/w

通过句法依存关系分析和语义依存关系分析，句中“人”有三个限制性定语：“第一个”，“在太空行走的”和“美国”。这种长距离语法与语义依赖关系识别出来以后，可以准确定位问句中心“人”，以及长距离的约束条件。

(2)依存关系能体现限定性疑问词(哪，哪个，什么等)所限定的名词成分，这对问句分类具有重要指示作用，例如以下两个问句：

(i)哪个总理毕业于清华大学?

(ii)朱镕基总理毕业于哪个大学?

抽取特征词的结果两个问句相似，疑问词和谓词也一样，主要的区别在于疑问词“哪个”所限定的名词成分不一样，因此发问中心和问句类型不一样。

(3)命名实体识别结合依存关系分析在问句分类中也有重要作用，例如以下两个问句：

(iii)谁是新中国的主要缔造者和领导人?

(iv)谁是毛泽东?

虽然这两个问句的疑问词都是“谁”，但问句(iii)是问人名，问句(iv)是描叙类型。两者的区别在于(iv)的表语是人名的命名实体。

实验表明，基于依存关系的问句理解，结合命名实体识别和词义消歧，进行问句特征抽取，比特征词抽取对问句分类可以获得更高的正确率。表 2 是分别使用特征词抽取和依存关系特征抽取，利用基于统计的机器学习方法支持向量机实现问句分类的结果。其中支持向量机分类器使用一对一模式，采用 LIBSVM 软件包[9]实现多值问句分类器的构造。

表 2　问句分类实验结果

问句类型	问句数量	正确	
		率特征词	依存关系特征
人名	92	67.3%	86.9%
地点	630	61.7%	88.1%
数字	878	71.4%	87.9%
时间	422	77.0%	88.6%
简称	88	70.4%	87.5%
描述	120	68.3%	88.3%

结论　本文以规则和统计方法相结合，实现了基于依存关系的中文问句理解与问句分类机制。以《知网》为汉语常识性知识库，提出了以谓词核心驱动，结合问句案例文法规则，对问句进行命名实体识别、词义消歧、句法依存关系分析和语义依存关系分析，实现问句理解过程。然后抽取问句特征，生成特征向量，使用基于统计的机器学习方法支持向量机实现问句分类。实验表明使用依存关系特征抽取比特征词抽取可以获得更高的问句分类正确率。

参 考 文 献

1　Zhang Dell, Lee Wee Sun. Question classification using support vector machines. In: Proceedings of the 26th annual international ACM SIGR Conference on Research and Development in Information Retrieval, New York, Jul. 2003. 26～32

2　Dan R, Chad C, Li Xin. Question-answering via enhanced understanding of questions. In: Proceedings of the 11th Text Retrieval Conference. Gaithersburg, Nov. 2002. 667～676

3　Kadri H, Wayne W. Question classification using support vector machines and error correcting code. In: Proceedings of HLT-NACCL 2003, Edmonton, Apr. 2003. 28～30

4　Metzler D, Bruce Croft W. Analysis of Statistical Question Classification for Fact-Based Questions, In: Proceeding of Information Retrieval, Jan. 2005. 481～504

5　Hermjakob U. Parsing and question classification for question answering. ACS-2001 Workshop on Open-Domain Question Answering, Toulouse, Feb. 2001. 255～262

6　Li Xin, Dan R. Learning question classifier. In: Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Aug. 2002. 556～562

7　Li Xin, Dan R, Kevin S. The role of semantic information in learning question classifiers. In: Proceedings of the 1st International Joint Conference on Natural Language Processing. Berlin, June 2004. 451～458

8　Lin Ming-Qin, Li Juan-Zi, Wang Zuo-Ying, Lu Da-Jin. A Statistical Model for Parsing Semantic Dependency Relations in a Chinese Sentence. Chinese Journal of Computers, 2004, 27(12): 1679～1687

9　Chang Chin-chung, Lin Chih-jen. LIBSVM: a Library for Support Vector Machines. http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001-05-15/2003-10-25