基于句法分析和答案分类的中文问答系统

孙 昂[1],江铭虎[1,3],贺一帆[1],陈 林[1],袁保宗[2]
(1. 清华大学人文学院计算语言学实验室,北京 100084;2. 北京交通大学信息科学研究所,北京 100044;3. 清华大学心理学与认知科学中心,北京 100084)

摘 要: 本文根据疑问词和谓语的距离信息对问句进行细致的句型分析,然后对答句进行浅层句法分析,在此基础上,抽取出问题特征集、答句特征集和组合特征集作为分类特征,引入最大熵模型和支持向量机训练答案抽取分类器. 基于不同特征组合训练得到的分类器在五类事实性问题上进行了测试,其 F 值分别达到 70.87 %和 85.75 %.

关键词: 中文问答系统; 句法分析; 答案抽取; 最大熵模型; 支持向量机
中图分类号: TP391 **文献标识码**: A **文章编号**: 0372-2112 (2008) 05-0833-07

Chinese Question Answering Based on Syntax Analysis and Answer Classification

SUN Ang[1], JIANG Ming-hu[1,3], HE Yi-fan[1], CHEN Lin[1], YUAN Bao-zong[2]
(1. *Lab of Computational Linguistics, School of Humanities and Social Sciences, Tsinghua University, Beijing* 100084, *China*;
2. *Institute of Information Science, Beijing Jiaotong University, Beijing* 100044, *China*;
3. *Center for Psychology and Cognitive Science, Tsinghua University, Beijing* 100084, *China*)

Abstract: This paper first conducts rigorous sentence pattern analysis of questions based on the distance between question word and predicate, and then conduct shallow parse of answer candidate sentences. Based on the analysis, we extract question feature set; answer sentence feature set and combined feature set as our features for answer classification. Then we apply maximum entropy model and support vector machine to these features to train answer classifiers. The F-Measures of the two classifiers' experiment conducted on five kinds of fact-based questions achieve 70.87 % and 85.75 % respectively.

Key words: Chinese question answering; syntax analysis; answer extraction; maximum entropy model (MEM); support vector machine (SVM)

1 引言

从海量信息中快速、准确地获得有用的信息,是信息时代人们的迫切需求,这一需求也推动了问答系统成为当前自然语言处理领域的一个研究热点.

问答系统和传统信息检索系统的主要区别于两个方面. 一是系统的输入,问答系统的输入不再是传统检索方法下的查询关键词,而是更为自然的日常语言的问句;二是系统的输出,传统检索系统的输出是和查询关键词相关的一系列文档,需要用户自己从文档中查找有用信息,而问答系统有能力直接返回针对用户提问的答案. 问答系统所关心的就是如何抽取出问题的正确答案.

目前,英文问答系统在答案抽取方面的研究已经比较深入. 最初,研究者们一般采用句模匹配的方法[1]. 该方法考虑答案出现的上下文信息,在问题类型确定的前提下,与此类型问题相关的答案抽取模板将被激活用来抽取答案. 该方法的弊端是需要精确的问题分类体系和性能良好的问题分类系统. 此外,答案抽取模版一般需要人工总结,费时费力.

近年来机器学习的方法逐渐成为主流[2,3]. 机器学习的答案抽取一般是基于这样一种假设,即包含正确答案的句子与问句的距离应该小于未含有正确答案的句子与问句的距离. 研究者们一般把问句和候选答句映射到不同的空间,然后计算问句和候选答句在此空间的距离. IBM[4,5]的系统把问句和候选答句映射到句法树和表层语义标记,然后在这个句法语义空间中计算问句与候选答句的相似度. 文献[6]则把问句和答句映射成逻

收稿日期:2007-10-24;修回日期:2008-01-25
基金项目:国家自然科学基金(No. 60673109)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

辑形式,然后用多种推理规则计算它们的相似性. 文献[7]把噪声信道模型应用到答案抽取中,将问句和答句的距离转化为计算条件概率 $P(Q|S_{iA_{i,j}})$,其中 S 为候选答句,$A_{i,j}$ 为 S 中可能的答案词.

