文章编号：1003-0077(2018)03-0135-08

基于词语关联的散文阅读理解问题答案获取方法

乔 霈[1]，王素格[1,2]，陈 鑫[1]，谭红叶[1]，陈 千[1]，王元龙[1]

（1. 山西大学 计算机与信息技术学院，山西 太原 030006；
2. 山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006）

摘 要：高考语文阅读理解问答题中的提问方式复杂多样，使用的词语语义抽象，而相关阅读材料的内容表达丰富和含蓄，造成问题中的词语与阅读材料中词语存在一定的语义鸿沟。为了解决这一问题，该文对词语关联进行相关研究。首先利用 LDA 主题聚类方法，将同一主题类的词语进行聚类，根据各类词语的词性、词频特征，筛选与主题相关联的词语，再利用 Word2Vec 的语义相似度计算，将每一个主题关联的词语扩展，获得与主题词语义关联的词语。最后，将所提出的方法应用于近 12 年北京高考题和模拟题的散文抽取类问答题解答中，实验结果表明该方法优于传统的词语扩展方法。

关键词：阅读理解；问答题；LDA 聚类；词语关联

中图分类号：TP391　　**文献标识码**：A

Word Association Based Answer Acquisition for Reading Comprehension Questions from Prose

QIAO Pei[1], WANG Suge[1,2], CHEN Xin[1], TAN Hongye[1], CHEN Qian[1], WANG Yuanlong[1]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: Substantial semantic gap exists between the questions words and the article words in the reading comprehension test for Chinese of the college entrance examination, which may derive from the complexity and diversity of questions, abstract semantic meaning of words, and the rich and implicit semantics of articles. To address this issue, this paper investigates the words association. Specifically, all the words in the corpus are clustered into topics through LDA, which is then filtered by the part-of-speech and frequency, and augmented by the lexeme-related words according to the similarity of word embedding. Experiments on prose reading comprehension datasets of the college entrance examination indicate that our method performances better than traditional methods of words extension.

Key words: reading comprehension; essay question; LDA clustering; words association

0 引言

随着国内外越来越多的机构投入到问答系统的研究中，使得自动问答技术取得了很大的进展。问答系统，即利用自然语言处理技术理解用户所提出的问题，再返回给用户正确的答案[1]。阅读理解属于问答任务中的一个重要分支，又与传统的问答存在区别，它是通过机器理解一篇文章，再根据文中信息对所提的问题做出回答，主要侧重于问题与阅读材料的语义相关性。面向高考散文阅读理解问答题，按照问题的提问方式我们将其归纳为特点（特色）类、感受类、认识（态度）类、原因类、列举类、其他类共六类。为了解答这些问题，首先需要理解题干的相关信息，然后从阅读材料中获取与题干中信息相关的词语或短语，最后将阅读材料中与词语或短

收稿日期：2017-08-09　**定稿日期**：2017-09-25

基金项目：国家“863”高技术项目(2015AA015407)；国家自然科学基金(61573231)

语相关的句子作为答案句。表1所示为阅读理解中问答题的问题、答案示例。

表1 阅读理解中问答题的问题、参考答案示例

【问题】作者从鲁迅的故乡环境中看出了哪些特色？（选自2013年北京市语文高考题）
【参考答案】 颜色：黑瓦白墙，白石板、黑柱子，搭配着绿水，非常素净； 空间：空间庞大，人有足够的地方可以移动； 物体：厚实可靠，石板下有流水，质朴中带有温情。

表1中问题的关键词为“特色”，描述对象为鲁迅的故乡环境，理解“特色”一词的抽象语义，需要从阅读材料中寻找鲁迅的故乡环境与“特色”相关联的词语，构成答案句。

针对散文阅读理解类问题中的词语较为抽象，在语义上难以与阅读材料中的信息联系，需要将问题中具有抽象含义的词语扩展为与其关联的具体词语，再进一步与阅读材料中的句子进行联系。

本文利用LDA方法将问题库中的问题词语与阅读材料涉及的内容进行主题聚类，然后按照词性、词频特征筛选出每个主题下相关的词语作为问题词语的主题关联词，再利用Word2Vec训练散文语料，将得到的主题关联词语进一步扩展为语义关联词语。通过获取语义关联词语，使问题关键词语与阅读材料中句子之间建立联系，从而丰富问题关键词语，提高问题答案句的抽取性能。

