

文章编号: 1003-0077 (2016) 00-0000-00

## 基于框架语义的高考语文阅读理解答案句抽取

李国臣<sup>1,2</sup>, 刘姝林<sup>1</sup>, 杨陟卓<sup>1</sup>, 李茹<sup>1,3</sup>, 张虎<sup>1</sup>, 钱揖丽<sup>1</sup>

(1.山西大学 计算机与信息技术学院, 山西 太原 030006;

2.太原工业学院 计算机工程系, 山西 太原 030008;

3.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相对普通阅读理解问答难度更大, 问句抽象表述的理解需要更深层的语言分析技术, 答案候选句抽取更注重与问句的关联分析, 答案候选句排序更注重答案句之间的语义相关性。为此, 该文提出借助框架语义匹配和框架语义关系抽取答案候选句, 在排序时引入流形排序模型, 通过答案句之间的框架语义相关度将排序分数进行传播, 最终选取分数较高的 Top-4 作为答案句。在北京近十二年高考语文阅读理解问答上的准确率与召回率分别达到了 53.65%与 79.06%。

**关键词:** 高考语文; 阅读理解; 框架语义; 答案句抽取; 流形排序

**中图分类号:** TP391

**文献标识码:** A

### Answer Sentences Extraction for Chinese Reading Comprehension of College Entrance Examination Based on Frame Semantic

LI Guochen<sup>1,2</sup>, LIU Shulin<sup>1</sup>, YANG Zhizhuo<sup>1</sup>, LI Ru<sup>1,3</sup>, ZHANG Hu<sup>1</sup>, QIAN Yili<sup>1</sup>

(1.School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006,China;

2.Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan,Shanxi 030008,China;

3.Key Laboratory of Computation Intelligence & Chinese Information Processing,Shanxi University,

Taiyuan,Shanxi 030006,China)

**Abstract:** Reading comprehension QA for College Entrance Examination on Chinese is much more difficult than general reading comprehension QA. It requires deeper linguistic analysis technology to understand the question. The answer extraction should pay more attention to the relation with question. The candidate answer ranking should pay more attention to the semantic correlation between them. Therefore, this paper proposes to extract the candidate answer sentences by frame semantic match and frame-frame semantic relation, and the ranking points are spread through frame semantic relevancy based on manifold-ranking model. At last, the top-four candidate answers are selected based on the ranking points. The accuracy and recall on the college entrance examination of Beijing in recent twelve years is 53.65% and 79.06% respectively.

**Key words:** College Entrance Examination on Chinese; Reading comprehension; Frame semantic; answers sentences extraction;Manifold-ranking

## 1 引言

阅读理解是人们从大量文本中获取信息的重要途径, 很多时候用户并不想浏览全文, 而是想知道某个问题的确切答案。尽管目前的问答系统发展有了长足的进步, IBM 的“沃森(Watson)系统”<sup>[1]</sup>、日本的“高考机器人”、Facebook 推出的在线问答系统成为国际智能信息处理学科取得显著进展的突出标志, 但是在国内, 面对中学教育群体测试的问答系统,

**收稿日期:**           **定稿日期:**

**基金项目:** 国家 863 计划项目(2015AA015407);国家自然科学基金(61373082, 61502287);山西省科技基础条件平台建设项目(2014091004-0103);山西省回国留学人员科研资助项目(2013-015);中国民航大学信息安全测评中心开放课题基金(CAAC-ISECCA-201402);山西省高校科技创新项目(201505)

已有的研究还远远不足。对此，国家 863 计划（语言问题求解和答案生成关键技术课题），提出研究面向大规模群体测试——高考，来推进类人智能问答系统的发展。

根据《2016 年普通高等学校招生全国统一考试大纲》的说明，高考语文论述类文章阅读理解必考内容及相应的能力层级如下：1.理解（1）理解文中重要概念的含义。（2）理解文中重要句子的含义。2.分析综合（1）筛选并整合文中的信息。（2）分析文章结构，把握文章思路。（3）归纳内容要点，概括中心思想。（4）分析概括作者在文中的观点态度。本文主要研究论述类文章阅读理解中问答题的解答，试题样例如表 1。

表 1 高考语文论述类文章阅读理解问答题样例

2004 年北京高考语文第 18 题
<b>问句：</b> 文中揭示的可能的记忆存储手段是什么？请简要回答。
<b>参考答案：</b> CPEB 在朊毒体状态下仍发挥它的正常功能——蛋白质合成，所合成的蛋白质会随着记忆的形成加强突触，使突触和神经细胞长期保存这些记忆。
<b>参考答案对应的答案句（来自阅读材料中）：</b>  更令人意想不到的，CPEB 在朊毒体状态下仍然发挥了它的正常功能——蛋白质合成。  CPEB 合成的蛋白质会随着记忆的形成加强突触，使突触可以长期保存这些记忆。  研究表明，在哺乳动物的神经突触中，CPEB 的朊毒体特征可能就是使突触和神经细胞存储长期记忆的机制，科学家计划对这一理论作进一步的研究。