与此相比,汉语问答系统在答案抽取方面的研究还非常不充分,很少有文献单独探讨汉语问答系统的答案抽取. 本文提出了一种新的答案抽取方法:在对问句和候选答句进行句法分析的基础上提取出若干语言学特征,使用这些特征训练一个判断候选答句正确性的二值分类器,根据分类器的分类结果抽取答案,并开发实现了这样一个问答系统.

问答系统关心的是问题的正确答案[8]. 通常一个问题要经过三个模块的分析和处理. 比如问句 1:"世界上人口最少的国家是哪个?". 首先是问题分析模块根据问句的句法特征和语义信息判别出其问题类型为 LOC_COUNTRY 类(地点类中的国家子类),并且抽取出"世界,人口,最,少,国家,是"作为查询关键词. 随后,文档检索模块利用第一步所提供的查询关键词进行传统的信息检索,其输出结果是和查询关键词相关的一些候选答句的集合. 最后,答案抽取模块根据一定的抽取策略抽取出正确答案"梵蒂冈".

2 基于分类的答案抽取

2.1 基于分类的答案抽取思想

自然语言处理中的许多问题,如分词标注,句法分析和语义分析等都可以形式化为分类问题. 给定一个上下文 x,我们可以估计某个类别 y 在此上下文中的概率 $p(y|x)$. 这种分类器可以通过对大规模的样本数据集合训练得到. 样本代表了分类任务的知识和信息,它和机器学习模型的结合可以预测随机过程将来的行为.

考虑到正确答案总是和一定的上下文相关联,本文提出把答案抽取的任务看成一个二分类问题. 即,给定关于某一个问题的一系列候选答句,我们对此候选答句集合中的句子进行二分类,把包含正确答案的句子识别为正例,未包含正确答案的句子识别为反例.

2.2 答案抽取与最大熵模型

最大熵模型在处理自然语言分类问题上的优势在于它关注上下文信息,其特征集不需要深层的语言学知识但仍然可以有效地近似语言关系的复杂性[9]. 因此,本文采用最大熵模型来训练答案抽取的分类器.

概率模型定义在 $X\times Y$ 上,X 是上下文特征集,Y 是一组类别标记. 就本文的问答系统而言,X 为从问答对中抽取的相关特征组成的向量,而 Y 有两个取值,$Y=1$ 表示候选答句为正例,$Y=0$ 表示候选答句为反例. 这样,可以利用最大熵模型来完成对候选答句进行分类的任务. 最大熵概率模型的定义如下:

$$p(x,y)=\pi\mu\prod_{j=1}^{k}\alpha_j^{f_j(x,y)} \tag{1}$$

其中,π 是常数,$\{\mu,\alpha_1,\cdots,\alpha_k\}$ 为模型特征参数,$\{f_1,\cdots,f_k\}$ 即所谓特征.

最大熵模型的原理是找到这样一个模型 $H(p)=-\sum_{x\in X,y\in Y}p(y|x)\log(p(y|x))$,$p(x|y)$ 在满足一定的约束条件下能够使得熵值 $H(p)$ 最大.

约束条件在问答系统中,即从问答对中抽取的相关特征集,用函数 f_i 来表示,它用来刻画任何一个样本对 (x,y) 的任何属性. 针对问答系统的答案抽取问题,我们定义一个二值特征函数:

$$f_i=\begin{cases}1, & \text{若 } x,y \text{ 满足一定的条件}\\ 0, & \text{否则}\end{cases} \tag{2}$$

给定特征集合后,首要的任务是基于训练集合计算每个特征的期望值,每个特征的约束条件都要求这个经验期望与模型中的理想期望值相同. 在所有满足限制的概率分布模型中,选取满足使熵值最大化的分布[10]. 本文中用于抽取答案的最大熵模型采取二值特征,并用最大熵中 improved iterative scaling(IIS)算法进行参数训练.

2.3 答案抽取与支持向量机

近年来,基于统计学习理论的支持向量机方法在文本分类等自然语言处理领域中得到了成功的应用[11]. 支持向量机能够找到使二分类问题分类间隔(Margin)最大的最优分类面. 由于 Margin 最大,支持向量机具有良好的外推能力,即使训练样本较小,也能得到良好的训练结果. 这些特性对于训练样本有限的问题答案分类来说具有重要的价值.

本文采用与最大熵模型相同的特征训练支持向量机分类器,进行答案抽取.

