基于知网的中文问题自动分类

孙景广，蔡东风，吕德新，董燕举

（沈阳航空工业学院 自然语言处理研究室，辽宁 沈阳 110034）

摘 要： 问答系统能用准确、简洁的答案回答用户用自然语言提出的问题。问题分类是问答系统所要处理的第一步，分类结果的正确率直接影响后续工作的进行。本文提出了一种使用知网作为语义资源选择分类特征，并使用最大熵模型进行分类的新方法。以问题的疑问词、句法结构、疑问意向词、疑问意向词在知网中的首义原作为分类特征。在知网中选择的首义原能很好的表达问题焦点词的语义信息，从而可作为问题分类的一个主要特征。实验结果表明，该方法能有效提高问题分类的精度，大类和小类的分类精度分别达到了 90.75%和 79.8%。

关键词： 问答系统；问题分类；知网；最大熵模型；分类特征；

HowNet Based Chinese Question Automatic Classification

Jingguang Sun, Dongfeng Cai, Dexin Lv, Yanju Dong

(Natural Language Processing Laboratory, Shenyang Institute of Aeronautical Engineering, Shenyang, Liaoning, 110034)

Abstract: Question answering system can provides a precise and concise answer to a natural language query. Question classification is the first task of Question Answering System must carry out, and the precision of question classification has great effect on the following work. In this paper, we present a new method on feature extraction that use HowNet as semantic resource and use Maximum Entropy Model realize. We choose the interrogative words, syntax structure, question focus words and its first sememes as classification feature. For the first sememes in HowNet can express the main meaning of the question focus words, so it can be as an important feature. The experiment result show that this method can improve the precision of question classification: the classification precision of coarse classes and fine classes reaches 90.75% and 79.8% respectively.

Key Words: Question Answering System; question classification; HowNet; Maximum Entropy Model; classification feature;

1 引言

问答系统(Question Answering System)是一种计算机信息检索的高级形式，它和一般的检索系统的最大不同是：一、它以自然语言的形式提出问题，二、返回给用户的是与问题直接相关的一句话或一段话，而不是网页列表。例如对于问题："谁是中国太空第一人？"，搜索引擎返给用户的是很多包含关键词"中国"、"太空"、"第一人"的网页，由用户自己寻找正确答案[1]。而问答系统直接返给用户的就是这个问题的答案——"杨利伟"。因此，问答系统要比传统的基于关键字检索的搜索引擎得到的结果更加方便、快捷、高效[2]。

作者简介：孙景广（1981－），男，山东临沂人，硕士研究生 E-mail: sunjingguang@gmail.com

蔡东风（1958－），男，辽宁沈阳人，教授，博士 E-mail: cdf@ge-soft.com

问答系统一般分为问题理解、信息检索、答案抽取三个部分，几乎所有的问答系统在问题理解阶段都有问题分类这一过程。问题分类就是对于给定的问题，根据问题的答案类型把该问题映射到给定的语义类别中。问题分类是问答系统所要处理的第一步，分类结果的正确率直接影响后续工作的进行。

对英文问题自动分类的研究开展的比较早。最初采用了基于规则的方法，具有代表性的是 Dell Zhang 等人提出的采用 SVM（支持向量机）进行分类的方法[3]。后来又分别采用层次分类思想[4]和 SNoW（Sparse Network of Winnow）分类器[5]进行分类。采用语义词典（WordNet）进行分类[6]，也取得了不错的分类效果。

对中文问题自动分类的研究还不是很多，主要有复旦大学和哈尔滨工业大学等分别采用了 SVM 算法和改进的贝叶斯模型进行问题分类。后者对大类和小类的分类准确率分别达到了 86.62%和 71.92%[7]。

本文提出了一种采用知网(HowNet)作为语义资源选择行分类特征，并且利用最大熵模型(Maximum Entropy，ME）进行分类的新方法。

2 知网

知网[8](HowNet）是一个以汉英双语来表示概念与概念之间以及概念的属性之间关系的知识库。知网将客观世界中的词汇所代表的概念分为四大类：实体、事件、属性、属性值，并通过义原来标注概念。在知网中，义原是指不可再分的语义单位，共包含了大约2200个义原。