从参考答案以及对应的答案句可以看出，答案不再是某一候选句的某一语块，而是多个相关答案句的归纳总结。更重要的是，某些答案句可能从词语相似度或句子相似度看起来与问句没有任何关系，从句子语义场景关联的角度看，却与问句密切相关，而且答案句与答案句之间也是彼此关联的。如何把这些答案句抽取出来是一个巨大的挑战，也是高考阅读理解问答技术的关键。

## 2 相关研究

在阅读理解问答中，答案句的抽取直接影响着最终的答案。目前，答案句抽取的方法主要是基于答案句与问句的相似度，即相似度越高该句子排序越靠前，成为答案句的可能性越大。其中有直接基于词语共现或关键词词频（TF）、反文档频率（IDF）等来度量句子间相似度的方法。文献[2]借助潜在语义分析理论，构建了潜在语义空间，并在此空间上计算问句和答案句的相似度。文献[3]利用问句的分类信息对句子进行主题聚类，通过 Aspect Model 能够获得对句子更精确的描述。为了获得词语的同义或更深层的语义信息，有基于 HowNet<sup>[4]</sup>、Chinese FrameNet<sup>[5][6]</sup>、同义词词林<sup>[7]</sup>等语义资源的相似度计算模型。文献[8]提出将问题和所有候选句的语义角色标注结果表示成树状结构，用树核的方法计算语义结构相似度。文献[9]给出了一种基于语料库的日文 why 型问题答案句抽取方法，用 why 型问题答案对语料训练，得到句子或者段落的排序模型，然后在测试集上将检索到的候选句子或段落，用排序模型进行排序，将 Top-1 作为答案。文献[10]针对 why 型问题，提出基于问题话题和话题间因果修辞关系识别的答案句检索方法。文献[11]基于框架语义提出加入框架篇章关系、框架关系及有定零形式线索三个语义特征，获得更优的答案句检索与答案抽取结果。文献[12]利用语法、框架、语义三方面的特征提高机器阅读理解的性能。

另外，2000 年 Riloff 搭建的 Quarc 系统<sup>[3]</sup>，面向 3-6 年级儿童阅读的短文，利用启发式规则发现问句和背景材料的词汇语义线索来抽取答案；文献[14]针对不同的问题类型构建了

大量规则，在中文阅读理解问答上，找到正确答案句的概率为 42%；文献[15]在 WIKIPEDIA(Root and Stem,2010)数据集上，准确率达到了 84%。

利用传统关键词组合匹配或句子相似度的方法只能找到与问句语法结构或语义表述相似度较高的句子，而不能找到与问句语义场景相关的句子；利用有限规则的方法对复杂关联推理问题难以有效求解。

本文针对高考阅读理解问答题的特点，在初步抽取答案句时利用了与问句出处句之间的框架语义相关性；答案句排序时引入流形排序模型，通过答案句之间的框架语义相关度将排序分数进行迭代传播，最终选取排序分数较高的候选句作为答案句。本文的方法不仅在高考阅读理解中具有较好的答题效果，也可以应用在自动文摘、文本推理、以及复杂事实问题求解等自然语言处理任务中。

### 3 汉语框架网

汉语框架网（Chinese FrameNet, CFN）是山西大学以框架语义学为理论基础的大规模词汇语义知识库构建工程，用于语言学、计算语言学以及自然语言处理研究<sup>[16]</sup>。通过框架语义可以挖掘到词语背后隐藏的概念结构和语义场景。

框架网络基于认知理论，以真实语料为研究背景，形式化地表示出句法、语义之间的关系，根本特点就是就是把认知结构引入了语义学。“框架”作为一个语言学术语，是跟一些激活性语境相一致的一个结构化的范畴系统，是储存在人类经验中的图示化情境，这种范畴系统所描述的既可能是一个实体，也可能是一种行为实践模式，甚至是一些社会制度、习俗等<sup>[17]</sup>。

框架是指由词元和它所联系的框架元素构成的表达特定场景的语义结构形式，是理解词语与句子的背景和动因。框架承担词包括动词、形容词、名词和成语，它们是标注工作的着眼点，统称为词元。在句子中能激起框架的词元称为目标词。在汉语框架网中，对于每个框架按照四个方面进行描述：（1）框架定义；（2）核心框架定义；（3）词元；（4）框架关系。汉语框架网能公开使用的框架有 483 个，词元有 5364 个。

框架网络之所以称之为“网”，表现之一就是框架与框架之间的相互关系<sup>[18]</sup>。框架关系从语义场景的角度给出了框架与框架之间的激活关系，区别于语法关系、篇章关系。目前 FrameNet 定义的框架关系有 8 种：继承、透视、使用、总分、起始、参照、因果、先后。