3 特征抽取

我们利用问答对作为抽取特征的资源. 选取的特征须能反映问题正确答案的性质,并能被运用到最大熵模型训练可识别出正确答案句子的分类器中. 考虑到特征资源的可用性,我们从问答系统的三个模块中抽取适合于用来训练分类器的特征,即三组特征:问句特征集,候选答句特征集和组合特征集.

3.1 问句特征集

3.1.1 问句的句型分析

对问句进行深层句法分析能帮助我们弄清用户的真正查询意图,有效地抽取答案. 鉴于现有汉语句法分析系统还不成熟[12],并考虑到问句句型的特殊性和有限性,本文在实验室原有句型系统的基础上,对问句进行了更为细致的句型分析.

我们首先采用距离疑问词最近原则确定问句的谓语动词,然后利用疑问词和谓词的距离信息,确定问句

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

中其他各词所担当的主要句法成分.在此句法分析的基础上,抽取出句子的主语,谓语和宾语作为特征.对于不是由动词担当谓语的情况,使用实验室原有的句型系统进行分析处理.

距离疑问词最近原则确定谓语动词

我们对汉语问句的语言学分析发现,距离疑问词最近的动词在问句中往往担当谓语.比如经过分词和标注后的问句:"发现/v 大庆/ns 油田/n 是/v 哪/r 一/m 年/n ?/w","是"距离疑问词"哪"最近,而且充当问句的谓语.再看一个比较复杂的问句:"日本/ns 天皇/n 裕仁/nr 以/p 广播/vn '/w 终/d 战/vg 诏书/n '/c 的/u 形式/n 正式/ad 宣布/v 日本/ns 无条件/d 投降/v 是/v 在/p 哪/r 一/m 年/n ?/w",尽管句子中有 5 个动词,但作谓语的还是距离疑问词最近的动词"是".

抽样统计结果也证实了我们的经验观察的正确性.本文从一个最大规模的问句集(约 20000 句)中随机抽取出 4000 句,对这 4000 句进行了人工谓语分析.分析结果如下:动词作谓语的问句数是 3369 句,占抽样问句总数的 84.22 %;距离疑问词最近的动词作谓语的问句数是 3249,占抽样总数的 81.22 %,占动词作谓语问句数的 96.44 %.因此,本文采用距离疑问词最近原则来确定问句的谓语.

利用疑问词和谓语动词的距离信息确定问句的其他主要成分

《现代汉语基本句型》[13]总结了疑问代词表示疑问的六种句型(中括弧内为疑问代词担当的句法成分,例子是从本文实验用句中抽取得到):

(1) 疑问代词[主] ‖ 谓

例 1:谁/r 发现/v 了/u 南极/ns 大陆/n ?/w

例 2:初/f 唐/nr 四杰/nr 中/f 的/u 谁/r 写/v 了/y《/w 滕/nr 王/n 阁/ng 序/n》/w ?/w

(2) 主 ‖ 动 + 疑问代词[宾]

例 3:欧盟/j 的/u 总部/n 是/v 哪/r ?/w

例 4:泉城/ns 是/v 中国/ns 的/u 哪/r ?/w

(3) 主 ‖ 疑问代词[谓语中心语](+"了""着")

例 5:他/r 怎么/r 了/y ?/w

(4) 主 ‖ 疑问代词[状] + 动/形/ ……

例 6:风/n 是/v 怎么/r 形成/v 的/u ?/w

(5) 主 ‖ 动/形 + 疑问代词[补]

例 7:那/r 家/q 公司/n 发展/v 得/u 怎么样/r 了/y 呢/y ?/w

(6) 疑问代词[定] + 中心语

例 8:什么/r 离子/n 使/j 水/n 呈现/v 酸性/n ?/w

例 9:什么/r 交通/n 工具/n 可以/v 在/p 雨/n 中/f 前行/v ?/w

例 10:什么/r 的/u 氧化物/n 是/v 玻璃/n 的/b 主要/a 成分/n ?/w

例 11:地球/n 大气/n 中/f 占/v 比例/n 最/d 大/a 的/u 是/v 什么/r 气体/n ?/w

例 12:UK/nx 是/v 哪/r 的/u 简称/n ?/w

例 13:最/d 笨拙/a 的/u 交通/n 工具/n 是/v 哪/r 种/q ?/w

由于本文的问题集主要是事实类问题,疑问词作谓语中心语、状语以及补语的问句很少.因此,本文主要研究疑问代词作主语、宾语和定语的情况.其中,疑问词作主语一般有谓语动词紧随其后,疑问词和谓语动词的距离是 0(句型 1).疑问词作宾语一般处于句末(句型 2).这两种情况比较特殊,易于处理.疑问词作定语的情况稍显复杂(句型 6),需要根据疑问词和谓语动词的位置信息做出相应的判断.

