(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103823794 B
(45)授权公告日 2016.08.17

(21)申请号 201410064433.3

(22)申请日 2014.02.25

(73)专利权人 浙江大学
地址 310027 浙江省杭州市西湖区浙大路38号

(72)发明人 黄妍 何莲珍

(74)专利代理机构 杭州天勤知识产权代理有限公司 33224
代理人 胡红娟

(51)Int.Cl.
G06F 17/27(2006.01)
G06N 5/00(2006.01)

(56)对比文件
US 2009/0306967 A1,2009.12.10,
CN 103049433 A,2013.04.17,
WO 2014/000764 A1,2014.01.03,
JACK MOSTOW et al..Using Automated Questions to Assess Reading Comprehension, Vocabulary, and Effects of Tutorial Interventions.《Cognition and Learning》.2004,第2卷103-140.
Jonathan C.Brown, Gwen A.Frishkoff, Maxine Eskenazi.Automatic question generation for vocabulary assessment.《HLT"05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing》.2005,819-826.

审查员 石爽

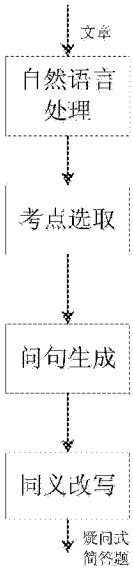
权利要求书3页 说明书12页 附图1页

(54)发明名称

一种关于英语阅读理解测试疑问式简答题的自动化命题方法

(57)摘要

本发明公开了一种关于英语阅读理解测试疑问式简答题的自动化命题方法,包括自然语言处理、考点选取、问句生成和同义改写四个步骤。该方法首先对输入的文章进行自然语言处理;然后基于词频密度、段落长度和句义近似度选取考点句子;根据词汇功能语法理论将陈述句转化为疑问句;最后对疑问句实施同义词替换和代词替换,形成疑问式简答题。本发明自动化命题方法由于加入了考点选取和同义改写,生成的疑问句可适用于阅读理解测试;同义改写部分由于采用了限定词汇范围和义项范围的方法,可突破语义消歧精度较低的瓶颈,实现准确率较高的同义词替换;疑问句生成部分由于同时参考了句法和语义信息,能够高效地生成类型多样、质量较高的问句。



CN 103823794 B

1.一种关于英语阅读理解测试疑问式简答题的自动化命题方法,包括如下步骤:

(1)自然语言处理;

1.1利用自动句法标注器对文章中的句子进行句法分析,得到句子的短语结构和词法信息;所述的词法信息包括句子中各名词的数以及各动词的时态和体态;

1.2利用自动语义角色标注器提取所述句子中述语动词指派给所在句中各句子成分的语义角色;所述的句子成分为单词、短语或从句;

1.3利用自动指代消解器提取所述句子中代词所指的句子成分;

1.4利用自动词汇范畴标注器提取所述句子中实词和固定短语的词汇范畴;

1.5利用语料库结合HAL法与LSA法,计算得到词典范围内所有单词的语义向量;

(2)考点选取;

2.1计算文章中句子的词频密度;

2.2计算文章中每个段落应选的考点数目;

2.3取所述句子中所有单词的语义向量的几何中心作为句子的语义向量,进而计算文章中每个句子与其他句子的句义近似度;

2.4按词频密度从高到低的顺序对文章中句子进行排序,依次判断每个句子是否被选为考点;

(3)问句生成;

3.1对于被选为考点的句子,根据句子的词法信息和各句子成分的语义角色建立句子基于词汇功能语法理论的功能结构;

3.2使功能结构中的独立功能体均作为提问对象;所述的独立功能体是指功能结构中以子功能结构作为明细的属性,其包括主语、宾语、间接宾语以及附加语;

3.3对于任一提问对象,确定该提问对象的中心语,进而根据中心语的词汇范畴以及提问对象的语义角色确定提问对象的疑问词;

3.4在被选为考点的句子中使该疑问词代替提问对象,进而根据所述的短语结构和功能结构对该句子中的句子成分做主谓倒装和时数一致性调整,生成以该疑问词引导的特殊疑问句;

3.5根据步骤3.3～3.4遍历每一个提问对象,生成多个特殊疑问句;

(4)同义改写;

4.1对文章中的实词或固定短语进行语义消歧,以确定实词或固定短语在特殊疑问句中的语义;

4.2对于特殊疑问句中的任一实词或固定短语,判断该实词或固定短语的语义是否为词典中该实词或固定短语的高频义项,若是则进入步骤4.3,若否,则不对该实词或固定短语做同义改写;

4.3根据语义利用词典获取该实词或固定短语的同义词集合,依次对集合中的同义词进行判断:对于集合中的任一同义词,判断该实词或固定短语的语义是否也是词典中该同义词的高频义项,若是则进入步骤4.4,若否,则判断集合中的下一个同义词;

4.4判断该同义词是否超出阅读理解测试所指定的词汇范围,若否,则将该同义词替换该实词或固定短语,若是,则判断集合中的下一个同义词;

4.5根据步骤4.2～4.4遍历特殊疑问句中的所有实词或固定短语;

4.6对于特殊疑问句中的任一代词，判断该代词所指的句子成分是否也在该特殊疑问句中，若是，则不对该代词做同义改写，若否，则进一步判断该特殊疑问句中是否存在另一个代词所指的句子成分与该代词的所指相同且该另一个代词在特殊疑问句中处于该代词前面的情况，若是，则不对该代词做同义改写，若否，则用该代词所指的句子成分替换该代词。