本文利用框架语义抽取答案候选句，一方面由于框架网络包含了比其它词典更为详细的“句法-语义”信息，利用这些信息能有效理解句子的深层语义与语境。一个框架就是一个图示化情境，而一个情境可能触发多个框架，也能抽取多个相关候选句。另一方面，框架语义关系针对的不是相邻句子之间的关系，而是阅读材料中所有句子所揭示的语义场景之间的关系。由于同一场景或者相关场景的描述会在阅读材料的不同位置出现，因此利用框架语义关系可以把相关场景中的句子抽取出来。

### 4 方法描述

阅读理解问答解答过程可以形式化描述为：给定问句  $Q$  与阅读材料  $D$ ，本文首先进行预处理，将背景材料切句： $D=\{S_1,S_2,\dots,S_n\}$ ， $S_i$  表示阅读材料中第  $i$  个句子；对问句提取关键词： $Q=\{K_1,K_2,\dots,K_m\}$ ， $K_j$  为第  $j$  个关键词；其次，利用问句关键词匹配定位问题在阅读材料中的出处句  $S_Q$ ；第三，通过框架语义抽取答案候选句集  $S_C$ ；第四，通过流形排序将答案锁定在候选集中排序较高的 Top-4，作为答案句集  $S_A$ ；最终，进行简单整合生成答案  $A$ 。具体问答系统流程如图 1 所示。

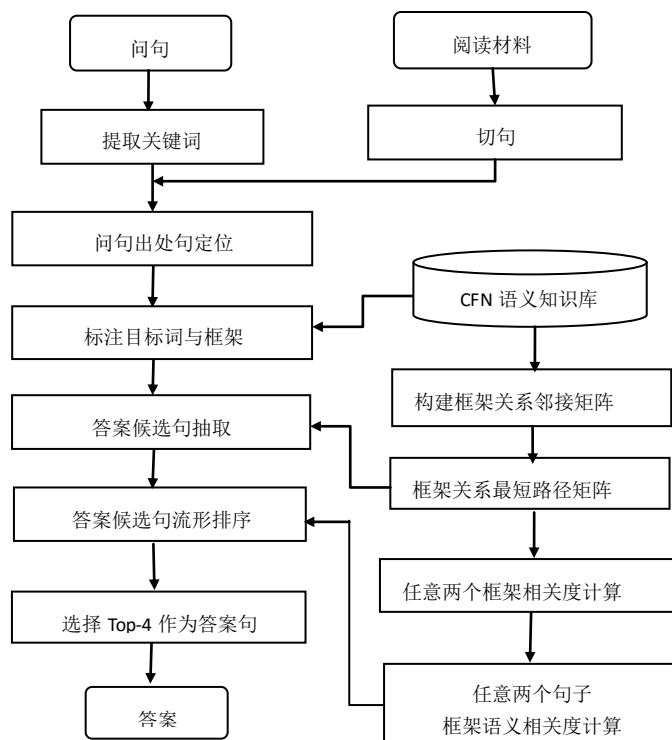


图 1 问答系统流程图

#### 4.1 问句出处句定位

定位问句出处有两个原因：（1）问句的出处比问句拥有更完整的语义场景，能激起多个框架，可以作为问句的查询扩展。（2）问句的出处与答案句有前后勾连的关系，问句出处句激起的框架，为相关答案句的抽取提供更多的知识与线索。一个比较直观的例子如下：

**例 1** 2004 年北京高考语文第 18 题

**问句：**文中揭示的可能的记忆存储手段是什么？请简要回答。

**问句出处句：**此次研究揭示了可能的存储手段，不过让人惊奇的是，‘脱毒体活动’竟然在其中发挥着作用。

目标词 1: 研究；框架 1: 研究

目标词 2: 揭示；框架 2: 证明

目标词 3: 存储；框架 3: 储存

目标词 4: 惊奇；框架 4: 心理刺激

从上面例子可以看出问句中有两个目标词，激起两个框架；问句出处句有四个目标词，激起四个框架，很好的扩展了问句的语义场景，为答案候选句抽取提供了更充足的语义信息。

本文利用关键词匹配定位问句的出处，即与问句关键词匹配数最多的为问句出处。问句提取的关键词将直接影响定位的结果，并不是问句中所有的词都可以作为关键词，经过大量的统计发现，很多停用词（“的”，“什么”，“在”等）、以及问句特有的一些词语（“根据”，“本文”，“简答”等），它们对于问句定位起不到任何作用，可以将这些词在问句分词后直接过滤。例 1 中的问句提取的关键词为：揭示、可能、记忆、存储、手段。

