第8卷 第3期 2008年2月
1671-1819(2008)3-0672-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol.8 No.3 Feb. 2008
© 2008 Sci. Tech. Engng.

中文阅读理解系统的研究与实现

邢雪峰 艾丽蓉*
(西北工业大学计算机学院,西安 710072)

摘 要 阅读理解系统作为开发、评估和比较自然语言问答方法的可控测试平台,引起了自然语言领域越来越多学者的关注,设计并实现了一个中文的阅读理解系统,着重分析了问题分析、答案定位和答案提取等关键技术;并根据不同的问题类型,设计了不同的答案提取策略。实验表明,系统的性能比基准测试方法提高了近5个%。

关键词 阅读理解系统 自然语言处理 答案提取
中图法分类号 TP391.1; **文献标志码** A

近几年来,人们开始关注自动阅读理解系统的研究。阅读理解系统能够分析一篇用自然语言表示的文章,接着给出一系列问题,人们期待这个系统能够从文章中汲取信息自动产生问题的答案。阅读理解系统的任务是评估机器的阅读理解能力,同时能够为自然语言处理技术的应用提供一个测试平台,推动自然语言处理技术研究的进步。进一步与传统的基于文档的网上搜索引擎研究相比,阅读理解系统展示了一种新颖的信息检索方法。其研究成果对信息提取[1]、自动文摘[2]和问答系统[3]的研究也具有借鉴作用。

1 相关工作

阅读理解系统最初是由 MITRE Corporation 的一个研究小组提出的[4]。他们开发了第一个阅读理解系统——Deep Read,同时开发了 Remedia Corpora 语料库,并定义了评价系统性能的三种指标:Precision & Recall(准确率和召回率)、HumSent 准确率和 AutSent 准确率。Riloff 等开发了一个基于规则的阅读理解系统 Quarc[5]。Charniak 等使用了统计的方法开发了 Qspecific 系统[6]。Ng 等把机器学习的方法应用到阅读理解系统中[7]。Anand 等首次提出把阅读理解系统任务分成问题分析、答案定位和候选答案评价三个子任务[8],同时开发了系统 Spot,并在 CBC4Kids Corpora 语料上做了测试。相比较国外,国内从事这方面的研究还比较少,Kui Xu 等做了一些研究,并开发了 ChungHwa Corpus 中英文双语语料资源[9]。因此,本文实现了一个中文的阅读理解系统,以期对中文阅读理解的研究有更深刻的认识。

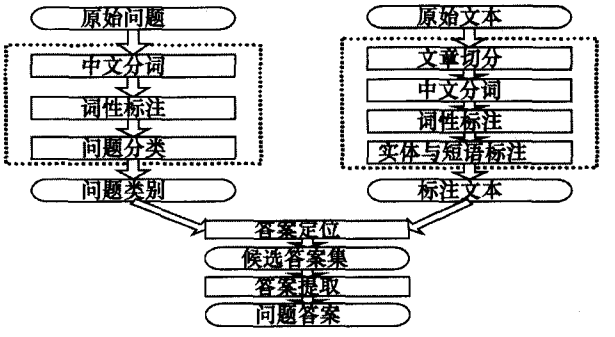

图1 系统结构

2 系统描述

系统结构如图1所示,原始文本按照标点符号(句号、叹号和问号)切分成若干个句子,经过分词、

2007年10月29日收到
第一作者简介:邢雪峰(1982—),男,硕士研究生,研究方向:智能信息处理。
*通信作者简介:艾丽蓉(1970—),女,副教授,研究方向:软件工程,信息安全与智能信息处理。

词性标注、短语与实体标注产生标注文本。同时,原始问题经过分词、词性标注和问题分类得到问题的类型,然后根据问题类型和标注文本,对问题可能的候选答案进行定位,得到问题的可能候选答案集。最后候选答案评价根据一定策略对候选答案集进行评价,评价最好的句子作为问题的最终答案。

系统所采用的语料资源是 CBC4Kids Corpus。由于该语料资源是纯英文的,首先进行了翻译工作。翻译由两个人进行,针对翻译不同的地方进行修正,给出最终的翻译结果,然后利用这个结果进行测试。