主语、谓语和宾语的抽取算法

(1) 根据距离疑问词最近原则确定谓语动词,并抽取出谓语动词.

(2) 如果疑问词和谓语动词的距离是 0 并且谓语动词紧跟疑问词之后,则可判定疑问词单独做主语,此时不抽取主语,抽取谓语动词之后的名词性成分(名词或名词短语)作为宾语(句型 1).

(3) 如果疑问词在句末,则可判定疑问词单独作宾语,此时不抽取宾语,抽取谓语动词之前的名词性成分作为主语(句型 2).

(4) 如果疑问词和谓语动词的距离不是 0 并且疑问词不处于句末,则可判定疑问词作定语(句型 6).

(a) 疑问词在谓语之前:如果疑问词单独修饰一个名词,则抽取该名词为特征主语词*(例 8)*;如果有多个名词紧跟疑问词,则抽取这些名词作为特征主语词项*(例 9)*;如果疑问词和"的"组合修饰名词性成分,则抽取此名词性成分作为特征主语词项*(例 10)*.如果疑问词之后并没有找到相应的名词性成分,则不进行特征主语词项的抽取.最后抽取谓语动词之后的名词性成分作为宾语词项.

(b) 疑问词在谓语之后:如果疑问词单独修饰一个名词,则抽取该名词为特征宾语词*(例 11)*;如果有多个名词紧跟疑问词,则抽取这些名词作为特征宾语词项;如果疑问词和"的"组合修饰名词性成分,则抽取此名词性成分作为特征宾语词项*(例 12)*.如果疑问词之后并没有找到相应的名词性成分,则不进行特征宾语词项的抽取*(例 13)*.最后抽取谓语动词之前的名词性成分作为主语词项.

*(*对于例外,算法实现中单独处理,限于篇幅,这里不赘述*)*.

3.1.2 问句特征集的抽取

问句词及词性序列(POS_Q):比如,问句 2"第一次

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

世界大战爆发于哪一年?"的 POS_Q 特征为 {第一/m 次/q 世界大战/n 爆发/v 于/p 哪/r 一/m 年/n ?/w}

查询关键词(Query Words,简称 QW):在经验分析和试验过程中,基于关键词对检索信息贡献度的大小,我们制定了三条选取查询关键词的标准.第一是抽取所有实词包括命名实体、名词、动词、形容词等;第二是选择性地抽取数词和量词,对于在句子的词及词性序列中,没有紧随疑问词之后的数词和量词加以抽取,如果数量词紧随疑问词之后则不加以抽取,比如对于问句 2,我们抽取"第一 次",不抽取"一";第三是过滤掉疑问词和停用词词表中的词(的,有…).

所以,问句 2 的 QW 特征为{第一 次 世界大战 爆发 年}.

疑问词(Interrogative Words,简称 IW):比如问句 2 的 IW 特征为{哪}.

主语(Sub):比如,利用本文的问句句法分析系统,问句 2 的 Sub 特征为{世界大战}.

谓语(Pred):比如问句 2 的 Pred 特征为{爆发}.

宾语(Obj):比如问句 2 的 Obj 特征为{年}.

3.2 候选答句特征集

这一特征集是从候选答句中抽取而来.如果把句子定义为 S,句子中的词定义为 W_i,那么 S 可以被表示为 $S = W_1 \cdots W_i \cdots W_n$.

候选答句词序列(W_i):此特征是对 S 的词进行列举,$\{W_1, W_2 \cdots W_{i-1}, W_i, W_{i+1} \cdots W_{n-1}, W_n\}$.