#### 4.2 基于框架语义的答案候选句抽取

答案候选句抽取是问答系统的关键步骤，考虑到问句出处与答案句之间激起的框架相同或框架距离很近，提出借助框架语义匹配和框架语义关系作为主要手段，指代消解作为辅助手段，进行答案候选句抽取。

### 方法一：基于框架语义匹配的答案候选句抽取

框架匹配从与问句语义场景相同的角度抽取相关候选句,当句子激起的框架与问句或问句出处句激起的框架有一对或一对以上相同时,被选为答案候选句。一个框架下有一个或多个词元,一个词元往往激发多个框架,本文使用山西大学 CFN 实验室开发的“汉语句子框架语义分析”平台对篇章中的句子自动标注,标注方法为:该词元在 CFN 例句库中激起次数最多的那个框架。CFN 篇章例句库有 484 篇中文文档,涉及科技、环境、历史等 15 个领域,共有 5351 条句子,由三个人手工标注,交叉验证。例 1 通过框架匹配抽取到的候选句如下:

S<sub>10</sub>: 因此研究人员在发现这种名为 CPEB 的蛋白质具有特定朊毒体特征时感到非常**惊讶**。

目标词:惊讶;框架 4:心理刺激

S<sub>14</sub>: CPEB 合成的蛋白质会随着记忆的形成加强突触,使突触可以长期**保存**这些记忆。

目标词:保存;框架 3:储存

S<sub>17</sub>: 更令人**意想不到**的是,CPEB 在朊毒体状态下仍然发挥了它的正常功能——蛋白质合成。

目标词:意想不到;框架 4:心理刺激

S<sub>18</sub>: **研究表明**,在哺乳动物的神经突触中,CPEB 的朊毒体特征可能就是使突触和神经细胞**存储**长期记忆的机制,科学家计划对这一理论作进一步的**研究**。

目标词:研究;框架 1:研究

目标词:表明;框架 2:证明

目标词:存储;框架 3:储存

将抽取到的这 4 个句子,与表 1 中的参考答案对比,可以发现准确率达到 75%,召回率达到 100%。

### 方法二:基于框架语义关系的答案候选句抽取

答案候选句不只是来源于框架语义匹配度较高的句子,还可能来源于与问句出处句有框架语义关系的句子。利用框架语义关系,从语义场景相关的角度抽取答案候选句,弥补了框架语义匹配的不足。当句子激起的框架与问句或问句出处句激起的框架有一对或一对以上框架距离小于等于 2 时,被选为答案候选句。图 2 为与“研究(Research)”框架距离在 2 以内的框架关系图。图中灰色的椭圆表示当前框架,它周围的椭圆表示与它有框架关系的框架,矩形表示子框架的个数,不同线型的线条表示不同的框架关系。

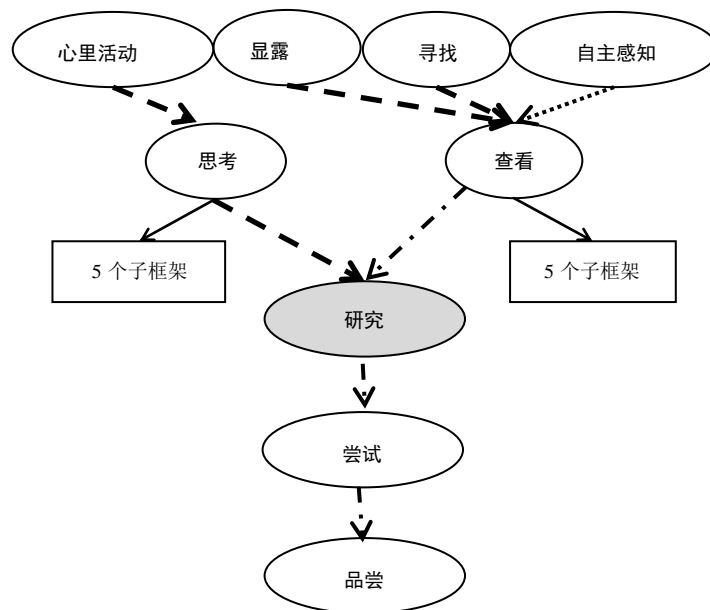


图 2 框架关系示例

表 2 给出了图 2 中与研究有直接关系的四个框架的具体定义，可以看出四个框架都是认知者对某一问题（或现象）的认知活动，“研究”的图示化情境离不开研究者的反复“尝试”、仔细“查看”与认真“思考”的过程。

表 2 框架定义示例

框架	定义
研究	<b>研究者</b> （个人，组织或机构）试图通过参考和 <b>问题</b> 所在 <b>领域</b> 相关的文献、观察或实验来 <b>解决问题</b> 。
查看	该框架指的是 <b>认知者</b> （人或其他智能生命体）对某事物，即 <b>现象</b> ，给予密切关注，比起一般的感知觉活动，查看隐含了人的分析、思考过程。
思考	一位 <b>认知者</b> 在一段时间内对一个 <b>问题</b> 进行考虑，可能是这个人将要进行的一系列行动，也可能是比较抽象的事物。
尝试	一位 <b>评价者</b> 参与对某一 <b>实体</b> （或转喻）的 <b>体验</b> ，往往是第一次尝试，目的是形成对其质量的评价。