2.根据权利要求1所述的自动化命题方法，其特征在于：所述的步骤1.5中结合HAL法与LSA法得到词典范围内所有单词的语义向量的具体过程为：首先，从语料库中获取单词共现关系矩阵，该矩阵中任一元素的取值为该元素所在行和列分别对应的两个单词在语料库中共同出现的次数；然后，对该共现关系矩阵进行奇异值分解，得到词典范围内所有单词的语义向量。

3.根据权利要求1所述的自动化命题方法，其特征在于：所述的步骤2.1中根据以下公式计算句子的词频密度：

$$d_{core}=\frac{\sum_{i=1}^{n}w_i}{n}$$

其中：d_{core}为句子的词频密度，n为句子中的单词个数，w_i为句子中第i个单词在整篇文章中出现的次数。

4.根据权利要求1所述的自动化命题方法，其特征在于：所述的步骤2.2中根据以下公式计算文章中每个段落应选的考点数目：

$$N_j=m_j\times\frac{N_{tf}}{m}$$

其中：N_j为文章中第j段落应选的考点数目，m_j为第j段落中的句子数目，N_{tf}为文章规定应选的考点数目，m为文章中的句子数目，j为段落序号。

5.根据权利要求1所述的自动化命题方法，其特征在于：所述的步骤2.3中根据以下公式计算文章中每个句子与其他句子的句义近似度：

$$sim=\frac{\mathbf{s}_1^{\mathrm{T}}\mathbf{s}_2}{\|\mathbf{s}_1\|\|\mathbf{s}_2\|}$$

其中：s_1和s_2分别表示文章中任意两个句子的语义向量，sim为这两个句子的句义近似度。

6.根据权利要求1所述的自动化命题方法，其特征在于：所述的步骤2.4中判断每个句子是否被选为考点的评判标准如下：首先，对于待判断的句子，若其所在段落已选的考点数目已达到应选的考点数目，则不将其选为考点；若其所在段落已选的考点数目未达到应选的考点数目且其与文章中所有已被选为考点的句子的句义近似度均低于预设的近似度阈值，则将其选为考点并存储记录，否则不将其选为考点。

7.根据权利要求1所述的自动化命题方法，其特征在于：所述的步骤3.1中根据以下规则建立句子基于词汇功能语法理论的功能结构：

使句子中述语动词的原形作为功能结构或直联子功能结构中谓词的明细；所述的直联

子功能结构是指包含该述语动词的最小句子成分在所述功能结构中所对应的子功能结构；

使句子中述语动词的时态和体态分别作为功能结构或直联子功能结构中时态和体态的明细；如果该述语动词的时态或体态不完整，相应的直联子功能结构中时态或体态的明细继承上一级子功能结构中时态或体态的明细；

根据Propbank的语义角色标注体系，使句子中述语动词指派的序号最低的核心语义角色所对应的句子成分作为以该述语动词作谓词的功能结构或直联子功能结构中主语的谓词的明细；进而确定该句子成分的中心语，若中心语为名词，则将中心语的数作为所述主语的数的明细，若中心语为非名词，则令所述主语的数的明细为单数；

若句子的述语动词指派了至少两个核心语义角色，则使其中序号次低的核心语义角色对应的句子成分作为以该述语动词作谓词的功能结构或直联子功能结构中宾语的谓词的明细；

若句子的述语动词指派了至少三个核心语义角色，则使其中序号第三低的核心语义角色对应的句子成分作为该述语动词作谓词的功能结构或直联子功能结构中的间接宾语的谓词的明细；

若句子的述语动词还指派了若干附属语义角色，则将该若干附属语义角色对应的句子成分作为以该述语动词作谓词的功能结构或直联子功能结构中各对应附加语的明细。

8.根据权利要求1所述的自动化命题方法，其特征在于：所述的步骤3.3中确定提问对象的疑问词的标准如下：

若提问对象的语义角色为核心语义角色，则进而判断提问对象的中心语的词汇范畴：若词汇范畴为\.person，则令提问对象的疑问词为who；若词汇范畴为其他，则令提问对象的疑问词为what；

若提问对象的语义角色为附属语义角色中的时间，则进而判断提问对象的中心语的词汇范畴：若词汇范畴为\.duration，则令提问对象的疑问词为how long；若词汇范畴为其他，则令提问对象的疑问词为when；

若提问对象的语义角色为附属语义角色中的场所，则进而判断提问对象的中心语的词汇范畴：若词汇范畴为\.location、\.address、\.factory、\.geography或\.organization，则令提问对象的疑问词为where；若词汇范畴为其他，则令提问对象的疑问词为how；

若提问对象的语义角色为附属语义角色中的原因或目的，则令提问对象的疑问词为why；

若提问对象的语义角色为附属语义角色中的方式，则令提问对象的疑问词为how。

一种关于英语阅读理解测试疑问式简答题的自动化命题方法

技术领域

[0001] 本发明属于语言自动化测试技术领域，具体涉及一种关于英语阅读理解测试疑问式简答题的自动化命题方法。

背景技术

[0002] 计算机化是现代教育测试的重要发展方向。目前语言测试在施测环节已能够实现计算机自适应测试，在评分环节能够实现主观题机器自动评分，然而在命题环节，自动化水平依然很低，命题者基本仅在文字编辑和词典查询方面获取计算机的辅助。