3 关键技术

阅读理解系统任务可分为问题分析、答案定位和答案提取三个子任务[8]。

3.1 问题分析

获得正确答案的一个重要步骤就是决定问题所期待的答案类型。可分为问题分类和问题归类。

问题分类应该有三个目标:第一,问题类型应该尽可能多地覆盖数据集上的问题;第二,每个类型应该很容易从问题中准确地提取出来;第三,每个类型所对应的答案应该能从文本中准确地提取出来。问题归类就是根据某些人工总结的规则把问题归结到某个类别中。这些规则可以总结如下:第一,可以根据某些诸如“谁”、“哪里”等特定的词和短语;第二,根据语义以及词法匹配知识。

考虑到现阶段阅读理解系统都是针对低年级学生的阅读理解水平进行测试的,简单地把问题分为时间,地点,人物,解释,数量和其他类型。对问题归类采取的是基于规则的方法,即通过问题中是否包含特定词语进行归类。

3.2 答案定位

答案定位就是在给定文本集中,根据问题类型标注问题的候选答案集。又称为语义标注,可分为命名实体标注和短语标注。

命名实体标注的任务是识别出文本中出现的专有名称和有意义的数量短语并加以归类。对实体识别的方法有基于规则和基于统计的方法,以及上述两种方法相结合的方法。短语标注就是识别出诸如动词短语,分句,甚至整个句子的可能答案。对标注解释型的短句可以利用以下规则:第一,词法规则:即找和解释相关的诸如“因为”、“由于”等特殊词汇;第二,篇章规则:即寻找诸如“这就是原因所在”等暗示其前面句子就是解释的短语。

命名实体标注是完备的,而短语标注是不完备的。在实体标注上,系统使用的是中回科学院的词法分析系统[10],该词法分析系统能够对文本中的人名,地点,时间,数量进行标注。在短语的标注上,使用是上述的词法规则对文本中的句子进行标注。

3.3 答案提取

答案提取就是依据某种策略从候选答案集中找到问题的最佳答案。为便于比较,使用了两个基本的评价方法:Bow (Bag of Words)方法[4]和标准信息检索中的 Tfidf 方法[6]。Bow 方法把文本和候选答案句子表示成词汇袋集合,计算候选答案和问题词汇重叠的个数,把重叠个数最大的候选答案句作为问题的答案。Tfidf 方法采用标准信息检索中的方法,把单个句子看作文档,如果问题中含有词汇 $\omega_1,\cdots,\omega_n$,词汇 i 在某个文档中的频率是 f_i,文档集中含有词汇 i 的个数是 n,那么该文档的得分是 $\sum_{i=1}^{n}\frac{f_i}{n_i}$。

上述方法都把整个文本看作是问题的候选答案集,Bow 方法把句子和问题匹配的词汇数作为选择问题答案的特征,而 Tfidf 方法也可看作是赋权值的词汇袋匹配方法,它们针对不同的问题,使用同一种答案提取策略,在实现上比较简单,易于操作。但没有考虑问题的类型和句子的语法和语义信息,易出现错报的情况。因此,针对不同的问题类型,这里设计了不同的答案提取策略。定义如下:Bow():词汇袋方法得出的问题答案;Tfidf():Tfidf 方法得出的问题答案;pS:人物型候选答案集;nS:数量型候选答案集;lS:地点型候选答案集;cS:解释型候选答案集;tS:时间型候选答案集;Q:问题;nS:问题的最终答案。

(1)人物类型问题设计的策略如图 2 所示:首先使用 Tfidf 方法得到问题的答案句,如果该答案句

不在人物候选答案集中,则依次查找其下一句和上一句。如果都不在人物候选答案集中,问题答案由Bow方法得到。

(2)地点类型问题设计的策略如图3所示:首先使用Tfidf方法得到问题答案,如果该答案句不在地点候选答案集中或者该答案句和问题匹配的词汇数小于6,就认为问题的答案不可靠,则用Bow方法得到问题的答案;如果该答案和问题的匹配词汇数小于4,就认为匹配词汇数太小不足以确定是问题答案,问题答案由Tfidf方法获得。

```
a nSentence = Tfidf ()
b If Aþ Contain (pS, nSentence)
  Then nSentence += 1;
c If Aþ Contain (pS, nSentence)
  Then nSentence -= 2;
d If Aþ Contain (pS, nSentence)
  Then nSentence = Bow ();
```

