基于多策略的单文档问答式信息检索技术

杜永萍 何 明
(北京工业大学计算机学院 北京 100124)

摘 要 单文档问答式信息检索,即是阅读理解(Reading Comprehension,简称RC)。该任务的目的在于理解一篇文档并对提出的问题返回答案句。提出了充分利用外部资源采用多策略技术来提高RC系统性能的方法,包括基于Web的答案模式匹配应用、词汇语义关联推理以及上下文辅助等策略。本方法使得RC系统性能在Remedia标准测试集上的性能得到提高。描述了不同策略对提高系统性能的有效性,t-test结果表明,运用答案模式匹配和词汇语义关联推理策略所得到的性能显著提高;同时分析了指代消解策略在系统中的关键作用;最后比较了RC任务和多文档问答式信息检索(Question Answering,简称QA)任务的差异性。

关键词 模式,阅读理解,问题回答,自然语言处理
中图法分类号 TP391 **文献标识码** A

Multi-strategy Based Single Document Question Answering

DU Yong-ping HE Ming
(Institute of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract Single document question answering is also called Reading Comprehension(RC), which attempts to understand a document and returns an answer sentence when posed with a question. We proposed an approach that adopted multi-strategy and utilized external knowledge to improve the performance of RC, including pattern matching with Web-based answer patterns, lexical semantic relation inference and context assistance. This approach gives improved RC performance on the Remedia corpus. The effectiveness of different strategy was analyzed and pairwise t-tests show the performance improvements due to Web-derived answer patterns and lexical semantic relation inference technique are statistically significant. In addition, the performance impact by the co-reference resolution was also discussed. Finally, the comparison between the task of RC and multi-document question answering(QA) was analyzed.

Keywords Pattern, Reading comprehension, Question answering, Natural language processing

1 引言

问题回答(Question Answering,简称QA)这一研究方向近年来受到了各研究机构广泛的关注。输入的查询是自然语言描述的问题、返回的问题的精确答案。问题回答融合了信息检索、信息抽取及自然语言处理技术,是一具有广泛应用前景的研究领域。该研究方向的发展也引发了另一相关领域的研究——阅读理解(Reading Comprehension,简称RC)。它是一个与问题回答非常类似的任务,目的是在给定的一篇文档中对于给定的问题返回正确答案,如同我们在英语测试(CET-4,CET-6等)中的阅读理解题目。这项研究在2000年后开始受到部分研究机构的关注,ANLP-NAACL workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems,以及另外一个在Johns Hopkins University召开的Summer Workshop: Technology for Reading Comprehension QA,吸引了自然语言处理研究领域的研究机构开始探讨这一新的研究方向。

阅读理解任务同一年一度的TREC会议(Text REtrieval Conference)(http://trec.nist.gov)中设立的问题回答任务类似。QA任务的目的在于从文档集中获取问题的答案,而RC任务则聚焦于单篇文档,理解一篇不受领域限制的文档,并对提出的问题返回一个答案句。Light等人对QA和RC 2个任务做了1s个详细的比较[1],发现TREC QA任务中大多数问题的答案在搜索的文档集中出现次数均多于1次。然而,RC任务在Remedia语料上,80%的问题答案句仅出现1次。也就是说,RC系统通常仅仅有1次机会获取到问题的答案。因此,RC系统迫切需要有外部扩展知识源来辅助进行深层次的文本分析,实现正确答案的获取。

MITRE Corporation的一个研究小组首先开始了阅读理解任务的研究,开发了第一个阅读理解系统Deep Read[2]。Deep Read系统的性能测试集是Remedia语料(该语料此后作为各研究单位阅读理解任务的标准测试集),该语料包含

到稿日期:2008-08-04 返修日期:2008-11-10 本文受国家自然科学基金青年基金(No. 60803086),北京工业大学博士科研启动基金(52007012200701)资助。

杜永萍(1977—),女,博士,讲师,主要研究方向为自然语言处理、信息检索,E-mail:ypdu@bjut.edu.cn;何 明(1975—),男,博士,讲师,主要研究方向为数据挖掘、人工智能。

115 篇英文文章，按照难易程度分为不同的等级 Level 2 到 Level 5。其中 Level 2 和 Level 5 的 55 篇文章用作训练，Level 3 和 Level 4 的 60 篇文章用作测试。每篇文章平均包含 20 条语句和 5 个问题（类型为：who，where，when，what，why）。MITRE 同时也定义了 HumSent 作为性能评价标准，即测试集中系统正确回答的问题比例。HumSent 评价中的标准答案是人工标注、检测文档并选择最合适的句子作为问题答案。经统计，Remedia 测试集上 11%的问题在文章中不存在正确答案。因此，RC 系统在 Remedia 语料上的性能上限只能达到 89% HumSent 准确率。