[0003] 语言测试命题环节的计算机化具有迫切性。在标准化阅读理解测试开发中，人工命题的成本很高，效率却比较低。命题者需接受专业培训，还要经历繁杂的命题环节，包括改编文章、寻找考点、编写和研磨题目，以及试测题目。即使是经验丰富的命题者，也无法准确预知题目质量，试测后只有部分题目得以采用，这些问题导致大规模题库难以建立，进而阻碍了计算机自适应阅读理解测试的发展。

[0004] 关于阅读理解测试自动化命题的研究较少。Ruslan Mitkov和Le An Ha在标题为Computer-aided generation of multiple-choice tests(Proceedings of the2003Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics Workshop on Building Educational Applications Using Natural Language Processing,2003,17-22)的文献中提出基于浅层句法分析识别短语，根据语料库词频和WordNet构造干扰项生成多项选择题的命题方法。其研究表明，与直接命题相比，命题者在计算机所生成题目的基础上修改时效率可提高十多倍，且最终编写出来的题目质量更高。然而，该研究中计算机产生的问题类型有限，提问对象只限于名词，疑问词只限于which和what。

[0005] Jack Mostow和Wei Chen在标题为Generating instruction automatically for the reading strategy of self-questioning(Proceedings of the2009Conference on Artificial Intelligence in Education:Building Learning Systems that Care:From Knowledge Representation to Affective Modeling,2009,465-472)的文献中提出基于情境模型和题目模板产生简答题的方法，该方法只能针对具有人物情节的语篇命题，提问对象的范围限于人物心理状态。

[0006] Michael Heilman和Smith Noah在标题为Good question!Statistical ranking for question generation(Human Language Technologies:The2010Annual Conference of the North American Chapter of the Association for Computational Linguistics,2010,609-617)的文献中提出基于短语结构生成问句的方法。该方法没有考虑句子成分间的语义关系，且只依据特定词语这种表层信息决定疑问词，准确度不够高。

[0007] PrashanthMannem,RashmiPrasad和AravindJoshi在题目为Question generation from paragraphs at UPenn:QGSTEC system description(Proceedings of Question Generation2010,2010,84-91)的文献中提出基于语义角色生成问句的方法。该方法虽然考

虑了语义关系，由于采用针对专有名词的命名实体识别方法区分词汇范畴，无法把普通名词作为提问对象。

[0008] Xuchen Yao，GosseBouma和Yi Zhang在题目为Semantics-based question generation and implementation(Dialogue&Discourse，2012，11-42)的文献中提出基于最小递归语义理论生成问句的方法，该方法利用了深层语义结构，可生成质量较高的问句，但是该方法计算量大，效率较低。

[0009] 上述方法生成的问句不适用于阅读理解测试，因为这些问句未经过同义改写，而且这些方法未包含针对测试筛选题目的机制。阅读理解指对信息通过字形、语音和语义编码抽象出意义的过程。如果题目仅仅是将考点从陈述句式转为疑问句式，被试可能无需理解，仅凭字形匹配就能回答题目。另外一方面，语言测试的本质是根据被试对有限题目的答题情况推测其语言能力；阅读理解测试主要考察被试从语篇中获取信息的能力，选择语篇中哪些部分作为考点应从该部分是否具有文章信息的代表性出发。合理的考点应能够体现文章的核心信息，全面但不重复地涉及各个语义群。

发明内容

[0010] 针对现有技术所存在的上述技术问题，本发明提供了一种关于英语阅读理解测试疑问式简答题的自动化命题方法，能够对输入的文章选出符合核心性、全面性和相互独立性的考点，通过疑问句转换和同义改写，生成事实型简答题。

[0011] 一种关于英语阅读理解测试疑问式简答题的自动化命题方法，包括如下步骤：

[0012] (1)自然语言处理；

[0013] 1.1利用自动句法标注器对文章中的句子进行句法分析，得到句子的短语结构和词法信息；所述的词法信息包括句子中各名词的数以及各动词的时态和体态；

[0014] 1.2利用自动语义角色标注器提取所述句子中述语动词指派给所在句中各句子成分的语义角色；所述的句子成分为单词、短语或从句；

[0015] 1.3利用自动指代消解器提取所述句子中代词所指的句子成分；

[0016] 1.4利用自动词汇范畴标注器提取所述句子中实词和固定短语的词汇范畴；

[0017] 1.5利用语料库结合HAL法(Hyperspace Analogue to Language，多维空间类比分析法)与LSA法(Latent Semantic Analysis，潜在语义分析法)，计算得到词典范围内所有单词的语义向量；

[0018] (2)考点选取；

[0019] 2.1计算文章中句子的词频密度；

[0020] 2.2计算文章中每个段落应选的考点数目；

[0021] 2.3取所述句子中所有单词的语义向量的几何中心作为句子的语义向量，进而计算文章中每个句子与其他句子的句义近似度；

[0022] 2.4按词频密度从高到低的顺序对文章中句子进行排序，依次判断每个句子是否被选为考点；

[0023] (3)问句生成；

[0024] 3.1对于被选为考点的句子，根据句子的词法信息和各句子成分的语义角色建立句子基于词汇功能语法理论的功能结构；

[0025] 3.2使功能结构中的独立功能体均作为提问对象；所述的独立功能体是指功能结构中以子功能结构作为明细的属性，其包括主语、宾语、间接宾语以及附加语；

[0026] 3.3对于任一提问对象，确定该提问对象的中心语，进而根据中心语的词汇范畴以及提问对象的语义角色确定提问对象的疑问词；

[0027] 3.4在被选为考点的句子中使该疑问词代替提问对象，进而根据所述的短语结构和功能结构对该句子中的句子成分做主谓倒装和时数一致性调整，生成以该疑问词引导的特殊疑问句；