图2 人物策略

```
a nSentence = Tfidf ();
b If Aþ Contain (lS, nSentence) or
    WordMatch (nSentence, Q) < 6
  Then nSentence += Bow ();
c If WordMatch (nSentence, Q) < 4
  Then nSentence = Tfidf ();
```

图3 地点策略

(3)数量类型问题设计的策略如图4所示:问题中是否包含量词对获得正确答案影响很大,如果问题中包含量词,就用图4所示的数量规则获得答案;如果该答案不在数量候选答案集中,则由Tfidf方法得到问题答案。如果也不在数量候选答案集中,则返回由Bow方法得到的答案句。在设计的数量规则中,采用的是给句子打分的方法获得问题的答案,根据句子中是否包含问题中的量词、动词和其它词来赋以不同的分值,得分最高的句子作为问题的答案。

```
a If Contain (Q, 量词)
  Then nSentence = NumRule ();
b If Aþ Contain (n S, nSentence)
  Then nSentence = Tfidf ();
c If Aþ Contain (nS, nSentence)
  Then nSentence = Bow ();
```

图4 数量策略

```
a nSentence = Bow ();
b If AþContain (tS, nSentence)
  Then nSentence = Tfidf ();
c If AþContain (tS, nSentence)
  Then nSentence = TimeRule ();
```

图5 时间策略

(4)时间类型问题 设计的策略如图5所示:首先用Bow方法获得问题的答案,如果该答案句不在时间候选答案集中,则答案由Tfidf方法获得,如果该答案句也不在时间候选答案集中,问题答案则由图7所示的时间规则获得。时间规则是用给句子打分的方法获得问题答案的。根据句子中是否包含“不久”、“年代”等隐含性时间词和句子中是否包含问题中的动词以及其他词汇进行打分,得分最高的句子将作为问题的答案。

```
a If Contain (S, Qq)
  Then Score (S) += 6;
b If Contain (S, Qverb)
  Then nSentence += 3;
c If Contain (S, Qother)
  Then nSentence += 1;
```

图6 数量规则

```
a If Contain (S, (不久, 年代等))
  Then Score (S) += 6;
b If Contain (S, Qverb)
  Then nSentence += 3;
c If Contain (S, Qother)
  Then nSentence += 1;
```

图7 时间规则

(5)解释类型问题对该问题的回答是困难的,因为很多答案都隐含在文本中,很难对这些问题的答案进行标注。因此,设计的策略如图 8 所示,并没有考虑解释型候选答案集,而是和其他类型的问题使用同一个答案提取策略,即首先使用 Tfidf 方法获得问题的答案句,如果该答案句和问题匹配的词汇数小于 4,则由 Bows 方法得到问题的答案句。

```
a  nSentence = Tfidf ();
b  If  WordMatch (nSentence ,Q) < 4
   Then nSentence = Bow ();
```

图 8　其他策略

4　实验结果及分析

我们选用了 CBC4Kids Corpora 语料库中的训练集和测试集的各 30 篇文章进行了测试,采用的评价标准是 HumSent 准确率。系统的整体性能表现如表 1 所示。可以看出, Tfidf 方法比 Bow 方法的结果有了一定的提高,这和英文阅读理解系统的测试结果是一致的。此外,本文方法在改善系统回答准确率方面提高了近 5%。

表 1　整体系能表现

	Bow/%	Tfidf/%	本文方法/%
Train	65.5	68.5	72.0
Test	71.4	70.8	74.5
Total	68.5	69.7	73.3

系统在回答不同问题类型上的表现如表 2 所示:

表 2　系统在回答不同类型问题的表现

		Person /%	location /%	number /%	time /%	cause /%	other /%
Bows	Train	63.0	60.9	65.2	71.4	68.0	64.0
	test	73.9	73.9	82.6	45.5	73.9	73.1
Tfidf	train	70.4	65.2	60.9	71.4	72.0	66.7
	test	82.6	60.9	69.6	63.6	78.3%	70.5
本文方法	train	68.1	65.2	73.9	76.2	80.0	68.3
	test	69.6	68.1	87.0	68.2	73.9	74.4