1 相关工作

自从1999年TREC(text retrieval conference)会议[2]开设QA Track以来，自动问答及阅读理解的研究就备受关注。早期问答系统的研究主要有以下三个方面：(1)基于统计方法是从文本集中抽取答案返回给用户。例如，IBM开发的基于统计的问答系统主要应用统计翻译、词汇模式等抽取方法。(2)基于知识库的方法是从知识库中抽取问题的答案。例如，芝加哥大学开发的FAQFinder[3]，用于解决一些地理、历史、文化等方面的简单问题。(3)基于语义的方法是通过计算词语间的语义相似度获得答案句。例如，台湾Sheng-YuanYang开发的FAQ-master[4]。目前阅读理解方面的研究大多针对简单文本和简单问题，例如，微软建立的一套面向儿童的开放域阅读理解数据集MCTest[5]，Smith等[6]针对此数据集提出了在文本上设置滑动窗口来与问题答案对中的词汇匹配打分的方法，引用一种基于RTE的方法将问题与答案按照启发式规则进行拼接，然后计算上述拼接结果与原文信息之间的相关性。Facebook的bAbI项目仿真生成了20个任务用于测试文本理解和推理[7]，Sukhbaatar等[8]提出端到端的记忆网络模型，用于解答上述20个任务中的短文本问题。由于问题和文本是自动生成，相应的数据简单，使得结果准确率高，但是该实验侧重信息推理，未考虑文本的语义信息，因此，难以应用到中文阅读理解任务中。王智强等[9]提出一种基于篇章框架语义分析的答案抽取方法，并将其应用于解答中文阅读理解问题。该方法主要依赖框架结构，而散文本身用词广泛，隐含语义丰富，且问题中的词语较抽象，框架关系中目前还未覆盖散文领域的抽象词语，因此，还难以利用框架关系建立问题中词语与文章之间的关系。

对于机器阅读理解问题，现有研究者的主要工作集中于问题分析、答案抽取及生成[10]。然而，由于问题中的关键词与答案句中的词语在表达方式上存在差异，导致问题中词语未能与文章中的词语相联系，这将影响答案句抽取的准确性，因此，在散文问答题中有必要进行词语关联方法研究。

词语关联，即寻找词语的潜在语义，解决词语的一词多义、多词同义现象，用于提高检索的准确率。目前，问答系统中采用的词语关联方法主要包括基于统计的方法和基于语义词典或特定扩展词表的方法。

基于统计的词语关联方法通常利用词语之间的共现概率或互信息等统计信息来选取关联词，该方法并没有深入分析原查询词与候选关联词间的语义关系。例如，Jones[11]提出词的聚类算法，根据词与词之间在语料库中的共现程度实现词聚类，并将查询词所在簇中的其他词作为关联词语。丁立恺[12]提出词关联度的概念，通过对文本语料库中词语出现的频率，以及任意两个词语共同出现的频率进行统计，获得各个词语之间的关联度。

基于语义词典的方法，需要借助词典中的词建立与查询词之间的语义关联。张华平等[13]通过使用WordNet的语义体系对词语进行语义关联性的扩展。Rothe等[14]结合深度学习方法以WordNet作为语义资源提出自动扩展的方法，构造了一个使用词嵌入扩展同义词集和语义嵌入的系统。史俊冰等[15]建立了同义词词典，并在此基础上实现了词语扩展。万静等[16]通过构建领域知识词典的方法扩

展用户输入的关键词。以上基于词典的扩展方法依赖于完备的语义体系，而目前并没有散文领域的相关体系。另外，基于语义词典的方法不依赖语料集，难以联系阅读理解的文本内容的特性。陈建超等人[17]通过建立包含上下文信息的同义词集解决文本中的一词多义和多词同义问题。他将词语的上下文信息视为特征词，根据特征词之间存在的关联性特点建立了一个评分机制，提取分数最高的特征词集对应的词汇作为一个同义词集，该方法比直接提取近义词或提取上下文相关词的准确率有所提高，但是考虑阅读材料中大多采用含蓄、隐式的词语来表达作者的情感，因此，难以直接将该方法应用于阅读理解当中。

上述研究主要侧重将词语扩展为与其表层语义相近的词，从而忽略了词语在特定语境和不同主题下的潜在语义信息。

LDA(latent dirichlet allocation)是 Blei 等[18] 2003 年提出的一种被广泛使用的主题模型，能够从海量语料库中获取核心语义或特征并对主题进行建模，它是一个离散数据集生成概率模型的过程[19]，其工作原理是将语料库中的每一个文档与一组潜在主题的概率分布进行对应，而每一个潜在主题同时与文档中词语的概率分布相对应。该模型基于三点假设[20]：(1)词袋模型，LDA 认定每篇文档是由一组词汇构成，且词汇之间无先后顺序关系，词语集合 $W=\{w_1,w_2,\cdots,w_n\}$；(2)训练集中的文档顺序也是随意的，无指定顺序，因此，每篇文档可以表示成一个词频向量关系集合 $d_i=<t_{i1},t_{i2},\cdots,t_{ij}>$，其中 t_{ij} 表示单词 j 在文档 i 中出现的次数；(3)它是一种基于参数的贝叶斯模型，在训练模型前需要先设定主题数量 K。因此，LDA 被广泛地应用于文本的特征选择、主题分类、文本聚类等[21]。本文利用 LDA 将大量未知的文本自动划分为适当的类簇，使同一类别中的文本尽可能含有相似的主题，而不同类别的文本间主题差异较大，以此方法从无序的文本信息中发现文本的分布特点。