[0028] 3.5根据步骤3.3～3.4遍历每一个提问对象，生成多个特殊疑问句；

[0029] (4)同义改写；

[0030] 4.1对文章中的实词或固定短语进行语义消歧，以确定实词或固定短语在特殊疑问句中的语义；

[0031] 4.2对于特殊疑问句中的任一实词或固定短语，判断该实词或固定短语的语义是否为词典中该实词或固定短语的高频义项，若是则进入步骤4.3，若否，则不对该实词或固定短语做同义改写；

[0032] 4.3根据语义利用词典获取该实词或固定短语的同义词集合，依次对集合中的同义词进行判断：对于集合中的任一同义词，判断该实词或固定短语的语义是否也是词典中该同义词的高频义项，若是则进入步骤4.4，若否，则判断集合中的下一个同义词；

[0033] 4.4判断该同义词是否超出阅读理解测试所指定的词汇范围，若否，则将该同义词替换该实词或固定短语，若是，则判断集合中的下一个同义词；

[0034] 4.5根据步骤4.2～4.4遍历特殊疑问句中的所有实词或固定短语；

[0035] 4.6对于特殊疑问句中的任一代词，判断该代词所指的句子成分是否也在该特殊疑问句中，若是，则不对该代词做同义改写，若否，则进一步判断该特殊疑问句中是否存在另一个代词所指的句子成分与该代词的所指相同且该另一个代词在特殊疑问句中处于该代词前面的情况，若是，则不对该代词做同义改写，若否，则用该代词所指的句子成分替换该代词。

[0036] 所述的步骤1.5中结合HAL法与LSA法得到词典范围内所有单词的语义向量的具体过程为：首先，从语料库中获取单词共现关系矩阵，该矩阵中任一元素的取值为该元素所在行和列分别对应的两个单词在语料库中共同出现的次数；然后，对该共现关系矩阵进行奇异值分解，得到词典范围内所有单词的语义向量。

[0037] 所述的步骤2.1中根据以下公式计算句子的词频密度：

[0038] $$d_{core}=\frac{\sum_{i=1}^{n}w_i}{n}$$

[0039] 其中：d_{core}为句子的词频密度，n为句子中的单词个数，w_i为句子中第i个单词在整篇文章中出现的次数。

[0040] 所述的步骤2.2中根据以下公式计算文章中每个段落应选的考点数目：

[0041] $$N_j=m_j\times\frac{N_{tf}}{m}$$

[0042] 其中：N_j为文章中第j段落应选的考点数目，m_j为第j段落中的句子数目，Nt_f为文章规定应选的考点数目，m为文章中的句子数目，j为段落序号。

[0043] 所述的步骤2.3中根据以下公式计算文章中每个句子与其他句子的句义近似度：

[0044] $$sim = \frac{\mathbf{s}_1^{\mathrm{T}}\mathbf{s}_2}{\|\mathbf{s}_1\|\|\mathbf{s}_2\|}$$

[0045] 其中：s_1和s_2分别表示文章中任意两个句子的语义向量，sim为这两个句子的句义近似度。

[0046] 所述的步骤2.4中判断每个句子是否被选为考点的评判标准如下：

[0047] 首先，对于待判断的句子，若其所在段落已选的考点数目已达到应选的考点数目，则不将其选为考点；若其所在段落已选的考点数目未达到应选的考点数目且其与文章中所有已被选为考点的句子的句义近似度均低于预设的近似度阈值，则将其选为考点并存储记录，否则不将其选为考点。

[0048] 所述的步骤3.1中根据以下规则建立句子基于词汇功能语法理论的功能结构：

[0049] 使句子中述语动词的原形作为功能结构或直联子功能结构中谓词的明细；所述的直联子功能结构是指包含该述语动词的最小句子成分在所述功能结构中所对应的子功能结构；

[0050] 使句子中述语动词的时态和体态分别作为功能结构或直联子功能结构中时态和体态的明细；如果该述语动词的时态或体态不完整，相应的直联子功能结构中时态或体态的明细继承上一级子功能结构中时态或体态的明细；

[0051] 根据Propbank(命题树库)的语义角色标注体系，使句子中述语动词指派的序号最低的核心语义角色所对应的句子成分作为以该述语动词作谓词的功能结构或直联子功能结构中主语的谓词的明细；进而确定该句子成分的中心语，若中心语为名词，则将中心语的数作为所述主语的数的明细，若中心语为非名词，则令所述主语的数的明细为单数；

[0052] 若句子的述语动词指派了至少两个核心语义角色，则使其中序号次低的核心语义角色对应的句子成分作为以该述语动词作谓词的功能结构或直联子功能结构中宾语的谓词的明细；

[0053] 若句子的述语动词指派了至少三个核心语义角色，则使其中序号第三低的核心语义角色对应的句子成分作为该述语动词作谓词的功能结构或直联子功能结构中的间接宾语的谓词的明细；