可以看出,系统在回答其他类型问题时,本文方法比 Bow 方法和 Tfidf 方法要好。尤其在回答数量和时间类型问题上表现很出色,这主要是由于该类问题的候选答案能全面准确地从文本中标注出来,有利于我们策略的实施。而系统在回答人物和地点类型的问题时,表现不尽如人意。我们认为,对人物类型问题的回答,采用的策略是答案句中必须包人名,但对于"他加入了六人的登上运动队"等这样包含团体组织的答案,系统不能够正确地返回该答案,这就降低了该方法回答人物类型问题的准确率。在回答地点类型问题时,对答案是"在他的骨骼里发现了放射性元素铀"等句子,无法在标注阶段对其进行标注,用本文的策略很难得到问题的正确答案。

此外,系统对问题的分类还存在着不细,不准的缺点,如何对问题进行正确的分类将是下一步工作的重点。系统对实体和短语的标注还有很多错报和漏报情况,这影响了系统系统性能的表现,需要进一步的深入研究。答案提取是正确答案获取的关键技术,从系统表现来看,找到不同问题的关键特征,制定不同的答案提取策略并非易事。如何找到问题类型的本质特征将是以后研究的一个重点。

5　结束语

中文阅读理解的研究还刚刚起步,分类正确,定位全面和提取准确是阅读理解系统的目标。文章实现了一个简单的中文阅读理解系统,并根据不同的问题类型设计了不同的答案提取策略,系统性能比基准方法有了改善。但系统在问题分类,答案定位和答案提取方面还有许多不足。

参 考 文 献

1　Riloff E. Information extraction as a stepping stone toward story understanding. The MIT Press,1999

2　Morros A H, Kasper, G M, Adams D A. The effects and limitations of automated text condensing on reading comprehension. Information Systems Research, 1992; 3(1): 17—35

3　Hirschman L, Gaizauskas R. Natural language question answering: the view from here. Natural Language Engineering,7(4); 275—300

(下转第 681 页)

Realization of an Object Classifier on the Basis of Image Semantics

SUN Ji-feng , YUAN Chun-lin* ,QIU Wei-dong ,YU Ying-lin
(Institute of Electronics and Information, South China University of Technology,Guangzhou 510640,P. R. China)

[**Abstract**] The research of image semantics is an active field in present time. It remains to be unsolved on the problem of "semantic gap" between the low-level image features and image semantics. A new method for the object classification is proposed, in which low-level features are extracted from some image, then inputed to a BP neural network and trained in the teacher's guide. Once the training effectively finishs, by the excellent extending capacity of the net, the system can classify the objects in the image correctly, and give the location information of every object. In this way, the semantics of the image can get . Experimental results show the effectiveness of the proposed method on the object classification.

[**Key words**] image semantics low-level features BP neural network location informatio

(上接第675页)

4 Hirschman L,*et al.* Deep read: a reading comprehension system. In: Proceedings of the 37 th Annual Meeting of the Association for Computational Linguistics,1999:325—332

5 Riloff E, Thelen M. A rule-based question answering system for reading comprehension test. ANLP/NAACL,-2000:13—19

6 Charniak E, *et al.* : Reading comprehension programs in a statistical-language-processing Class. In : ANLP-NAACL 2000 Workshop: Reading Comprehension Testsas Evaluation for Computer-Based Language Understanding Systems,2000

7 Ng H T, Teo L H, Kwan L P. A machine learning approach to answering questions for reading comprehension tests. In: Proceedings of the 2000 Joint SIG-DAT Conference on Empirical Methods in Natural Language Processing and VeryLarge Corpora,2000

8 Anand P, Break E, *et al.* Fun with reading comprehension. Final Report of the Workshop 2000 of Language Engineering for Students and Professionals Integrating Research and Education, Reading Comprehension, in Johns Hopkins University,2000

9 Xu K, *Meng H , Design and development of a bilingual reading comprehension corpus*, submitted to the International Journal of Computational Linguistics and Chinese Language Processing, January 2005

10 http://www.i3s.ac.cn

Research and Realization of Chinese Reading Comprehension System

XING Xue-feng, AI Li-rong*
(School of Computer, Northwestern Polytechnical University, Xi'an 710072,P. R. China)

[**Abstract**] Reading comprehension systems have received increasing attentions within the NLP community as a controlled test-bed for developing, evaluating and comparing robust question answering methods. A chinese reading comprehension system is designed and realized, focused on the key technologies of the system, such as question analysis, answer location and answer extraction, and designed the different answer extraction strategy, according to different type of question. As the experiment illustrates, the performance of system improve approximately five percent than benchmark method.

[**Key words**] reading comprehension system natural language process answer extraction