除此之外，指代是语篇中常见的语言现象，是指篇章中的一个语言单位（通常是词或短语）与之前或将要出现的语言单位存在特殊的语义关联。指代消解是自然语言处理领域的一项重要重要且复杂的任务。在论述类文章中，常见的是指示代词“这”与人称代词“它”。为了削弱由于指代对答案候选句抽取的影响，本文基于答案候选句的篇章衔接性，对这两个代词做了一个简单指代消解处理：如果通过方法 1 和 2 抽取到的答案候选句句首为“这”或“它”，则将其在阅读材料中的前一句也加入答案候选句集。示例如下：

**例 2** 2007 年北京语文高考第 17 题

**问句：**文章认为核心能力既是一个公司的“福音”，却也可能是“诅咒”，为什么？

**S<sub>9</sub>：**这就是“核心能力的诅咒”。

加入 **S<sub>10</sub>：**当环境发生巨变时，公司难以应对而猝然倒下。

#### 4.3 基于流形排序的答案句选择

为了找到简洁地答案，必须对召回的答案句进行排序筛选。本文使用流形排序算法进行答案候选句排序，充分利用了答案候选句之间的相关性。流形排序算法<sup>[19]</sup>是一种半监督全局排序算法，最初用来对满足流形结构假设的数据进行排序。流形排序基于如下假设：（1）相近的节点倾向于有相近的排序分数；（2）相同结构（特别是被称为类或流形）中的节点倾向于有相近的排序分数。流形排序的直观描述如下：在数据集上构建带权图，对已标记节点赋正值，对其它待排序节点赋零值，然后所有节点通过带权图将它们分数传递给相邻近的节点。迭代地传播节点分数，直到全局稳定状态，最终所有的节点都得到了合理的排序分数。

本文利用 CFN 语义知识资源中的框架关系，构建了框架关键邻接矩阵，并使用 Floyd 算法求出了最短路径矩阵。首次提出了利用最短路径计算两个框架相关度的计算公式。基于最短路径的任意两个框架 F1、F2 框架语义相关度计算公式如下：

$$R(F_1, F_2) = e^{-\frac{d^2}{10}} \quad (1)$$

式中，d 为框架 F<sub>1</sub>、F<sub>2</sub> 的最短路径。框架语义相关度是[0,1]之间的实数。两个框架相同，即完全匹配时，d 为 0，框架相关度为 1。当距离大于 5 时，框架相关度就趋于 0。则基于框架相关度的句子相关度计算公式如下：

$$\text{Rel}(A, B) = \frac{\sum_{i=1}^m \max(R(A_i, B))}{m} + \frac{\sum_{j=1}^n \max(R(B_j, A))}{n} \quad (2)$$

式中  $A, B$  为任意两个框架,  $\max(R(A_i, B))$  是句子  $A$  中框架  $A_i$  与句子  $B$  中任意框架相关度的最大值, 同理  $\max(R(B_j, A))$  是句子  $B$  中框架  $B_j$  与句子  $A$  中任意框架相关度的最大值。

本文将问句出处句和答案候选句作为节点, 流形排序的过程可以形式化为: 给定节点集  $X = \{x_0, x_1, \dots, x_n\}$ ,  $x_0$  是问句出处句, 其余  $n$  个节点是答案候选句。函数  $f: X \rightarrow R$  定义一个排序函数, 该函数对各个节点  $x_i$  赋予其排序分数值  $f_i$ 。  $f$  可以看做向量  $f = [f_0, f_1, \dots, f_n]^T$ 。定义向量  $y = [y_0, y_1, \dots, y_n]^T$  表示各个节点的初始值, 其中  $y_0 = 1$ , 其余  $n$  个  $y_i = 0$ 。算法如下:

**输入:** 答案候选句集合  $X$ , 候选句初始值向量  $y$

**输出:** 答案候选句排序分数

1. 标注每个句子激起的框架。把每个句子表示成它所激起的框架的集合。
2. 利用公式 (2) 计算集合  $X$  中任意两个句子的框架语义相关度。
3. 构建答案候选句相关度矩阵

将候选句作为节点, 候选句之间的框架相关度作为节点之间边的权值。节点之间的相关度大于 0, 就在这两个节点之间产生一条以相关度作为权值的边。候选句集合可以生成一个无向图  $G$  来反应句子之间的关系。定义对称矩阵  $W$  作为候选句相关度矩阵, 其中  $W_{ij} = R(x_i, x_j)$ 。为了防止节点在第 4 步中自加强, 定义  $W_{ii} = 0$ 。