[0054] 若句子的述语动词还指派了若干附属语义角色，则将该若干附属语义角色对应的句子成分作为以该述语动词作谓词的功能结构或直联子功能结构中各对应附加语的明细。

[0055] 所述的步骤3.3中确定提问对象的疑问词的标准如下：

[0056] 若提问对象的语义角色为核心语义角色，则进而判断提问对象的中心语的词汇范畴：若词汇范畴为\.person，则令提问对象的疑问词为who；若词汇范畴为其他，则令提问对象的疑问词为what；

[0057] 若提问对象的语义角色为附属语义角色中的时间，则进而判断提问对象的中心语的词汇范畴：若词汇范畴为\.duration，则令提问对象的疑问词为how long；若词汇范畴为其他，则令提问对象的疑问词为when；

[0058] 若提问对象的语义角色为附属语义角色中的场所，则进而判断提问对象的中心语的词汇范畴：若词汇范畴为\.location、\.address、\.factory、\.geography或\.organization，则令提问对象的疑问词为where；若词汇范畴为其他，则令提问对象的疑问词为how；

[0059] 若提问对象的语义角色为附属语义角色中的原因或目的，则令提问对象的疑问词为why；

[0060] 若提问对象的语义角色为附属语义角色中的方式，则令提问对象的疑问词为how。

[0061] 本发明自动化命题方法由于加入了考点选取和同义改写，生成的疑问句可适用于阅读理解测试；同义改写部分由于采用了限定词义的方法，可突破语义消歧精度较低的瓶颈，实现准确率较高的同义词替换；疑问句生成部分由于结合了句法和语义信息，能够高效地生成类型多样、质量较高的问句。

附图说明

[0062] 图1为本发明命题方法的步骤流程示意图。

具体实施方式

[0063] 为了更为具体地描述本发明，下面结合附图及具体实施方式对本发明的技术方案进行详细说明。

[0064] 本实施方式从1989年至2006年间的CET4阅读真题中提取所有完全以简答题形式命题的文章共7篇作为输入。

[0065] 如图1所示，一种关于英语阅读理解测试疑问式简答题的自动化命题方法，包括如下步骤：

[0066] (1)自然语言处理

[0067] 1)利用自动句法分析器Charniak's Parser对文章中的句子进行句法分析，得到句子的短语结构和词法信息(该词法信息包括句子中各名词的数以及各动词的时态和体态)；

[0068] 2)利用自动语义角色标注器llinois Semantic Role Labeler提取所述句子中的述语动词指派给句子各成分的语义角色；所采用的Propbank的语义角色标注体系如表1所示：

[0069] 表1

[0070]

A#(核心语义角色)			
A0	典型施事	A1	典型受事

[0071]

A2~A6	其他		
AM(附属语义角色)			
LOC	场所	CAU	原因
TMP	时间	EXT	程度
PNC	目的	DIS	话语连接
MNR	方式	MOD	情态
NEG	否定	ADV	其他修饰
DIR	方向		

[0072] 3)利用自动指代消解器emPronoun提取句子中代词所指的句子成分；

[0073] 4)利用自动词汇范畴标注器Super-sense Tagger提取句子中实词和固定短语的词汇范畴；

[0074] 5)由1989年至2010年间CET4阅读真题的所有文章组成外部语料库。该语料库共有55592个单词形符，4008个单词类符。利用Perl程序模块Text-SenseCluster获取4008×4008的单词共现关系矩阵，对该矩阵进行奇异值分解，保留前100个奇异值，获得4008个单词的语义向量。

[0075] (2)考点选取

[0076] 本实施方式对每篇文章选4个考点。对于一篇文章，按照句子词频密度，自然段长度和句义近似度选取考点，具体执行如下步骤：

[0077] 1)计算文章中句子的词频密度；

[0078] 由于核心信息在文章中通常以较多词汇反复强调，对文章进行词频统计和排序，序位越前的单词与核心信息相关的可能性越高。在以句子为考点选取单位的情况下，本发明用句子的词频密度衡量句子核心度，根据以下公式计算句子的词频密度：

[0079] $$d=\frac{\sum_{i=1}^{n}w_i}{n}$$

[0080] 其中：d为句子的词频密度，n为句子中的单词个数，w_i为句子中第i个单词在整篇文章中出现的次数。

[0081] 2)计算文章中每个段落应选的考点数目；

[0082] 由于在通常情况下，自然段是由多个语义紧密关联的句子构成的语义单位，为保证所选考点尽可能全面地涉及各个语义群，本发明以自然段为考点选取的基础，并使各自然段应选考点数目与该自然段的句子数成正比，根据以下公式计算文章中每个段落应选的考点数目：

[0083] $$N_j=m_j\times\frac{N_{tf}}{m}$$

[0084] 其中：N_j为文章中第j个段落应选的考点数目，m_j为该段落中的句子数目，N_{tf}为文章规定应选的考点数目，m为文章中的句子数目。

[0085] 3)计算文章中每个句子与其他句子的句义近似度；

[0086] 由于文章总是围绕核心信息展开，文章中各语义群间往往会存在语义相近的句子。为保证考点之间语义的相互独立性，本发明在选取考点时计算句义近似度，剔除与已选考点的近似度超过预设阈值的候选考点。由于句子单词数目较少，两个近义句子可能不存在重合单词，无法以单词重合度衡量句子近似度。本发明基于语义向量计算句义近似度。首先对所有句子，取句子所含单词的语义向量的几何中心作为句子的语义向量，具体公式如下：