2　面向问题解答的词语关联方法

考虑到问题与答案集、问题与阅读材料中词语间具有主题相关性和语义相关性，我们利用 LDA 主题聚类方法，确定各类别问题词语的主题，再利用词语重要度选择各类别相应主题下重要度高的词语作为该问题类别的主题关联词语。在此基础上，利用 Word2Vec 对主题关联词语与材料中词语进行向量表示，用于度量词语间的语义相关性。最后利用上述两种方法，扩展问题的抽象词语，建立问题与散文材料中的词语联系。

2.1　基于 LDA 的问题主题词语扩展

为了获取主题相关的词语，以散文阅读理解为背景，从所有的阅读材料—问题—答案集中整理出抽取类问答题(抽取类，即答案句是从文中摘取的句子)，以这些问题-答案集为数据，通过 LDA 聚类方法将数据集下的词语聚集在不同的主题之下，使各类别问题词语对应各自的主题。例如，文本“作者故乡植物的生命具有哪些特点?”其示意图如图 1 所示。

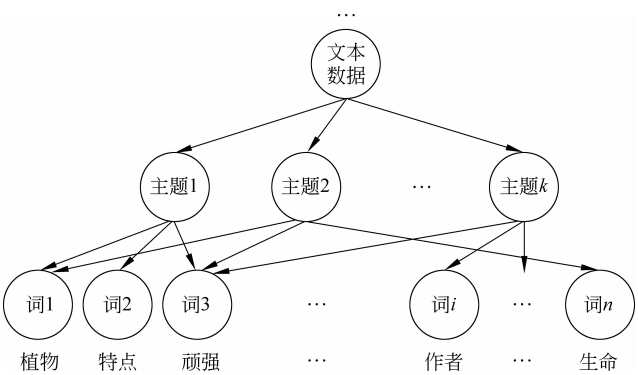

图 1　LDA 聚类主题—词汇分布举例

通过 LDA 主题聚类，可以计算各数据(一条数据指的是一个问题—答案对)在每个主题下的概率，计算方法如式(1)、式(2)所示。

$$N(TW_j,W_i)=W_i\cap TW_j \tag{1}$$

$$R(i,k)=\underset{j}{\operatorname{argmax}}\frac{|N(W_i,TW_j)|}{|W_i|} \tag{2}$$

其中，W_i 表示第 i 条数据的词集，TW_j 表示主题 j 中的词集，$N(W_i,\ TW_j)$ 为第 i 条数据与主题 j 中共同出现的词语，$R(i,k)$ 表示第 i 条数据对应的主题为 k。

针对上文引言中提到的阅读理解中六类问题所关联的词语，可以确定各数据所属的类别，利用各类别数据在不同主题下的比例可以获取类别为 ty_n 对应的最优主题 $k(\mathrm{ty}_n)$，如式(3)所示。

$$k(\mathrm{ty}_n)=\underset{k}{\operatorname{argmax}}\frac{m(k,\mathrm{ty}_n)}{n(\mathrm{ty}_n)} \tag{3}$$

其中，$n(\mathrm{ty}_n)$ 是 ty_n 类的数据总数，$m(k,\mathrm{ty}_n)$ 表示 ty_n 类中的主题为 k 的数据个数。

根据式(3)，可将各类问题与主题对应，然而通

过对大量数据考察，发现各主题下的部分词语集与该类问题中词语关联性不强，例如，各类问题的描述性词语一般多为名词和形容词，而动词“分析”“具有”“选择”等为不具有特定意义的词语。因此，需要进一步对六类问题中的词语进行筛选。各类问题的形容词和名词部分描述如表2所示。

表2 各类问题的部分描述性词语

问题类别	描述性词语
特点类 ty_1	名词：颜色、环境、生命、生活…… 形容词：顽强、美丽、古老、温暖、静谧、坦然、真实……
感受类 ty_2	名词：自然、时间、心灵、童年…… 形容词：真实、震撼、专注、无奈、遗憾、乐趣、寂寞……
态度类 ty_3	名词：文化、科学、生活、人生…… 形容词：震惊、欣喜、痛苦、得意、反感、郁闷、惆怅……
列举类 ty_4	名词：方面、因素、文化、精神、思想、成就…… 形容词：严谨、勤奋、扎实、深刻、崇高……
原因类 ty_5	名词：原因、问题、方式、空间、环境…… 形容词：广泛、强烈、不易、独到……
其他类 ty_6	名词：关系、情怀、意蕴、影响、意义、内涵…… 形容词：年轻、广博、巨大、觉悟……