Hirschman 等人[2]在 Remedia 测试集上，得到的 HumSent 准确率为 36.6%。Ng 等人[3]采用决策树的机器学习方法得到了 39.3% HumSent 准确率；Rillof 等人[4]和 Charniak 等人[5]分别将性能提高到了 39.7%和 41%。他们采用了人工规则集。

本文介绍了采用模式匹配、语义关联推理以及上下文规则集技术的单文档问答式信息检索-阅读理解系统，系统结构如图 1 所示。

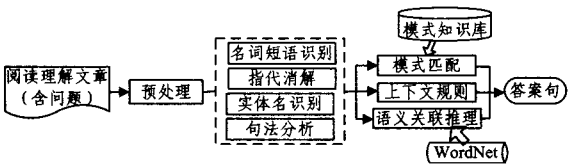

图 1　RC 系统结构组成

本文第 2 节介绍问题分析；第 3 节介绍系统的核心组成：采用多策略的答案句选取；第 4 节介绍答案句验证；第 5 节是实验结果分析与评价；第 6 节分析 QA 与 RC 之间的差异；最后是结论。

2　问题分析

问题分析是问答系统首要进行的工作，该过程的分析结果将对 RC 系统后续模块的性能有着重要的影响。图 2 描述了该模块的具体构成，主要由 3 个子模块 M1，M2 和 M3 构成。

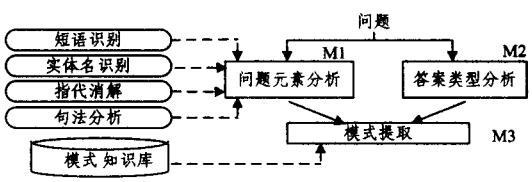

图 2　问题分析模块组成

M1：对问题中的各不同组成成分利用名词短语识别器、实体名识别器及句法分析器（Link Parser）[6]进行分析，以更好地服务于查询生成与模式提取。

示例问题：When did Lincoln free the slaves?

分析结果：动词（Q_Verb）：free

名词短语（Q_BNP）：the slaves

实体名（Q_PRN）：Lincoln

句子成分：主语（subject）-Lincoln

谓语动词（Verb）-free

宾语（object）-the slaves

状语（AdverbialModifier）-when

M2：确定问题的答案类型，这将对系统的答案获取提供一定的语义类型限制。系统中采用的答案类型分类体系如表 1 所列，为事实性问题设计。不同系统采用的答案类型分类体系是有差异的。分类类别多且细致，可以有效地排除似是而非的候选答案，但是也不能保证覆盖所有未见的新问题，这是绝大多数分类体系所面临的问题。

表 1　答案类型分类体系

LCN（地名）	PRN（人名）	ORG（机构名）
NUM（数字）	DAT（日期）	PCT（百分比）
MNY（货币）	ABBR（缩写）	BNP（名词短语）

示例问题答案类型：日期（DAT）

M3：从模式知识库中选取与该问题相应的答案抽取模式，以用作实现 QA 系统的答案抽取。模式知识库是经过模式自动学习和评价而得到的[7]。

示例问题答案抽取模式：

（1）Q_PRN Q_Verb Q_BNP in 〈A〉

（2）in 〈A〉，Q_PRN Q_Verb Q_BNP

（2）Q_PRN Q_Verb Q_BNP 〈A〉

……

这里，各种有关问句中的标识如 M1 模块分析所示；〈A〉标记表示问题的答案。以上不同的答案抽取模式分别表示包含答案的不同上下文出现的形式。

3　多策略答案句选取

RC 系统的关键模块是答案句选取。该模块实现阅读理解任务中答案句的选择、对篇章中的句子片段进行分析、采用不同的策略选取问题的候选答案句。以 BOW（Bag of Words）方法作为基本方法，而后逐步应用模式匹配、上下文规则以及语义关联推理技术提高准确率。图 3 所示为答案句选取模块的流程，由 M4，M5，M6，M7 4 个子模块构成。