在知网中对于概念的定义是采用知识描述语言（Knowledge Database Mark-up Language，KDML）来描述。KDML 对概念的定义采用 DEF 语义表达式，DEF 描述了词语详尽的语义特征，如：

生日：DEF={time|时间:TimeSect={day|日},{ComeToWorld|问世:time={~}}}

词语在知网中的首义原就是该词语在 DEF 定义中出现的第一个义原，例如，“生日”的首义原就是“time|时间”。

3 问题分类体系

我们参考了国内外的一些问题分类体系，同时为了便于比较实验结果，本文采用了哈尔滨工业大学信息检索研究室的问题分类体系[7]，这种分类体系能较好的满足实际问题分类的需要，一共含有 7 个大类，每个大类根据实际情况又定义了一些小类，共 60 小类。

4 问题分类特征的选取

本文一共选择了四种分类特征：1、问题疑问词（IW） 2、句法结构（SS） 3、疑问意向词（QFW） 4、疑问意向词在知网中的首义原（FS）。其中进行大类分类时使用了这四种特征，而小类分类时只使用了其中的问题疑问词和疑问意向词在知网中的首义原。在下面的具体介绍中以问题 Q：“**CNN 第一次广播是什么时候**”为例进行说明。

4.1 问题疑问词（IW）

疑问词是一个问题中非常重要的信息，正确选择疑问词非常关键。首先根据中文中经常用使用的一些疑问词，例如“什么”、“为什么”，“怎么样”，“谁”，等建立一个疑问词词表 **T**，然后对问题 Q 进行分词和词性标注，得到“CNN/nx 第一/m 次/q 广播/vn 是/v 什么/r 时候/n”，选择其中标记为“/r”的词“什么”到 T 中进行查找，最后可以确定，问题 Q 的疑问词是“什么”。

4.2 句法结构的选择（SS）

对于用疑问词“什么”来提问的问题，研究发现这类问题在表达上有一些比较固定的结构，经过总结，选取了一些有代表性的句法结构。选取的具体依据如下：

首先，疑问词以及疑问词附近的具有名词特性的词和动词含有重要的信息，我们把具有名词特性的词性如：n、nx、ng、vn 等统一用“n”来表示，把动词的词性和疑问词的词性分别用“v”和“r”来表示。

其次，根据汉语句子表达的特点，“的”字结构比较常见，所以如果句子中包含有“的”字时，把“的”字的词性用“D”记录下来。

根据上面的两条规则，包含有疑问词“什么”的问题最终选择的句法结构就是 nvrD 的一个不同组合。然后对结构相似的句子进行了规约，最后选择了 12 种具有代表性的句法结构。

4.3 疑问意向词的选择（QFW）

疑问意向词[10]是表达“问题问的到底是什么”这样一个含义的概念。关于疑问意向词目前还没有明确的定义，一般认为，用户的疑问意图就是要得到一个未知信息，也就是问题中最能体现答案类型的词。比如例句中的“时候”。

根据汉语句子的表达习惯，在疑问词附近的词常常具有重要的信息，更能表达整个句子所要表达的语义信息，其中具有名词特性的词，也就是上面所说的标记为“n”的词显得更加重要。

通过进一步的分析发现，疑问词右边的标记为“n”的词比疑问词左边标记为“n”的词更加重要。本文疑问意向词的具体选择方法为：选择疑问词右边标记为“n”的词，如果疑问词的右边没有标记为“n”的词，就到疑问词的左边寻找。如果存在有多个标记为“n”的词，选择最多两个标记为“n”的词作为疑问意向词。比如，按照上面的方法，最终选择的疑问意向词为“时候”。

4.4 疑问意向词在知网中的首义原（FS）

义原的含义在前面有关知网的简介中已经进行了说明。汉语在具体表达中，多个表达方式不同的句子往往包含有相同或相似的语义信息。通过上面的分析，我们知道疑问意向词包含有重要的语义信息，如果两个问题的疑问意向词所表达的语义是一样的，就能在很大程度上判定这两个问题所表达的语义也一样。比如：对于例句 Q，也可以这样表达：“CNN 第一次广播是什么日期？”。通过上面的所说的方法首先能够判断它们的疑问意向词分别为“时候”和“日期”，判断它们在语义上表达的意思是否一样，就是接下来需要解决的问题。