比如,给定一个候选答句(S_1):"第一次世界大战爆发于 1914 年,战争的导火索是萨拉热窝事件".S_1 的 W_i 特征是{第一 次 世界大战 爆发 于 1914 年 战争 的 导火索 是 萨拉热窝 事件}.

候选答句词性序列(POS_W_I):$\{W_1/POS\_TAG_1, \cdots, W_i/POS\_TAG_i, \cdots, W_n/POS\_TAG_n\}$.

比如,S_1 的 POS_W_i 特征是{第一/m 次/q 世界大战/l 爆发/v 于/p 1914 年/t ,/w 战争/n 的/u 导火索/n 是/a 萨拉热窝/ns 事件/n ./w}.

正确答案词的词性(POS_A):这里的词性也包括命名实体的标记.它也作为候选答句的一个约束条件.比如 S_1 的 POS_A 特征是{/t}.

3.3 组合特征集

组合特征是问题特征集和候选答句特征集的组合.以问题 2 为例,对于每一个从文档检索模块返回的句子,计算如下特征值:

f_1(QW_Match_Result):如果 W_i 在 QW 未匹配到任何词,则 f_1 为假,否则 f_1 为真.

例如 S_1,f_1 特征为真;但另一个候选答句:这一事件被称为萨拉热窝事件,其 f_1 特征为假.

f_2(IW_Match_Result):如果 W_i 匹配上了 IW,那么 f_2 为假,否则 f_2 为真.

比如 S_1,f_2 特征为真;但另一个候选答句:第一次世界大战爆发于哪一年,1914 年还是 1915 年? 其特征 f_2 为假,因为它包含疑问词"哪".

f_3(Sub_Match_Result):如果 W_i 匹配上了 Sub,那么 f_3 为真,否则 f_3 为假.

例如 S_1 的 f_3 特征为真,但是另一个候选答句:大战的导火线表面上是萨拉热窝事件.其 f_3 特征为假.

f_4(Pred_Match_Result):如果 W_i 匹配上了 Pred,那么 f_4 为真,否则 f_4 为假.

f_5(Obj_Match_Result):如果 W_i 匹配上了 Obj,那么 f_5 为真,否则 f_5 为假.

f_6(POS_A_Match_Result):如果 POS_TAG_i 和 POS_A 相匹配,那么 f_6 为真,否则 f_6 为假.

例如 S_1 的 f_6 特征为真;但另一候选答句"这/r 一/m 事件/n 成为/v 第一/m 次/q 世界大战/l 的/u 起/v 源/ng ./w",其特征 f_6 为假.

4 实验

4.1 文档检索模块

我们使用本实验室的搜索引擎 *Web Search* 作为文档检索模块.*Web Search* 可以自动从 Google 返回的文本片断中获取信息.对于每个问题,它的查询词(QW)将用来在网上检索信息,它的返回结果是 100 个网页的文本片断.然后我们对这些文本片断进行预处理,包括过滤垃圾信息(比如广告信息)和根据标点符号对文本片断进行句子切分.预处理后的结果即为我们的候选答句集合,其数量一般在 300 到 400 之间,为了实验结果的分析统一切分到 200 句.

这一切分主要是考虑到正例和反例的数据平衡,因为一般来讲文本片断中的反例是多于正例的.具体切分原则:首先利用 QW 和 POS_A 对候选句子进行检索,保留匹配成功的句子为正例,然后从未匹配成功的句子中增补候选答句数到 200 句.抽样统计结果中最理想的情况下正反例比例数为 1 比 1,不理想的情况下可达到 1 比 3.

4.2 训练集和测试集

表 1 训练数据集

问题类型	问题答案对数目(对)	网页文本片段数目(篇)	候选答句数目(句)
TIME_YEAR	150	15,000	30,000
HUM_PERSON	100	10,000	20,000
OBJ_SUBSTANCE	50	5,000	10,000
LOC_CONTINENT	50	5,000	10,000
LOC_COUNTRY	100	10,000	20,000

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2　测试数据集

问题类型	问题答案对数目(对)	网页文本片段数目(篇)	候选答句数目(句)
TIME_YEAR	50	5,000	10,000
HUM_PERSON	50	5,000	10,000
OBJ_SUBSTANCE	50	5,000	10,000
LOC_CONTINENT	50	5,000	10,000
LOC_COUNTRY	50	5,000	10,000

4.3　实验步骤

4.3.1　训练步骤

训练步骤是样本$((x^{(1)},y^{(i)}),\cdots,(x^{(n)},y^{(i)}))$训练基于最大熵模型(MEM1、MEM2)和支持向量机(SVM)的候选答句分类器.为了测试不同特征对于分类器的贡献度,我们首先利用特征 f_1,f_2 和 f_6(浅层句法信息)训练 MEM1,然后利用全部 6 个组合特征(即加入深层句法信息)训练 MEM2 和 SVM.