4. 关联度矩阵归一化

定义对角矩阵  $D$ , 其中  $D_{ii}$  表示相关度矩阵第  $i$  行的和, 相关度矩阵归一化为  $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ 。

5. 迭代  $f(t+1) = \alpha S f(t) + (1-\alpha)y$ , 直到收敛, 其中  $\alpha \in [0, 1]$ 。
6. 用  $f_i^*$  表示收敛序列  $\{f_i(t)\}$ , 这样每个句子就得到了它的排序分数  $f_i^*$ 。

图 3 流形排序算法

算法第二步定义  $S_{ii} = 0$  保证各个节点在迭代过程中不会自加强; 第三步对相关度矩阵归一化保证了迭代算法的收敛性; 第四步是算法的关键步骤, 迭代过程中使用参数  $\alpha$  来平衡临近节点和初始分数节点对其他节点分数的影响: 参数  $\alpha$  值越接近 1, 临近节点对节点分数的影响越大, 参数  $\alpha$  值越接近 0, 初始分数节点对节点分数的影响越大。根据流形排序算法, 当连续的两次迭代计算的分数之间的差异低于一个给定的阈值时, 认为收敛, 且收敛序列在有限次内就可以得到。利用极限的思想, 假设在某一时刻  $f(t)$  收敛于  $f^*$ , 我们用  $f^*$  代替了第五步  $f(t+1) = \alpha S f(t) + (1-\alpha)y$  中的  $f(t)$  与  $f(t+1)$ , 则

$$f^* = \alpha S f^* + (1 - \alpha)y \quad (3)$$

推导出,  $f^* = (1-\alpha)(1-\alpha S)^{-1}y$ , 又因为  $(1-\alpha)$  不影响排序的结果, 故最终的排序分数公式为:

$$f^* = (1 - \alpha S)^{-1}y \quad (4)$$

经过对高考试题答案句的统计发现, 答案句为 2 句的比例约为 15.3%; 答案句为 3 句的比例约为 38.6%; 答案句为 4 句的比例约为 46.1%。因此, 利用流形排序得到所有节点的排序分数后, 选择排序分数较高的 Top-4 作为答案句。为了答案的逻辑可读性, 把答案句按照在阅读材料中出现的自然顺序输出, 并且去除句首连词, 生成答案。

## 5 实验

## 5.1 实验数据

由于高考论述类文章阅读理解真题语料中问句比例较小，本文把一些选择题进行改造，形成问答题。题型改造示例如表 3。

表 3 题型改造示例

2015 北京高考选择题	改造后成为问答题
下列对中国印刷业面临“巨大的危机和挑战”的原因，概括不正确的一是	中国印刷业面临“巨大的危机和挑战”的原因是什么？
A. 研究进程的落后	【参考答案】
B. 计算机的历史短	(1) 研究进程的落后
C. 汉字自身的特征	(2) 汉字自身的特征
D. 技术条件不具备	(3) 技术条件不具备
【参考答案】B	

从表中可以看出，通过这种方法扩充问答语料真实可靠。最终在各省近十二年 106 篇高考真题语料上形成 117 个问句，作为训练集进行参数调节，在 574 篇高考模拟语料上形成了 593 个问句，作为测试集。

## 5.2 实验评价与结果分析

### 5.2.1 评价指标

本文咨询了资深的高考语文判卷专家，目前阅读理解问答题判卷按答题要点给分，即答中要点就得分。本文按照试题所给的参考答案人工找到其在阅读材料中对应的多个句子，标记为答案句的集合  $A'$ ，集合  $A'$  的大小，就是答案句的句子总数。 $S_A$  是使用本文方法，排序后选择的 Top-4 个句子构成的集合，本文的评价指标如下：

$$P(\text{准确率}) = \frac{S_A \text{中正确答案句的个数}}{S_A \text{中句子总数}} \times 100\% \quad (5)$$

$$R(\text{召回率}) = \frac{S_A \text{中正确答案句的个数}}{A' \text{中句子总数}} \times 100\% \quad (6)$$

### 5.2.2 不同方法实验结果比较

为了验证本文方法的有效性，设置了三个基于相似度的 baseline 方法在测试集上进行对比。问句  $Q$  的关键词为  $A_i$ ，共  $m$  个，句子  $S$  的关键词为  $B_j$ ，共  $n$  个，则