2005 版知网中对 81447 个词汇进行了语义描述，定义了 157185 个概念记录。应该说这个规模能够较好的覆盖目前的开放域问答系统中出现的问题中的词。所以我们选择了知网作为语义资源进行分析。

“时候”和“日期”在知网中的 Def 定义分别为：“DEF={time|时间}”和“DEF={time|时间:TimeSect={day|日}}”，通过分析发现，它们的首义原能够很好的表达这两个词的语义，所以我们选择这两个词的首义原“time|时间”，作为这两个问题的一个非常重要的特征。在本文中，对有多个 Def 定义的词，对常用的词建立了词表，其他的词按照知网中的排列顺序选择第一个。

5 实验结果及错误分析

5.1 问题集

本实验使用了哈尔滨工业大学信息检索研究室和中科院自动化研究所模式识别国家重点实验室提供的问题集，并且按和文献[7]同样的方法将问题集划分为训练集和问题集，最后经去重整理后得到的句子分布情况如表 1，可以看出各种问题的分布大致均匀。

表 1 训练语料和测试语料中的问题分布

大类（Coarse）	训练集问题数目	测试集问题数目
人物（HUM）	320	179
地点（LOC）	876	352
数字（NUM）	1062	238
时间（TIME）	619	148

实体（OBJ）	982	225
描述（DES）	457	155

5.2 评价标准

对大类和小类的分类准确率采用公式（1）进行评价：

$$\text{分类准确率} = \frac{\text{测试集中正确分类的问题数}}{\text{测试集中总的问题数}} \times 100\% \qquad (1)$$

5.3 实验结果

本文在总结已有方法的基础上，采用先对大类进行分类，再对小类进行分类的方法。大类和小类的训练和测试也可以并行进行。具体的问题分类流程如图 1 所示：

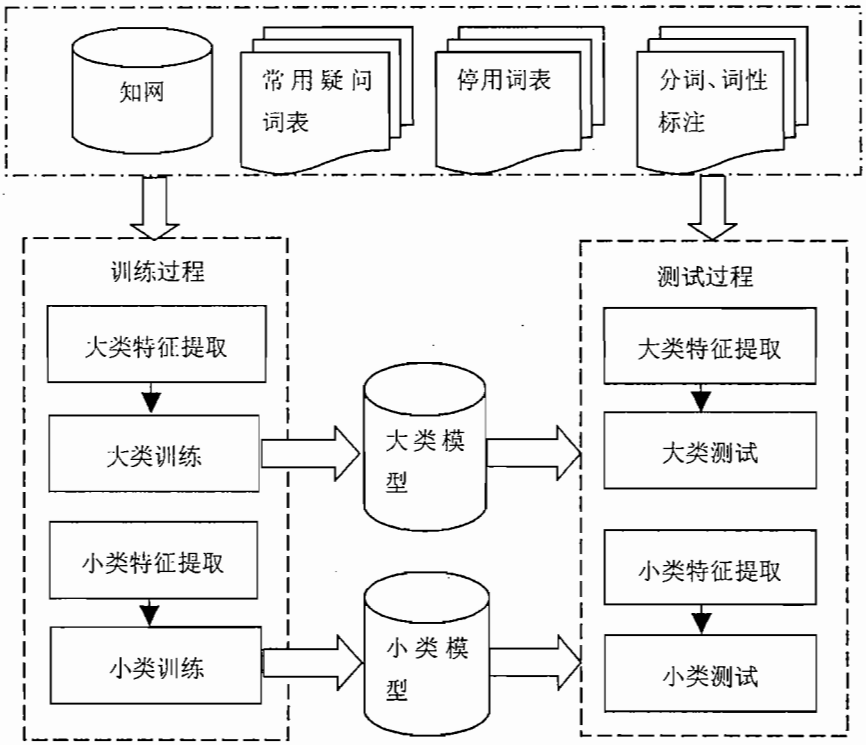

图 1 问题分类流程示意图

最大熵（Maximum Entropy，ME）模型是一个比较成熟的统计模型，适合于分类问题的解决。目前已成功应用于自然语言处理的多个领域，如：文本分类、词性标注、组块识别等等[9][10]。我们实验室对最大熵模型进行了比较深入的研究，本文也使用最大熵模型进行问题分类。把问题的疑问词（IW）、句法结构（SS）、疑问意向词（QFW）、疑问意向词在知网中的首义原（FS）作为分类特征，针对不同的特征组合进行了比较实验，结果如表 2 所示。