本文中最大熵模型和支持向量机的训练,分别使用 YASMET 和 SVMlight 工具包完成.

(1) 给定一个问题 q 和它的正确答案 a.

(2) 使用 ICTCLAS(中科院计算所的词法分析系统)对 q 和 a 进行标注,得到 POS_Q 和 POS_A,生成特征 QW 和 IW.

(3) 使用本文的问句句法分析系统抽取特征 Sub,Pred 和 Obj.

(4) 把(2)得到的 QW 提交给 *Web Search*,对其返回的文本片断进行预处理,得到一堆候选答句.

(5) 对每一个候选答句使用 ICTCLAS 得到 W_i 和 POS_W_i.

(6) 计算 $f_i(i=1\sim6)$.

(7) 生成样本$(x^{(i)},y^{(i)})$,$x^{(i)}$ 和 f_i 对应,$y^{(i)}$ 是候选答句正确性标记 0 或 1.

(8) 对于每一个问题 q,执行步骤(1)～(7).

(9) 把样本集提交给分类器 MEM1、MEM2 和 SVM.

4.3.2　测试步骤

给定输入 $x^{(i)}$,测试步骤使用分类器 1 和分类器 2 计算输出 $y^{(i)}$ 的值.

(1) 对每一个问题,生成一个样本$(x^{(i)},y^{(i)})$.

(2) 使用 MEM1 对候选答句进行分类.

(3) 使用 MEM2 对候选答句进行分类.

(4) 使用 SVM 对候选答句进行分类.

4.4　实验结果及分析

使用准确率(P)、召回率(R)和 F-Measure 对实验结果进行分析.

$$\text{Precision}=\frac{\text{Number of correctly classified candidates}}{\text{Total number of candidates}} \tag{3}$$

$$\text{Recall}=\frac{\text{Nuber of correctly classified Positive Examples}}{\text{Number of Positive Examples}} \tag{4}$$

$$F_\beta\text{ score}=\frac{(\beta+1)\cdot\text{Recall}\cdot\text{Precision}}{\beta\cdot\text{Recall}+\text{Precision}} \tag{5}$$

(在所有实验中,$\beta=1$)考虑到目前汉语问答系统尚未有统一的评测平台,也没有相关系统可以参照评比,本文的比较基准(Baseline)是通过单纯计算 6 个组合特征,不和最大熵模型结合得到:即如果一个样本的 6 个特征全部为真,那么我们认为其为正确答句;否则不是正确答句.这个基准可以看作是一个简单的句模匹配方法(当然并非真正意义上的句模匹配).

表 3　实验结果

问题类型		TIME_YEAR	OBJ_SUBSTANCE	LOC_CONTINENT	LOC_COUNTRY	HUM_PERSON	平均值
基准(%)	准确率	69.16	28.75	26.52	55.38	25.68	41.1
	召回率	13.04	9.42	8.30	14.42	9.07	10.85
	F 值	21.94	14.19	12.64	22.88	13.41	16.96
MEM1(%)	准确率	61.54	31.18	29.90	68.42	26.49	43.51
	召回率	97.47	66.72	88.42	87.56	30.57	74.15
	F 值	75.45	42.37	44.69	76.82	28.38	53.54
MEM2(%)	准确率	81.24	70.35	84.68	79.90	25.97	68.43
	召回率	95.42	68.56	89.62	83.27	31.65	73.70
	F 值	87.76	69.44	87.08	81.55	28.53	70.87
SVM(%)	准确率	97.80	100.00	100.00	96.51	89.74	96.81
	召回率	97.80	32.56	80.00	94.86	79.55	76.95
	F 值	97.80	49.12	88.89	95.68	84.34	85.75