(1) 基于词共现的句子相似度<sup>[20]</sup>

$Q$  与  $S$  词共现的个数为  $c$ ，则句子相似度为：

$$Sim(Q, S) = \frac{2c}{m+n} \quad (7)$$

(2) 基于《知网》的句子相似度<sup>[4]</sup>

两个词基于知网的相似度为  $S(A_i, B_j)$ ，令  $a_i = \max\{S(A_i, B_1), S(A_i, B_2), \dots, S(A_i, B_n)\}$ ； $b_j = \max\{S(B_j, A_1), S(B_j, A_2), \dots, S(B_j, A_m)\}$ ，则句子相似度为：

$$Sim(Q, S) = \frac{\sum_{i=1}^m a_i + \sum_{j=1}^n b_j}{2} \quad (8)$$



(3) 基于 word2vector 的句子相似度<sup>[21]</sup>

利用北京理工大学在百度百科上训练得到的词向量 word embedding 库 (54 万),  $A_i^v$  与  $B_j^v$  分别为  $A_i$  与  $B_j$  的词向量, 两个词的相似度为对应的词向量的余弦相似度:

$$Cos(A_i^v, B_j^v) = \frac{A_i^v \cdot B_j^v}{\|A_i^v\| \times \|B_j^v\|} \quad (9)$$

令  $a_i = \max\{Cos(A_i^v, B_1^v), Cos(A_i^v, B_2^v), \dots, Cos(A_i^v, B_n^v)\}$ ;  $b_j = \max\{Cos(B_j^v, A_1^v), Cos(B_j^v, A_2^v), \dots, Cos(B_j^v, A_m^v)\}$ , 再利用公式 (8) 计算句子相似度。

在表 4 中, 将本文抽取候选句的方法也逐级进行对比, (1) ~ (6) 都是按照句子与问句的相似度或相关度大小直接排序, (7) 使用流形排序模型。用 Top-4 的准确率、召回率来评价答案句抽取的结果, 实验结果如表 4 所示。

表 4 不同方法答案句抽取结果比较

实验方法		Top-4	
抽取方法	排序方法	P (%)	R (%)
(1)基于词共现的句子相似度	直接排序	28.54	36.78
(2)基于《知网》的句子相似度	直接排序	29.16	39.17
(3)基于word2vector句子相似度	直接排序	31.69	43.74
(4)框架匹配	直接排序	34.50	47.08
(5)框架匹配+框架关系	直接排序	39.34	55.75
(6)框架匹配+框架关系+指代消解	直接排序	40.36	57.41
(7)框架匹配+框架关系+指代消解	流形排序	43.49	62.28

实验结果表明本文的方法比三个 baseline 的方法答案句抽取效果都要好, 利用框架匹配准确率与召回率分别提高 2.81% 与 3.34%, 加入框架匹配后, 又分别提高了 4.84% 与 8.67%, 加入指代消解后, 又分别提高了 1.02% 与 1.66%, 加入流形排序后分别提高了 3.13% 与 4.87%, 说明本文提出的基于框架语义的答案候选句抽取方法和基于流形排序的答案候选句筛选方法是切实有效的。

### 5.2.3 不同省份考题实验结果比较

各个省份高考论述类文章阅读理解问答题存在以下差异:

(1) 阅读材料体裁不同: 北京高考为科技类文本, 分为自然科学类与社会科学类。其他省份为时

评、短评、书评、传记、新闻、报告、科普文等。

(2) 考查难点不同: 北京地区问答题大多考查筛选并整合文中信息, 别的省份大多考查对重要词句的理解、把握文章结构与整体思路。因此, 本文的方法在不同省份的实验结果差异较大。各个省份以及模拟题 Top-4 的准确率与召回率如图 4 所示。

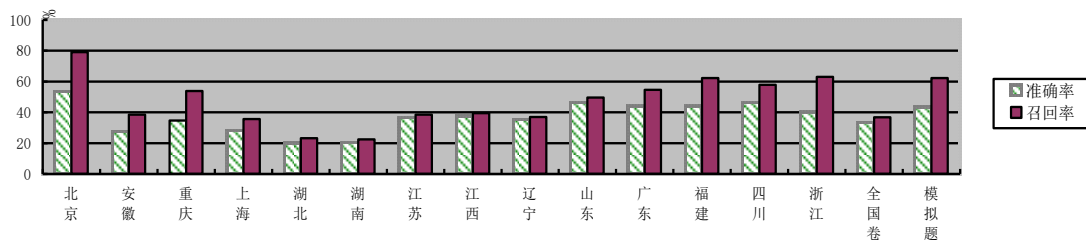


图 4 各省份答案句抽取结果比较

在详细分析每个省份实验结果的过程中，发现本文的方法在北京高考题获得了最高 79.06% 的召回率与 53.65% 的准确率。此外，本文发现有些省份召回率与准确率非常低。如例 3。

### 例 3 2005 年湖南语文高考第 8 题

**问句：**新生儿的哪些行为能证明“人类大脑的自我构成功能并非只从诞生时刻才开始”？

**答案句：**1. 母亲只要觉察到腹中胎儿在动，每天大声念两遍同一则童话，到分娩的那一天，婴儿生下来后就**知道**这个童话。2. 新生儿通过改变吮吸奶头的速度，对他母亲念的本原童话录音和另一则在形式上作了改动的童话录音进行**选择**。3. 如果母亲会两种语言，她的新生儿能够**区分**这两种语言，并喜欢从中**选择**自己还在母腹中时听得最多的那种语言。