为了准确地获得与解题相关的关联词语，分别统计六类问题中的词语对应的主题下的名词和形容词出现的次数。按照高频词数均值法确定主题中抽取词语的数量 t_n，计算方法如式(4)所示。

$$t_n = \frac{\sum_{n=1}^{m} |f_{ty_n}|}{m} \tag{4}$$

其中，m 表示问题类别总数，f_{ty_n} 表示 ty_n 类问题对应的主题中频次高于 l 的词语集。

由于每类词语中名词和形容词的重要度不同，进一步设置参数 α 和 $1-\alpha$ 分别代表名词和形容词在每类问题词中所占的比重，以此获得每个类别中名词和形容词保留的个数，计算方法如式(5)所示。

$$\begin{cases} N'(ty_n, w_{n.}) = \alpha_{ty_n} \cdot N \\ N'(ty_n, w_{adj.}) = (1-\alpha_{ty_n}) \cdot N \end{cases} \tag{5}$$

其中，$N'(ty_n, w_{n.})$ 为 ty_n 类下名词的数量，$N'(ty_n, w_{adj.})$ 为 ty_n 类下形容词的数量。

2.2 基于 Word2Vec 的问题语义词语扩展

由于高考阅读材料的有限性，仅仅利用2.1节中方法获得每类词语的主题关联词语不能满足问题解答的要求，需要进一步获取与散文领域中词语语义相关联的词语。Word2Vec 是2013年由 Google 公司开发的将词表示为向量形式的工具[22]，这些向量中含有潜在丰富的语义信息。本文将散文阅读材料与主题相关的词语通过 Word2Vec 训练，使它们转化为特定维度的向量表示，然后再计算词语间相关性，该方法记为 TWE。

假设 PC 为散文材料库，T 为主题词语集合。

词语相似度计算过程：利用 Word2Vec，将词语 $p \in PC$ 和主题关联词语 $q \in T$ 分别表示成向量 $\boldsymbol{w}(p)$，$\boldsymbol{w}(q)$，PC 中所有词语的向量集合记为 PC'。通过计算 $\boldsymbol{w}(p)$ 与 $\boldsymbol{w}(q)$ 之间的余弦夹角，可获得 $\boldsymbol{w}(p)$ 与 $\boldsymbol{w}(q)$ 的相似度矩阵 $\{\cos(\boldsymbol{w}(p), \boldsymbol{w}(q))\}_{|T||PC|}$。

词语关联度排序函数：为了获取 PC 中与 q 语义相似度高的词语，我们定义 $\boldsymbol{w}(p)$ 与 $\boldsymbol{w}(q)$ 余弦值的排序函数 Rank，具体如式(6)所示。

$$Top\text{-}h(\boldsymbol{w}(q)) = \underset{w(p) \in PC'}{\text{argRank}}\{\cos(\boldsymbol{w}(p), \boldsymbol{w}(q))\} \tag{6}$$

这里 $Top\text{-}h(\boldsymbol{w}(q))$ 为余弦值排序在前 h 个对应的词语序列。

3 散文问答题答案抽取方法

根据高考语文相关专家分析，通常阅读理解问答题的得分是按照给出的答案要点进行评判。因此，针对散文问答题的答案抽取任务，需要计算词语间的相关性。通常采用词语的词形匹配和语义相似计算，而词形匹配一般使用词语匹配的句子相似度计算方法[23]，语义相似计算采用 HowNet 的句子相似度计算[24]方法。

假设问句 q 中的关键词集为 $W(q) = \{key_i(q)\}_{i=1}^{n}$，根据问句确定的问题类别，再利用第2.1～2.2节的方法获得关联词集为 $CO(q) = W(q) \cup Top\text{-}h(\boldsymbol{w}(q))$。假设问题句 q 的第 j 个关联词为 $co_j(q)$，且 $co_j(q) \in CO(q)$。

设阅读材料中句子集为 $S = \{s_t\}_{t=1}^{l}$，任意句子 $s \in S$ 对应的关键词集为 $KW(s) = \{key_i(s)\}_{i=1}^{n}$，这里的 $key_i(s)$ 为句子 s 的第 i 个关键词。