[0087] $$\mathbf{s}=\frac{\mathbf{v}_1+\mathbf{v}_2+\ldots+\mathbf{v}_n}{n}$$

[0088] 其中：s是句子的语义向量，v_i(i=1,2,…,n)表示句子第i个单词的语义向量。

[0089] 然后根据以下公式计算文章中每个句子与其他句子的句义近似度：

[0090] $$sim=\frac{\mathbf{s}_1^{\mathrm{T}}\mathbf{s}_2}{\|\mathbf{s}_1\|\|\mathbf{s}_2\|}$$

[0091] 其中：s_1和s_2分别表示文章中任意两个句子的语义向量，sim为这两个句子的句义近似度。

[0092] 4)设置文章中所有句子两两组合的句义近似度的中值为考点选取的近似度阈值；对句子按照词频密度从高到低排序；

[0093] 5)根据词频密度序号依次判断：首先，对于待判断的句子，若其所在段落已选的考点数目已达到应选的考点数目，则不将其选为考点；若其所在段落已选的考点数目未达到应选的考点数目且其与文章中所有已被选为考点的句子的句义近似度均低于预设的近似度阈值，则将其选为考点并存储记录，否则不将其选为考点。

[0094] 6)如果已遍历完所有句子，但已选考点总数还未达到预设考点总数，则降低句义近似度阈值，重新执行步骤5)。

[0095] (3)问句生成

[0096] 1)对于被选为考点的句子，根据句子的词法信息和句子各成分的语义角色建立句子基于词汇功能语法理论的功能结构；

[0097] 本实施方式以文章CET4021的一个句子“The study shows that after school begins children are far more influenced by parents”为例建立该句子的功能结构。该句子中单词的词汇范畴和句子各成分的语义角色如表2所示(语义角色一栏中的V表示述语动词)。

[0098] 表2

[0099]

句子的单词	词汇范畴	语义角色		
		1	2	3
The		A0		
study	n.act			
shows	v.cognition	V		
that		A1		
after				AM-TEMP
school	n.group		A1	
begins	v.change		V	
children	n.person			A1
are	v.stative			
far	d.all			AM-MNR
more	d.all			
influenced	v.social			V
by				A0
parents	n.person			

[0100] 1.1)将处于短语结构中最外层的述语动词的原形show作为句子功能结构的谓词PRED的明细,根据短语结构中shows的词性标注VBZ(表示动词一般现在时),将句子功能结构的时态TENSE和体态ASPECT分别设置为现在时PRES和一般式CONT;将被show指派语义角色A0的句子成分the study作为句子功能结构中主语SUBJ的谓词PRED的明细,根据Michael Collins在题目为Head-driven statistical models for natural language parsing (Computational Linguistics,2003,29(4):589-637)的文献中提出的中心语判定规则确定主语the study的中心语为study,根据短语结构中study的词性标注NN(表示名词单数),将主语SUBJ的数NUM设置为单数SG;将被shows指派语义角色A1的句子成分that after school begins children are far more influenced by parents作为句子功能结构的宾语OBJ的谓词PRED的明细。

[0101] 1.2)将处于短语结构次外层的述语动词的原形influence作为其直联子功能结构的谓词PRED的明细。所述直联子功能结构为该句子功能结构的宾语OBJ的明细。根据短语结构中influenced的词性标注VBN和influenced前面的系动词are(表示动词一般现在时),将所述直联子功能结构的时态TENSE和体态ASPECT分别设置为现在时PRES和一般式CONT;将被influence指派语义角色A0的句子成分parents作为所述直联子功能结构中主语SUBJ的谓词PRED的明细。单词的中心语必然是该单词,根据短语结构中parents的词性标注NNS(表示名词复数),将主语SUBJ的数NUM设置为复数PL;将被influence指派语义角色A1的句子成分children作为所述直联子功能结构中宾语OBJ的谓词PRED的明细。将被influenced指派语义角色AM-TEMP的句子成分after school begins和指派语义角色AM-MNR的句子成分far more分别作为所述直联子功能结构中附加语ADJ_{TEMP}和ADJ_{MNR}的谓词PRED的明细。

[0102] 1.3)将处于短语结构最内层的述语动词的原形begin作为其直联子功能结构的谓词PRED的明细。所述直联子功能结构为该句子述语动词influenced的直联子功能结构的附加语ADJ_{TEMP}的明细。根据短语结构中begins的词性标注VBZ(表示动词一般现在时),将所述

直联子功能结构的时态TENSE和体态ASPECT分别设置为现在时PRES和一般式CONT；将被begins指派语义角色A0的句子成分school作为句子功能结构中主语SUBJ的谓词PRED的明细。单词的中心语必然是该单词，根据短语结构中school的词性标注NN(表示名词单数)，将主语SUBJ的数NUM设置为单数SG。

[0103] 2)以句子功能结构中的宾语OBJ为例，选择该OBJ为提问对象。

[0104] 3)确定提问对象的中心语为that，根据中心语的词汇范畴确定疑问词。疑问词判断条件如表3所示：

[0105] 表3

[0106]