从平均值来看:基准的准确率、召回率和 F 值都比较低,从一定程度上说明单纯的关键词匹配技术在面对新问题的时候缺乏预测能力,抽取答案的性能比较低;MEM1 和基准相比,系统的召回率明显提高,提高了 63.3%,说明浅层句法信息和最大熵模型的结合可以有效地对新问题的候选答句进行判断和预测,系统识别正例(正确答句)的能力明显提高;MEM2 和 MEM1 相比,二者在召回率上相差不大,但其准确率比 MEM1 提高了约 25 个百分点,说明加入深层句法信息对于识别反例(错误答句)的贡献度更大,从而在总体上进一步提升了答案抽取的性能. SVM 和 MEM2 相比,平均 F 值进一步上升,但在 OBJ_SUBSTANCE 和 LOC_CONTINENT

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

两类的召回率有比较明显的下降. 总之,以平均 F 值为评判标准,基于问题答案分类的答案抽取方法取得了良好的性能表现,尤其是加入深层句法信息使系统的性能得到明显提升.

从问题类型的角度看:比较特殊的是 HUM. PERSON 类和 OBJ. SUBSTANCE 类. 通过错误分析发现,HUM. PERSON 类的 MEM 分类器之所以没有取得较好的实验效果,是因为系统中人名识别的准确率比较低. HUM. PERSON 类的答案多为人名,原则上应该有一个统一的人名标记作为特征 POS. A,但目前 ICTCLAS 内嵌的命名实体识别功能中并没有做到这一点. 比如以下三个人名在句子中的词性标记没有一定之规,"毛/ nr 泽东/ nr ","赵/ nr 匡/ vg 胤/ g "和"卢/ j 照/ v 邻/ vg ",尽管都是人名但因为没有一个一致的标记,导致 MEM 分类器并不能从 POS. A 特征中学习到足够有用的信息,有时反而会学习到错误信息误导分类器的判别能力. 而 SVM 分类器在 OBJ. SUBSTANCE 类召回率较低则是由于此类问题的 f_3、f_5 匹配情况较差,原因是问句的主、宾语虽然在语义上与答句词序列密切相关,却不是严格的词与词对应,造成 SVM 将许多正例误判为反例. 这启示我们,如果要解决事实类的问题,要求问答系统必须有一个比较好的命名实体工具,并具有一定的语义分析能力.

其他三类的结果表现基本一致,其中 TIME. YEAR 类表现最为突出. 分析中我们发现,这一类的问句句型和答句句型都比较固定,另外它的 POS. A 特征也很一致,基本上都是"/ t ". 这种特征上的规则性是系统取得良好表现的主要原因.

和英文问答系统相比,目前最好的英文问答系统一般能够回答对 2/ 3 多一点的事实类问题,本文的答案抽取分类器达到或超过了这一水平,证明基于分类的答案抽取方法取得了初步的成功. 同时系统中的问题也提示我们面对事实类问题,应该有一个比较好的命名实体抽取工具. 另外,除了考虑正确答案的词性标记外,还应当考虑答案的其他特性(例如语义特征),从而使分类器能学习到更多有用的知识.

5 结论

本文采用了对候选答句进行分类的方法进行答案的抽取. 该方法首先对问句进行了细致的句型分析,在此基础上抽取出对分类有指导意义的问句特征集. 其次在对候选答句进行浅层句法分析的基础上提取出候选答句特征集. 然后把最大熵模型和支持向量机应用到由问句特征和候选答句特征组合得到的特征集上,通过训练得到答案分类器. 分类器良好的实验表现证明了这种新的答案抽取方法的可行性,该方法取得了初步的成功.

同时实验中的不足也启示我们今后应该在解决事实类的问题时开发一个面向针对问答系统的命名实体识别工具. 此外,深入探讨正确答案的其他特性将为我们改进该方法提供更多的指导. 最后,如何将该方法运用到更多的问题类型上,如何抽取出更有效的对分类有指导意义的特征是我们今后努力的方向.

致谢 本文使用了哈工大信息检索研究室语言技术平台中的问答系统问题集,特此致谢.