(1) 词语的词形匹配计算方法 sim_1

问题句 q 与阅读材料中句子 s 的相似度算法如式(7)所示。

$$sim_1(q,s) = \frac{2|CO(q) \cap KW(s)|}{|CO(q)| + |KW(s)|} \tag{7}$$

(2) 词语的语义相似计算方法 sim_2

对于问句 q 与阅读材料中句子 s 的相似度计算，如式(8)所示。

$$sim_2(q,s)=\frac{1}{2}\left(\frac{a(s)}{|KW(s)|}+\frac{b(s)}{|CO(q)|}\right) \tag{8}$$

这里 $a(s)=\sum_{i=1}^{|KW(s)|}\max_{1\leqslant j\leqslant|CO(q)|}\{si(key_i(s),co_j(q))\}$，$b(s)=\sum_{j=1}^{|CO(q)|}\max_{1\leqslant i\leqslant|KW(s)|}\{si(key_i(s),co_j(q))\}$，$si(key_i(s),co_j(q))$ 为利用 HowNet 的词语相似度计算方法[24] 得到的相似度。

(3) 词语的词形匹配与语义相似混合计算方法 sim

对于词语的词形匹配，仅利用词语的表层信息，而词语的语义相似计算方法考虑词语的深层语义信息，因此，本文将两者有机结合。利用式(7)和式(8)，获得问句 q 与阅读材料中句子 s 的词语的词形匹配与语义相似混合计算方法 $sim=sim_1(q,s)+sim_2(q,s)$，选择问句 q 与阅读材料中句子相似度高的 N 个句子作为答案句。

4　实验结果与分析

4.1　实验数据及评价指标

本文的实验数据分为训练数据和测试数据。

训练数据：主题聚类所用的数据集是从人工整理的各省高考题(不包含北京卷)共 1 647 篇文章，包含 6 117 个问题—答案集中的约 600 个抽取类试题的问题—答案集；内容关联词语扩展所用的数据集是从网络爬取的近七万篇文学作品的阅读理解，规模大约 320 MB。

测试数据：选择北京市近 12 年的高考题和网上收集的 1 000 套高考模拟题作为方法验证，其中抽取类问答题有 400 个。

评价指标：(1)本文采用信息熵来度量聚类结果对各类问题的影响；(2)根据题目所给的参考答案人工从材料中寻找对应的句子，并记为答案句集合 A，T 为使用本文方法得到的答案句子集合，按如下公式计算准确率(P)、召回率(R)和 F 值。其中，

$$\begin{aligned}P&=\frac{|A\cap T|}{|T|}\times100\%\\R&=\frac{|A\cap T|}{|A|}\times100\%\\F&=\frac{2PR}{P+R}\end{aligned} \tag{9}$$

4.2　参数设置

4.2.1　主题个数的选择

利用 2.1 节中介绍的方法进行聚类，实验分别选择主题数 $k=5,7,10$，用于比较聚类结果中六类问题词语在不同主题下的分布比例，实验结果如表 3～表 5 所示。

根据表 3～表 5，获得六类问题在不同主题下聚类结果。首先，计算各类问题在各主题中的信息熵，然后加和取平均作为该主题数聚类下的整体信息熵值，最终得到主题数 $k=5,7,10$ 时信息熵值分别为 $H(k=5)=2.14$、$H(k=7)=2.35$，$H(k=10)=2.94$。熵值越小，说明聚簇结果越好。因此，选择主题数 $k=5$ 最佳。下面的实验主题数均为 $k=5$。

表 3　各类问题在主题数 $k=5$ 的聚类结果的分布比例/%

ty_n	Topic①	Topic②	Topic③	Topic④	Topic⑤
ty_1	18.49	40.34	8.40	21.01	11.76
ty_2	10.71	42.86	7.14	3.57	35.71
ty_3	6.45	35.48	16.13	9.68	32.26
ty_4	24.90	13.88	19.59	31.02	10.61
ty_5	24.29	22.86	10.71	15.00	27.14
ty_6	12.50	17.50	10.00	27.5	32.50

表 4　各类问题在主题数 $k=7$ 的聚类结果的分布比例/%

ty_n	Topic①	Topic②	Topic③	Topic④	Topic⑤	Topic⑥	Topic⑦
ty_1	16.90	16.20	15.49	7.75	2.82	20.42	20.42
ty_2	48.00	4.00	32.00	4.00	0	0	12.00
ty_3	40.74	7.41	33.33	3.70	0	3.70	11.11
ty_4	12.45	21.79	6.23	12.45	15.56	20.23	11.28
ty_5	14.91	14.29	21.12	19.25	8.70	14.91	6.83
ty_6	28.57	16.67	9.52	4.76	9.52	14.29	16.67