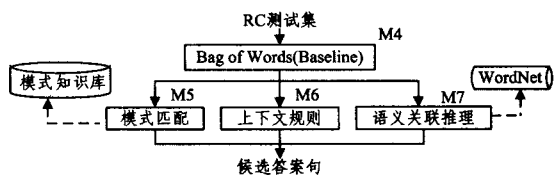

图 3　答案句选取模块组成

M4：将问题和文章中的句子表示为词的集合（Bag of Words）。对于给定的问题，BOW 方法选择与问题有最多匹配单词的候选答案句作为问题的最终答案。运用该方法时进行了 stemming 并去除了停用词。也就是说，如果两个词有相同的词根，则认为它们匹配。该方法是 RC 任务中常用的方法，本文以此方法的实验结果作为系统的 Baseline。

M5：由 M1 模块分析得到的各问题元素，对 M3 模块抽取到的答案模式进行实例化。在阅读理解文章句子中进行匹配，选取候选答案句。

示例问题答案抽取模式实例化：

（1）Lincoln freed the slaves in 〈A〉

（2）in 〈A〉，Lincoln freed the slaves

（3）Lincoln freed the slaves 〈A〉

模式匹配得到的候选答案句：Lincoln freed the slaves after the Civil War

M6：将上下文信息、答案类型的实体名信息进行融合。问题的答案类型可能是实体名（如表 1 所列），系统将在具有

最多匹配单词的候选答案句的上下文(前后各 2 个句子)中搜索包含实体名的句子。数据分析表明,所选择上下文窗口大小为 2 对于 when,who 和 where 类型的问题合适。

下面是根据问题的答案类型,运用上下文辅助策略用到的规则示例:

PRN rule: if ! *contain*(*S*, *PERSON*)
then *Search*(*Window*(*S*), *PERSON*)

TIM rule: if ! contain(*S*, *TIME*)
then Search(Window(*S*), *TIME*)

这里,"*S*"代表同问题具有最多匹配单词数目的候选答案句;"Window(*S*)"代表 *S* 的上下文,即前后各 2 个句子。运用示例如下:

Eg.:Who has a foolish idea ?

片段 t:Some man have foolish idea. Take Robert H. Goddard, for example.

最多匹配单词句子:Some man have foolish idea.

正确答案句:〈Anser〉Take Robert H. Goddard, for example.〈/Answer〉

M7:WordNet 是结构化词汇知识库,系统中它用作一个外部知识源,利用它包含的多种语义信息(同义、相似、上下位和蕴含)在词汇级挖掘问题和候选答案句之间的语义关联,进一步定位正确答案句。

Eg:What are the buildings in Washington like?

W= {buildings, Washington, like}

Answer Sentence: When you come to American capital, you will first find many of the beautiful, shining white buildings are built in the noble style of the ancient Greek temples and stand in wide avenues amid trees and fountains.

在正确答案句中用到了原问句中词汇"Washington"在 WordNet 中的同义词"American capital",而非原单词本身。由此可见,通过词汇蕴含关系确定了问句和答案句之间的语义关联。

4 答案句验证

在答案句选取模块,应用不同的技术可能抽取到不同的候选答案句。为了获得正确的答案句,需要进一步进行答案句验证。图 4 显示了答案句验证的流程,由 M8,M9 模块组成。

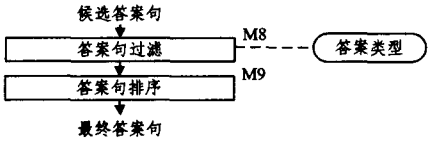

图 4 答案句验证模块组成

M8:由 M2 分析得到的答案类型对由答案句选取模块所选定的候选答案句进行过滤,排除不符合答案类型的候选答案句。该模块的实现需要和实体名识别技术结合使用。如:"who"类型问题,其答案类型是人名(PRN),而系统选取的候选答案句中不存在 PRN,不包含期望的答案类型,在该模块将其排除。

M9:当由前续模块识别出多个候选答案句时,需要一个好的排序机制将正确答案排在第一位,得分最高者作为问题的最终唯一答案。其中,模式匹配策略优先。可信率值高于 0.6 的答案模式同某候选答案句匹配时,该句将作为问题的最终正确答案。针对不同策略技术选取的候选答案句,采用 voting 机制实现唯一答案句的选取。

5 性能评价

在 Remedia 语料上测试 RC 系统性能。Remedia 测试集包含 60 篇文章,每篇文章包括 5 个问题。评价指标采用目前通用的 HumSent 准确率。