参考文献:

[1] Harabagiu S, Moldovan D, et al. FALCON: boosting knowledge for answer engines[A]. Proceedings of Ninth the Text Retrieval Conference[C]. Gaithersburg, Maryland, USA: NIST, 2000. 479 - 488.

[2] Lita L V, Carbonell J. Instance-based question answering: a data driven approach[A]. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)[C]. Morristown, NJ: ACL Press, 2004. 396 - 403.

[3] Sun A, Jiang M, Ma Y. An instance-based approach for pinpointing answers in Chinese question answering[A]. In Proceedings of 8th International Conference on Signal Processing[C]. Piscataway, NJ, USA: IEEE press, 2006. 1620 - 1623.

[4] Chu-Carroll J, Czuba K, Duboue P, et al. IBM's PIQUANT II in TREC2005[A]. Proceedings of the Fourteenth Text Retrieval Conference[C]. Gaithersburg, Maryland, USA: NIST, 2005.

[5] Ittycheriah A, Roukos S. IBM's statistical question answering system-TREC 11[C]. Proceedings of the TREC-2002 Conference[C]. Gaithersburg, Maryland, USA: NIST, 2002.

[6] Moldovan D, Harabagiu S, et al. LCC tools for question answering[A]. Proceedings of the Eleventh Text Retrieval Conference[C]. Gaithersburg, Maryland: NIST, 2002. 144 - 155.

[7] Echihabi A, Marcu D. A noisy-channel approach to question answering[A]. Hinrichs, E, Roth, D (Eds) Proceedings of 41st Annual Meeting of the Association for Computational Linguistics[C]. Morristown, NJ: ACL Press, 2003. 16 - 23.

[8] 郑实福,刘挺,秦兵,等. 自动问答综述[J]. 中文信息学报,2002,16(6):46 - 52.
Zhi Shifu, Liu Ting, Qin Bing, et al. Overview of question answering[J]. Journal of Chinese Infomation Processing, 2002, 16(6): 46 - 52. (in Chinese)

[9] 徐延勇,周献中,井祥鹤,等. 基于最大熵模型的汉语句子分析. 电子学报,2003,31(11):1608 - 1612.
Xu Yanyong, Zhou Xianzhong, Jing Xianghe. Chinese sentence parsing based on maximum entropy model[J]. Acta Electronica Sinica, 2003, 31(11): 1608 - 1612. (in Chinese)

[10] Berger A, Pietra S, Pietra V. A maximum entropy approach to natural language processing[J]. Computational Linguistics,

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1996,22(1):39 - 71.

[11] Joachims T. Text categorization with support vector machines: Learning with many relevant features[A]. Lecture Notes in Computer Science[C]. Berlin/ Heidelberg: Springer, 1998. 1398:137 - 142.

[12] 李蕾,钟义信. 自动文摘系统中基于全信息词典的复杂语句分析方法及其实现[J]. 电子学报,2000,28(8):104 - 109.
Li Lei, Zhong Yixin. CIL Based Algorithm Simplifying Analysis of Complex Chinese Sentences[J]. Acta Electronica Sinica,2000,28(8):104 - 109. (in Chinese)

[13] 北京语言学院句型研究小组. 现代汉语基本句型(续完)[J]. 世界汉语教学,1991,1:23 - 29.
Sentence pattern research group of Beijing Language Institute. Basic sentence pattern of Modern Chinese (end)[J]. Chinese Teaching in the world,1991,1:23 - 29. (in Chinese)

作者简介:

孙　昂　男,1982 年生于山东平邑,山东大学语言学学士,清华大学计算语言学硕士,现在美国纽约大学计算机系攻读博士学位. 主要研究方向:问答系统、信息检索等.

江铭虎　男,1962 年生于山东济南,电子工程博士,清华大学计算机系和比利时鲁汶大学电机工程系曾做博士后研究,德国海德堡大学医学院交叉学科计算中心访问教授. 现为清华大学人文学院计算语言学和清华大学心理学与认知科学中心教授. 主要研究方向:模式识别与人工智能、神经网络语言处理、自然语言处理.
E-mail:jiang.mh@tsinghua.edu.cn

贺一帆　男,1983 年生于上海,同济大学管理学学士,清华大学计算语言学硕士生. 主要研究方向:自然语言处理.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net