表 5　各类问题在主题数 $k=10$ 的聚类结果的分布比例/%

ty_n	Topic①	Topic②	Topic③	Topic④	Topic⑤	Topic⑥	Topic⑦	Topic⑧	Topic⑨	Topic⑩
ty_1	5.41	8.11	1.80	9.91	1.80	2.70	44.14	9.91	12.61	3.60
ty_2	10.71	17.86	3.57	7.14	0	3.57	3.57	25.00	25.00	3.57

续表

ty_n	Topic①	Topic②	Topic③	Topic④	Topic⑤	Topic⑥	Topic⑦	Topic⑧	Topic⑨	Topic⑩
ty_3	3.03	27.27	3.03	12.12	12.12	3.03	3.03	12.12	21.21	3.03
ty_4	12.26	6.51	10.34	19.16	8.43	11.11	4.21	6.90	6.13	14.94
ty_5	3.13	16.25	5.00	8.75	5.00	11.88	6.25	14.38	10.00	19.38
ty_6	4.17	10.42	12.5	22.92	8.33	2.08	10.42	12.5	8.33	8.33

4.2.2 词语筛选

考虑词语的覆盖度，本实验设置主题下高频词阈值 $l=6$，4，2 三组实验，利用 2.1 节中式(4)获得主题关联词语数 $t_n=13,24,60$，而实验中主题关联词语的数量 $t_n=24$ 时答题效果最好，因此，本实验将高频词的阈值设置为 $l=4$。

当 t_n 确定后，利用 2.1 节中式(5)确定每类问题中名词和形容词的比例，本实验取 $\alpha=0,0.1,0.2,\cdots,0.9,1$，共 11 组实验，针对每类问题的答题准确率，选择各类别主题下名词和形容词的最优个数，结果如图 2 所示。

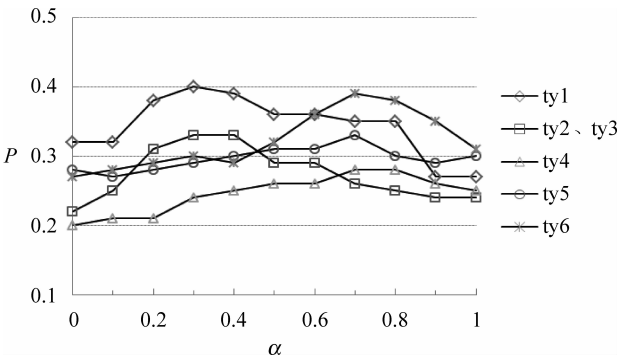

图 2 词语重要度参数 α 选择实验结果

从图 2 可以看出 ty_1、ty_2、ty_3 的问题在 $\alpha=0.3$ 时效果最好，ty_4、ty_5 和 ty_6 在 $\alpha=0.7$ 时效果最好。

4.2.3 语义关联词语 $Top\text{-}h(w(q))$ 中 h 的选择

散文语料中词向量的训练选取了 Word2Vec 的 Skip-gram 模型[25]，参数设定为默认值，即文本窗口为 5，向量维度为 300 维。训练后得到 80 000 多个词向量。

利用 2.2 节中介绍的方法扩展词语，词汇的数量 h 分别取 5，10，15，20，25，30，方法验证时答案句的个数取 $N=4,6,8$，通过测试，得到最好结果为 $N=6$。因此，抽取答案句子数 $N=6$ 计算准确率，结果如图 3 所示。

由图 3 可知，当最终确定扩展词汇数量 $h=5$ 时，实验结果较好。

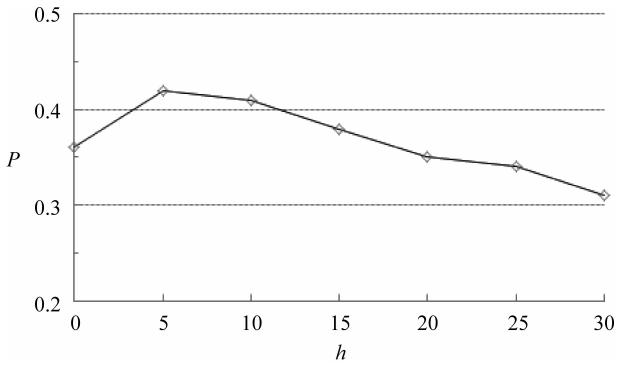

图 3 扩展词汇数量 h 的选取

4.3 答案句抽取实验结果与分析

为了验证本文扩展词语对答案句抽取的有效性，设置了三个 Baseline 方法进行对比。

(1) 直接抽取答案句(DE)：即问题中的关键词不进行词语扩展。

(2) 基于 Word2Vec 的词语关联抽取答案(WE)：将问题词集 $W(q)$ 直接利用 Word2Vec 余弦相似度得到扩展词集 $WV(q)$，不进行主题聚类。