语义角色	中心语的词汇范畴	疑问词	例子
A#	\.person	who	the little girl
		what	the device
TMP	\.duration	how long	for five years
		when	on Friday
CAU		why	because…
PNC		why	to earn money
LOC	\.location| address| factory| geography| organization	where	near the park
		how	in this way
MNR		how	carefully

[0107] 本实施例中，由于提问对象的语义角色是A1，且该提问对象的中心语的词汇范畴未定义，判定疑问词为what。

[0108] 4)用疑问词代替提问对象，根据短语结构和功能结构对被选为考点的句子做主谓倒装和时数一致性调整，生成以该疑问词引导的特殊疑问句。

[0109] 本实施方式用what代替句子成分that after school begins children are far more influenced by parents，形成句子Study after study shows what。将疑问词提前，根据句子功能结构的时态TENSE、体态ASPECT和主语SUBJ的数NUM，在疑问词后添加适用于一般现在时和主语为单数情况下的助动词does，并将述语动词shows用其原形代替，形成问句What does the study show?

[0110] (4)同义改写

[0111] WordNet中可将单词的多个义项根据它们在词典参考语料库中所出现的次数从高到低排序，获得义项的义频序号。本实施方式设置高频义项的判断条件为单词的词义在WordNet中的义频序号小于或等于2。同义改写步骤如下：

[0112] 1)对文章中的实词或固定短语进行语义消歧，以确定实词或固定短语在特殊疑问句中的语义；

[0113] 本实施方式采用Satanjeev Banerjee和Ted Pedersen在题目为Extended gloss overlaps as a measure of semantic relatedness(International Joint Conferences on Artificial Intelligence，2003，805-810)的文献中提出的拓展型Lesk算法。该算法不需要语料库训练，在Senseval-2数据上的细粒度语义消歧准确率为34.6%，具体方法为：对

于一个单词或固定短语w，以WordNet为词典，获取其n个义项的定义；对于w第i(i=1,2,…,n)个义项，计算w上下文单词的语义向量的几何中心量context(w)分别与该义项定义所含单词的语义向量的几何中心量S_i以及该义项的上义、下义和近义词义的定义所含单词的语义向量的几何中心量S_i'的句义近似度，然后取所有近似度之和作为该语义的分值，即：

[0114] $$\mathrm{score}(S_i)=\sum_{S'\equiv S \text{ or } S'\xrightarrow{ret}S} sim\left(context(w),\ gloss(S_i')\right)$$

[0115] 取n个义项中分值最高者作为该单词或固定短语在当前句子中的词义。

[0116] 2)对于特殊疑问句中的任一实词或固定短语，判断其在当前句子中的词义在WordNet中是否为该实词或固定短语的高频义项，若是进入步骤3)，若否，则不对该实词或固定短语做同义改写；

[0117] 3)根据语义利用WordNet获取该实词或固定短语的同义词集合，依次对集合中的同义词进行判断：对于集合中的任一同义词，判断先前所述的实词或固定短语的语义是否也是WordNet中该同义词的高频义项，若是进入步骤4.4，若否，则判断集合中的下一个同义词；

[0118] 4)判断该同义词是否超出阅读理解测试所指定的词汇范围，若否，将该同义词替换该实词或固定短语，若是，则判断集合中的下一个同义词；

[0119] 5)根据步骤2)～4)遍历特殊疑问句中的所有实词或固定短语；

[0120] 6)对于特殊疑问句中的任一代词，判断该代词及其所指的句子成分是否也在该特殊疑问句中，若是，则不对该代词做同义改写，若否，则进一步判断该特殊疑问句中是否存在另一个代词，其所指的句子成分与该代词的所指相同且其在特殊疑问句中处于该代词前面，若是，则不对该代词做同义改写，若否，则用该代词所指的句子成分替换该代词。

[0121] 以下对本实施例的输出结果进行分阶段评估。阶段1)的评估员为具有英语阅读理解测试命题经验的大学英语教师，其余阶段的评估员为通过了英语专业八级考试，具有较高英语水平的英语专业研究生。评估员在评估前均接受了关于评分培训。本实施例算法在表格中以ASIG(Automatic Short-answer Item Generation)表示。

[0122] 1)考点选取阶段：为了验证考点选取算法的有效性，设置另一种考点选取算法作为基线算法，具体算法如下：文章句子按原文顺序排序，选出第一句以及序号为句子总数四分位数的三个句子作为考点。

[0123] 由一名评估员评估ASIG算法和基线算法所选考点的质量。对于每一个考点，首先根据核心性、全面性和独立性分别评分，分值1表示符合，分值0表示不符合；其次，对文章原题所涉及的考点句子进行定位，然后对不同算法所选的每一个考点，判断是否存在原题考点与其重合，分值1表示存在，分值0表示不存在。不同算法对每篇文章所选考点的得分之和如表4所示。由该表可知，本发明算法在各项指标上的得分均远高于基线算法。

[0124] 表4

[0125]

文章编号（CET4-考试年份末尾2位数-序号）	核心性		全面性		独立性		与原题考点重合度	
	基线	ASIG	基线	ASIG	基线	ASIG	基线	ASIG
CET4021	3	4	2	4	0	4	1	3
CET4031	2	4	3	4	2	4	3	3
CET4032	3	4	2	4	4	4	1	2
CET4951	1	3	4	4	4	4	3	3
CET4971	2	3	3	4	2	4	2	3
CET4991	3	4	3	3	4	2	3	2
CET4992	1	4	3	4	4	4	1	4
总数	15	26	20	27	20	26	14	20