表 2 选用不同的特征时得到的分类结果

分类特征 / 准确率	IW	IW +SS	IW +QFW	IW +FS	IW+SS+QFW+FS
7 大类准确率	65.92%	69.01%	76.32%	84.51%	90.75%
60 小类准确率	35.31%	37.78%	67.54%	79. 8%	73.01%

5.4 实验结果分析

由表 2 可以看出，当选择不同的特征时，采用最大熵模型的分类结果有明显的不同。当综合所有特征时，7 个大类的准确率达到了最高值 90.75%，而 60 个小类的准确率为 73.01%并不是最高值，最高值出现在选择 IW 和 FS 作为特征，达到了 79.8%。分析原因在于：由于小类分的更细，过多的引入分类特征反而会造成数据稀疏，更不利于小类分类。所以，我们根据本文所选择的特征及最大熵模型的特点，进行大类分类时使用所有的特征，而进行小类分类时仅使用其中的两项——问题疑问词和疑问意向词在知网中的首义原。

通过对实验中的错误问题进行研究后发现，主要由以下原因造成：

第一：由于在选择疑问意向词时选择的是标记为“n”的词，在进行分词和词性标注时产生的错误，会对选择疑问意向词以及该词在知网中的首义原时产生错误。

第二：训练集中有些问题的分类存在错误或分类标准不一致，有些问题的分类可能不唯一。

第三：由于训练集中的问题有限，不能全面的覆盖所有的提问方式，对于训练集中没有出现过的问题类型，如果在测试集中出现，就很难正确分类。

第四：由于中文问题的表达方式中存在着省略的现象，很难判断省略的是什么内容。

6 总结和展望

从实验结果可以看出，本文采用的最大熵模型表现出了较好的性能。本文主要运用知网作为语义资源，从语义的角度进行特征选择，最终使 7 个大类的准确率达到 90.75%，60 个小类的准确率达到 79.8%，比同类实验结果[7]分别提高了 4.13%和 7.88%。

目前，疑问意向词以及疑问意向词在知网中的首义原的选择方法还可以进一步改进。并可以考虑加入知网提供的其它语义信息，进一步提高问题分类的准确率。

7 致谢

本文使用了哈尔滨工业大学信息检索研究室和中科院自动化研究所模式识别国家重点实验室提供的问题集。在此，对他们表示诚挚的感谢！

参考文献：

[1] 崔桓，蔡东风，苗雪雷. 基于网络的中文问答系统及信息抽取算法的研究[J]. 中文信息学报，2004，18(3):24-31

[2] 郑实福，刘挺，秦兵 et al. 自动问答综述[J]. 中文信息学报，2002，16(6)：46-52

[3] Dell Zhang,Wee Sun Lee. Question classification using support vector machines[A].In:the 26th ACM SIGIR[C].2003

[4] Xin li,Dan Roth.Learning Question classification using support vector machines[A].In: the 26th ACM SIGIR[C]. 2003

[5] Carlson,C.Cumby,J.Rosen,etal. The SNoW learning architecture[A]. In:UIUCDCS-R-99-2101, UIUC Computer Science Department[C] ,2004,451-458.

[6] Xin Li, Dan Roth. The Role of Semantic Information in Learning Question Classifiers[A].In: First International Joint Conference on Natural Language Processing[C],2004,451-458

[7] 文勖，张宇，刘挺 et al. 基于句法结分析的中文问题分类[J]. 中文信息学报，2006，20（2）：33－39

[8] 董振东,董强.知 网. http://www.keenage.com/zhiwang/c_zhiwang.html.

[9] Darroch, J.N, Ratcliff, D. Generalized Iterative Scaling for Log-Linear models. Annals of Mathematical Statistics, 43(5):1470-1480, 1972.

[10] 吕德新 中文自动问答系统中问题理解技术的研究[D] 沈阳航空工业学院硕士论文，2006 年 3 月