(3) 基于同义词词林的词语关联抽取答案句(SE)：将问题词集 $W(q)$ 利用同义词词林扩展词集 $SW(q)$。

利用 DE、WE、SE 以及本文方法 TWE 获取关联词语，再分别使用第三节的三种答案抽取方法获得前六个答案句子数，得到不同词语关联方法和不同的答案抽取方法间的比较结果，如表 6 所示。

表 6 不同词语关联与不同的答案句抽取方法的 F 值结果比较/%

答案句抽取方法	词语关联方法			
	方法 1	方法 2	方法 3	本文方法
sim_1	27.16	29.69	30.08	31.46
sim_2	29.54	33.34	33.34	35.11
sim	31.85	35.55	40.02	50.45

由表 6 中结果可知：

① 词语的词形匹配方法 sim_1 抽取答案句的 F 值比词语语义相似方法 sim_2 抽取答案句的 F 值低，主要原因是由于散文问答题的特殊性，它更强调抽象词语的语义。

② 词语的词形匹配与语义相似混合计算方法 sim，得到的结果比单独的方法 sim_1(sim_2)抽取答案句的 F 值高，主要原因是同时考虑了词语的词形和语义。因此，本文答案句抽取方法选择 sim 方法。在此基础上，不同词语关联方法在抽取前六句答案句时的准确率、召回率和 F 值如表 7 所示。

表 7　不同词语关联方法抽取答案句的结果比较

方法	准确率/%	召回率/%	F 值/%
DE	25.54	42.28	31.85
WE	29.32	45.16	35.55
SE	32.24	52.74	40.02
TWE	41.95	63.25	50.45

由表 7 中实验结果可知：

(1) 本文方法 TWE 比 Baseline 方法在答案句抽取的三项指标结果均好。方法 DE 没有对问题中的关键词扩展，使问题与阅读材料中相关句子难以联系。方法 WE 和方法 SE 虽然对解题起到一定作用，但是问题中抽象词语扩展为其抽象的近义词，未从根本上解决抽象词与具体词之间的语义鸿沟，导致准确率不及方法 TWE。

(2) 与方法 DE、WE、SE 相比，本文方法 TWE 从主题角度扩展抽象词的关联词语，使得答题准确率和召回率有了显著提升，说明词语关联方法对散文抽取类问题的解答确实起到了作用。

(3) 为了验证方法 TWE 的显著性，从统计学角度分析，采用配对样本的 t 检验方法衡量数据的统计意义，当 p 小于 0.05 时，说明两组数据的平均值在小于 5%的概率上是相等的，在大于 95%的几率上不相等，两组实验存在显著性差异。将方法 DE、WE、SE 与 TWE 对比，分别获得的概率值为：

$$p(1)=0.012<0.05$$
$$p(2)=0.036<0.05$$
$$p(3)=0.043<0.05$$

由于三组数据的概率 p 均小于 0.05，因此，方法 TWE 的实验结果具有统计显著性。

4.4　高考题答题结果及分析

对于引言中表 1 的高考题，利用 TWE 方法，可以获得问题的解答结果，如表 8 所示。

表 8　本文方法解答问答题示例

【问题】作者从鲁迅的故乡环境中看出了哪些特色？（本例选自 2013 年北京市语文高考题）

【问题关键词】鲁迅；故乡；特色

【“特色”的关联词语】美丽、古老、快乐、宁静、忧伤、坦然、神秘、生机勃勃、狭小、挺拔、欢快、和谐、生命、自然、内容……

【答案句】

你看他用笔何等经济，总是短短几句话就勾画出一个实实在在的人生处境，而同时他又总是把这处境放在一片抒情的气氛之内；

空间是庞大的，人有足够的地方可以移动，物件也是厚实可靠的，像那件大大的厨房里的那口大大的腌菜缸，在朴质的生活里有温厚的人情……

由表 8 结果可知，利用本文方法 TWE 扩展问题词语，再抽取答案句，可以获得“空间……；物件……”两句正确答案。

如果采用方法 DE 解答本题，未能获得正确答案句。方法 WE 和方法 SE 均获得一句正确答案。因此，本文方法 TWE 在一定程度上提高了散文阅读理解的答题准确率。

5　总结

散文阅读理解问题中的关键词具有抽象含义，导致问题与答案句之间具有较大的语义鸿沟，为了解决该类问题，本文提出词语关联方法。首先基于 LDA 聚类的主题—词汇分布，确定各数据的主题，然后根据各类数据在主题下的分布比例为每类数据分配最优主题，对该主题下的词语重要度进行选择，得到各问题类别的主题关联词语；接着，利用 Word2Vec 相似度方法将主题关联词语扩展为语义关联词语，最后利用词形匹配和语义相似混合计算方法抽取答案句。方法 TWE 有效提高了散文问答题的答题准确率和召回率。另外，方法 TWE 不仅适用于高考阅读理解的问题解答，也可以应用于信息检索任务中。

由于散文阅读材料往往带有作者的情感信息，因此在未来工作中，将考虑情感词的重要性，结合句子中的情感信息进一步获得词语的关联词语。

备注　本文使用了哈尔滨工业大学计算与信息检索中心研发的 LTP 进行分词及词性标注；使用了知网提供的语义相似度计算方法。

参考文献

[1] 张宁，朱礼军. 中文问答系统问句分析研究综述[J]. 情报工程，2016，2(1)：32-42.