5.1 不同策略对系统性能的影响

对测试集经过加工(即短语识别、实体名识别以及指代消解)后,在采用 BOW(Bag-of-Words)基本方法(M4 模块)的基础 Baseline 上,不同的答案句抽取策略应用到系统中,完成 RC 任务,包括系统中的 M5(模式匹配)、M6(上下文规则)和 M7(语义关联推理)模块中的技术。在 Remedia 测试集上的结果如图 5 所示。

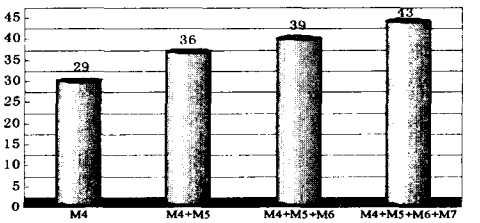

图 5 Remedia 测试集系统性能

如图 5 所示,模式匹配、上下文辅助以及语义关联推理 3 种不同策略对于系统相对性能的提高分别是 24.1%,8.3%和 10.3%。我们同时做了 t-test 来测试系统性能提高的显著性,如表 2 所列。据统计,模式匹配策略和语义关联推理显著提高系统性能,($p<0.05$),而上下文辅助策略则不然。

表 2 测试系统性能提高显著性的 *t-test* 结果

Pairwise comparison	*M4 && M4+M5*	*M4+M5 && M4+M5+M6*	*M4+M5+M6&& M4+M5+M6+M7*
t-test	$t(4)=3.24$	$t(4)=1.87$	$t(4)=3.92$
Results	$p=0.0316$	$p=0.135$	$p=0.017$

5.2 指代消解技术对系统性能的影响

阅读理解语料的一个显著特征是指代的出现较为频繁。指代是自然语言中常见的语言现象,指代消解是文本信息处理中的一个重要任务,是识别篇章中对现实世界中同一实体不同表达的过程,为篇章分析和自然语言理解系统提供完整详实的信息。我们在 Remedia 测试集上检测指代消解技术对系统性能的影响。其中,代词指代消解指代词和名词/名词短语间的指代问题;名词短语指代消解指解决名词短语间的指代问题。

系统应用不同的答案句抽取策略,完成 RC 任务,以不解决指代问题的系统结果作为 RC 系统的 Baseline,随后将代词指代消解、名词指代消解技术应用到系统。在 Remedia 测试集上,系统结果如表 3 所列。

表 3 Remedia 测试集上不同疑问词类型及系统总体(Overall)的 HumSent 准确率

疑问词类型 / 系统结果	When	Who	What	Where	Why	Overall
Baseline	43%	39%	36%	34%	26%	36%
Baseline+代词指代	51%	45%	39%	38%	28%	40%
Baseline+代词指代+名词指代	56%	49%	42%	41%	30%	43.3%

我们采取不同的策略完成了 RC 任务，但发现有一普遍现象使性能难以提高，示例问题如下：

Question: *When did James Barry become a doctor?*

Answer Sentence: *Dr. Barry finished her training in medicine at the young age of* 15.

该类型问题，如果不采用推理技术和背景知识源很难正确回答。而在 Remedia 语料上，该类问题出现比较频繁。在 TREC QA 任务中，Dan Moldovan 等人[8]和 Sanda Harabagiu 等人[9]成功地利用 WordNet 进行知识推理中的逻辑验证。推理技术也是 RC 系统中的有效策略。

6 阅读理解任务与问题回答任务的差异性

虽然 TREC QA 任务和阅读理解任务都是有关回答问题，但无论从模块流程、技术运用及评价等方面都存在着差异性。最显著的差别在于，TREC QA 任务的目的是从文档集合中寻找问题的答案，而阅读理解则是从和问题相关的单篇文档中寻找问题的答案。

Pranav 等人[10]比较了这 2 个不同任务的性能，发现答案的出现频率与系统性能有着密切的相关性：出现频率越高，系统越有可能返回正确答案。

面向指定文档的问题回答与面向文档集的问题回答任务有不同的挑战性，前者由于一个正确答案在该指定文档中通常只出现 1 次，因而通常也只有 1 次机会找到正确答案。而后者由于面向大规模的文档集，问题的答案可能会出现在不同的文档中，系统则有多次的机会找到正确答案。为了验证这一点，Pranav 等人[10]随机选取了 TREC8 的 51 个问题，人工在整个文档集中标识出包含问题答案的文档片段。表 4 所列为不同问题的答案出现频率的统计结果。