[0126] 2)同义改写阶段:为增加样本数量以提高评估的可靠性,取文章所有句子作为同义改写阶段的输入,统计同义词替换次数,计算单位句子的同义词替换次数;由一名评估员对所有同义词替换的正确与否进行判断,统计正确率。如表5所示,同义词替换的平均正确率为81.3%,远高于拓展型Lesk算法34.6%的语义消歧精度,证明通过限定同义词替换的义项范围和词汇范围能有效克服语义消歧准确度低的难题。虽然同义词替换次数因此而减少,平均每个句子进行了2.5次同义词替换。

[0127] 表5

[0128]

文章	替换次数	正确次数	正确率(%)	总句数	覆盖度(次/句)
CET4021	30	26	86.7	12	2.5
CET4031	51	45	88.2	18	2.8
CET4032	41	32	78.0	21	2.0
CET4951	34	27	79.4	19	1.8

[0129]

CET4971	41	28	68.3	16	2.6
CET4991	41	35	85.4	12	3.4
CET4992	40	33	82.5	15	2.7
合计	278	226	81.3	113	2.5

[0130] 3)问句生成和同义改写阶段:为增加样本数量以提高评估的可靠性,评估对象取对7篇文章所有句子所产生的疑问句(共469个句子)。由四名评估员以缺陷归类方式评估简答题问句,其中三名为主要评估者,一名为仲裁者。对于不一致的分类,首先以评估相同的两名评估者为准;如果三名评估者的评估互不相同,则由第四名仲裁者评估决定。该评估标准与Michael Heilman和Smith Noah在题目为Good question!Statistical ranking for question generation(Human language technologies:The2010annual conference of the North American chapter of the association for computational linguistics,2010,609-617)的文献中对其基于短语结构的问句生成方法所生成的疑问句的评估标准一致,本实施方式取该文献的实验数据作为评估基线,结果如表6所示。三名评估者的一致性

为Fleiss=4.7(基线的评估一致性为4.2),属于中等一致。由于评估等级达9项之多,该一致性可以接受。从表6可见,与单纯由陈述句生成问句的基线算法相比,本发明的方法虽然包含了可能引入错误的同义改写模块,所生成的问句在语法错误、不符合逻辑、语义模糊和答案缺失等方面的缺陷率较低,有效问句比例较高。本实施例中问句缺陷的主要来源是自然语言处理模块中自动标注器的错误。在自动标注器的精度得到提高后,根据本发明方法构造的实施系统能够产生比例更高的有效问句。由于本发明通过转换原文陈述句为疑问句得到简答题,除极少数题目由于原句与其他句子存在共指或归纳关系而属于推理型或归纳型简答题,绝大部分问句属于事实型简答题。

[0131] 表6

[0132]

问句缺陷	例子	比例(%)	
		基线	ASIG
语法错误	(1)What do many adult females select whether?	14.0	5.0
不符合逻辑	(2)How do mental attitudes instruct English?	20.6	4.5
语义模糊	(3)What is the extent of their safety value?	19.6	10.5
疑问词错误	(4)What faces a difficult decision?	0.9	5.1
语序/标点错误	(5)What did cities as time passed begin to do?	1.4	1.7
答案明显	(6)Who educated their kids? (Parents)	4.9	4.5

[0133]

答案缺失	(7)How did hotel rooms (cities) handle to make a net income?	8.9	1.5
其他错误	(8)Who saw how much I could have gained at dwelling (home)?	1.2	9.0
无		27.2	58.2

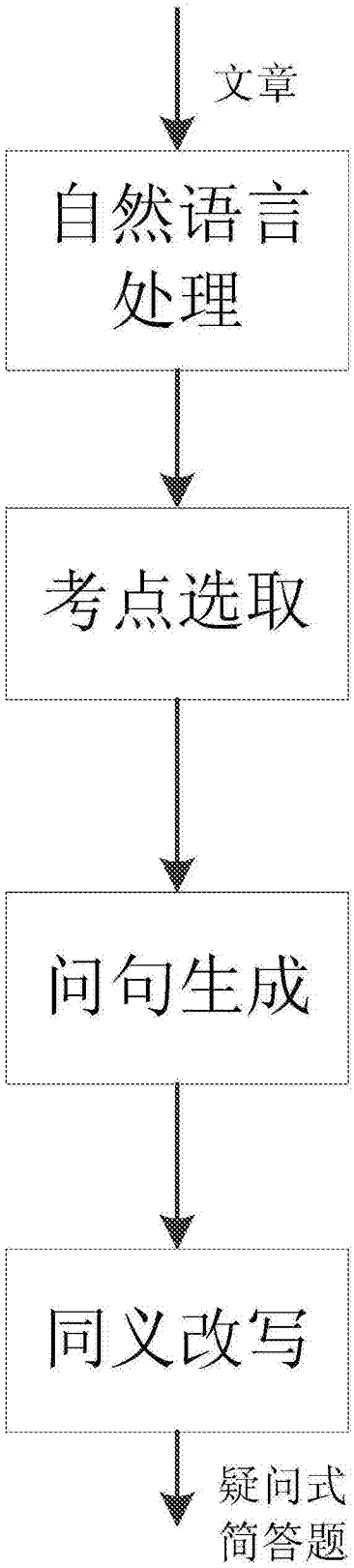


图1