[2] Katz B. Annotating the World Wide Web using natural language[C]//Proceedings of Computer-Assisted Information Searching on Internet，1997：136-155.

[3] Hammond K，Burke R，Martin C，et al. FAQ Finder：A case-based approach to knowledge navigation[C]//Proceedings of Conference on Artificial Intelligence for Applications，1995：80-86.

[4] Yang S Y. An ontological multi-agent system for web FAQ query[C]//Proceedings of the 2007 International Conference on Machine Learning and Cybernetics，2007：2964-2969.

[5] Matthew R，Christopher J C Burges，Eric Renshaw. MCTest：A challenge dataset for the open-domain machine comprehension of Text[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing，2013：193-203.

[6] Smith E，Greco N，Bosnjak M，et al. A strong lexical matching method for the machine comprehension test[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing，2015：1693-1698.

[7] Weston J，Bordes A，Chopra S，et al. Towards ai-complete question answering：A set of prerequisite toy tasks[J]. arXiv preprint arXiv：1502.05698，2015.

[8] Sukhbaatar S，Weston J，Fergus R. End-to-end memory networks[C]//Proceedings of Advances in Neural Information Processing Systems，2015：2440-2448.

[9] 王智强，李茹，梁吉业，等. 基于汉语篇章框架语义分析的阅读理解问答研究[J]. 计算机学报，2016，39(4)：795-807.

[10] 吴友政，赵军，段湘煜，等. 问答式检索技术及评测研究综述[J]. 中文信息学报，2005，19(3)：1-13.

[11] Jones S. Automatic keyword classification for information retrieval[J]. The Library Quarterly：Information，Community，Policy，1971，25(4)：33-98.

[12] 丁立恺. 基于词关联度的信息检索系统[D]. 上海：复旦大学硕士学位论文，2010.

[13] 张华平. 语言浅层分析与句子集新信息检测研究[D]. 北京：中国科学院研究生院博士学位论文，2005.

[14] Rothe S，Schütze H. Autoextend：Extending word embeddings to embeddings for synsets and lexemes[J]. arXiv preprint arXiv：1507.01127，2015.

[15] 史俊冰. 问答系统中词义消歧与关键词扩展研究[D]. 太原：太原理工大学硕士学位论文，2011.

[16] Wan J，Wang W C，Jun-Kai Y I. Semantic extended search approach based on ontology in knowledge base[J]. Computer Engineering，2012，38(6)：19-24.

[17] 陈建超，郑启伦，李庆阳，等. 基于特征词关联性的同义词集挖掘算法[J]. 计算机应用研究，2009，26(7)：2517-2519.

[18] Blei D M，Ng A Y，Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research，2003(3)：993-1022.

[19] Blei D，Carin L，Dunson D. Probabilistic topic models[J]. IEEE Signal Processing Magazine，2010，27(6)：55-65.

[20] 魏强，金芝，许焱. 基于概率主题模型的物联网服务发现[J]. 软件学报，2014(8)：1640-1658.

[21] 董婧灵. 基于 LDA 模型的文本聚类研究[D]. 武汉：华中师范大学硕士学位论文，2012.

[22] 宁建飞，刘降珍. 融合 Word2Vec 与 TextRank 的关键词抽取研究[J]. 现代图书情报技术，2016(6)：20-27.

[23] 王荣波，池哲儒，常宝宝，等. 基于词串粒度及权值的汉语句子相似度衡量[J]. 计算机工程，2005，31(13)：142-144.

[24] 刘青磊，顾小丰. 基于《知网》的词语相似度算法研究[J]. 中文信息学报，2010，24(6)：31-37.

[25] Mikolov T，Chen K，Corrado G，et al. Efficient estimation of word representations in vector space[J]. arXiv：1301.3781.

乔霈(1992—)，硕士研究生，主要研究领域为自然语言处理。

E-mail：1336822954@qq.com

陈鑫(1992—)，博士研究生，主要研究领域为情感分析。

E-mail：1315614497@qq.com

王素格(1964—)，博士，教授，主要研究领域为自然语言处理。

E-mail：wsg@sxu.edu.cn