我们可以看到，有 13 个问题的正确答案在文档集中出现了 1 次，10 个问题的正确答案在文档集中出现了 2 次，依此，有 1 个问题的答案出现次数最多为 67 次。

表 4 TREC8 答案出现的频率统计

答案出现频率	1	2	3	4	5	6	7	8	9	12	18	27	28	61	67
问题数目	13	10	6	4	2	3	5	1	1	1	1	1	1	1	1

我们在 Remedia 的 300 个测试问题集中同样随机选取了 51 个问题做了相同的分析，有 90%的问题的正确答案句仅出现 1 次，其余问题的正确答案句出现频率最高为 4 次。统计结果如表 5 所列。

表 5 Remedia 答案出现频率统计

答案出现频率	1	2	3	4	5
问题数目	46	4	0	1	0

以上 2 项统计结果表明，对于阅读理解和 TREC QA 任务，答案出现的频率有很大的差异性，毕竟它们面向的是不同规模的语料。如上统计，在 TREC 语料集上，答案出现的平均频率为 7.1，75%的问题的答案出现 2 次或 2 次以上，55%的问题的答案出现 3 次或 3 次以上；在 Remedia 语料上，答案出现的平均频率约为 1.14，仅有 10%的问题出现 2 次或 2 次以上。

由上所述，阅读理解需要对文章内容进行深层次的分析，有着更大的挑战性。

结束语 阅读理解任务明显不同于基于 Web 的搜索引擎，该任务的目标是理解单篇文档，并能自动回答基于该文档提出的问题。RC 任务类似于 TREC QA 任务，后者在文档集包含的信息中寻找问题的答案，而 RC 任务仅仅面向单篇文档，因而需要利用外部知识源来辅助进行文章内容的深刻理解，实现问题回答。本文采用了不同的策略，利用不同的资源，在采用 BOW 方法的系统 Baseline 基础上提高 RC 系统的性能。

本文实现的 RC 系统中运用了模式匹配、语义关联推理和上下文辅助规则集。将不同的策略运用在 Remedia 标准测试语料集上，均使系统性能得到了提高。特别地，在 Remedia 语料集上，24.1%的相对性能提高是基于 Web 答案模式的运用，10.3%的相对性能提高是由于语义关联推理的运用，8.3%的相对性能提高是由于上下文辅助策略的运用，而系统总体性能达到 HumSemt 准确率为 43%，优于之前出现的最好结果。实验结果的分析可以帮助我们了解系统在不同的模块存在的具体问题，找到发生错误的问题根源，以便采取措施对其进行改进，达到优化系统的目标。

RC 任务是自然语言处理研究领域中更富有挑战性的研究方向，它需要深层次的技术及丰富的知识资源。今后的研究应探讨深层推理技术。该技术在 RC 任务中非常重要。

参 考 文 献

[1] Light M, Mann G S, Riloff E, et al. Analyses for Elucidating Current Question Answering Technology[J]. Natural Language Engineering, 1998, 7(4)

[2] Hirschman L, Light M, Breck E, et al. Deep Read: A Reading Comprehension System[C]// Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. 1999

[3] Ng H T, Teo L H, Kwan J L P. A Machine Learning Approach to Answering Questions for Reading Comprehension Tests[C]// Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 2000

[4] Riloff E, Thelen M. A Rule-based Question Answering System for Reading Comprehension Test[C]// ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-based Language Understanding Systems. 2000

[5] Charniak E, Altun Y, de Salvo Braz R, et al. Reading Comprehension Programs in a Statistical-Language-Processing Class[C]// ANLP/NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-based Language Understanding Systems. 2000

[6] Sleator D, Temperley D. Parsing English with a Link Grammar [C]// Third International Workshop on Parsing Technologies. 1993

[7] 杜永萍，黄萱菁，吴立德. 模式学习在 QA 系统中的有效实现[J]. 计算机研究与发展，2006，43(3)：449-455

[8] Moldovan D, Harabagiu S, Girju R, et al. LCC Tools for Question Answering[C]// Proceedings of the Eleventh Text REtrieval Conference. NIST, Gaithersburg, MD, 2002: 144-154

[9] Harabagiu S, Moldovan D, Pasca M, et al. Answering complex, list and context questions with lcc's question answering server [C]// Proceedings of the Tenth Text Retrieval Conference. NIST, Gaithersburg, MD, 2001: 355-361

[10] Anand P, Breck E, Brown B, et al. Fun with Reading Comprehension [C]// ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-based Language Understanding Systems. Seattle, Washington, 2000