不难发现，问句与答案句之间存在巨大的语义鸿沟，问句中抽象的“行为”与答案句中的“知道”、“选择”、“区分”等具体动作难以建立语义关系，而且也无法判断阅读材料中出现过的行为哪些能证明“人类大脑的自我构成功能并非只从诞生时刻才开始”。对于一些深层的抽象概念与经验知识的理解与推理，目前汉语框架网还不能完全做到，这直接导致了这类问句难以获得良好的自动答案句抽取效果。

#### 5.2.4 $\alpha$ 对实验效果的影响

在对参数  $\alpha$  的实验中，本文令  $\alpha$  从 0.2-1 以 0.2 为间隔的 5 个不同的值，记录 Top-4 在训练集上的准确率与召回率，如图 5 所示。从图中可以看出， $\alpha$  为 0.8 时，准确率与召回率都最高，说明句子间框架关系对节点分数的影响很大。在实验中  $\alpha$  的取值为 0.8。

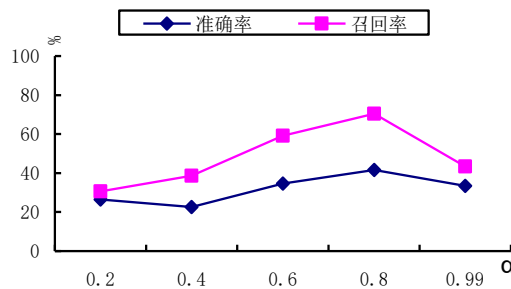


图 5 不同  $\alpha$  对实验效果的影响

## 6 结论与展望

本文针对高考语文论述类文章阅读理解，从认知的角度，提出一种基于框架语义的高考语文阅读理解方法，该方法以框架语义匹配和框架语义关系进行答案句抽取，最后利用流形排序模型选取答案句。基于框架距离的框架语义相关度的计算有效的利用了汉语框架语义知识资源。句子框架语义相关度弥补了基于相似度抽取答案句方法的缺点，从语义场景逻辑相关的角度抽取与筛选答案候选句，有效提高了高考阅读理解答案句抽取的准确率与召回率。目前，CFN 的词元、框架关系还较少，不足以覆盖所有的图示化情景，特别是在一些领域性较强的文本上存在一定的局限性。在未来的工作中，一方面要扩充 CFN 框架、词元知识库，预测更多的框架语义关系，另一方面还需利用上下文结构信息、哈工大开发的 LTP“语言云”技术平台以及构建一些高考语文知识库，逐步提升答题效果。

**致谢** 本文实验用到中国科学院计算技术研究所的 NLP/ICTCLAS2015 分词系统，北京理工大学提供的词向量库，山西大学 CFN 项目组开发的“汉语句子框架语义分析”平台 (<http://115.24.12.89:8080/autoanno/>)，在此表示感谢！

## 参考文献

- [1] 黄昌宁. IBM 深度问答系统战胜顶尖人类选手所想到的[J]. 中文信息学报, 2011, 25(6):21-25.
- [2] 余正涛, 樊孝忠, 郭剑毅, 耿增民. 基于潜在语义分析的汉语问答系统答案提取. 计算机学报, 2006, 29(10): 1889-1893.
- [3] 吴友政, 赵军, 徐波. 基于主题语言模型的句子检索算法. 计算机研究与发展, 2007, 44(2): 288-295.
- [4] 刘群, 李素建. 基于《知网》的词汇语义相似度[J]. 中文计算语言学. 2002, 7(2): 59-76.
- [5] Li Ru, Li Shuanghong, Zhang Zezheng. The Semantic Computing Model of Sentence Similarity Based on Chinese FrameNet[C]// 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, 2009:255-258.
- [6] 李茹, 王智强, 李双红, 等. 基于框架语义分析的汉语句子相似度计算[J]. 计算机研究与发展, 2013, 50(8):1728-1736.
- [7] 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报:信息科学版, 2010, 28(6):602-608.
- [8] 张志昌, 张宇, 刘挺等. 基于浅层语义树核的阅读理解答案句抽取. 中文信息学报, 2008, 22(1):80-86.
- [9] Higashinaka R, Isozaki H. Corpus-based Question Answering for Why-questions. Proceedings of IJCNLP08. Hyderabad, India: AFNLP, 2008: 418-425.
- [10] 张志昌, 张宇, 刘挺, 李生. 基于话题和修辞识别的阅读理解 why 型问题回答. 计算机研究与发展, 2011, 54(2): 216-223.
- [11] 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(04): 795-807.
- [12] Wang H, Bansal M, Gimpel K, et al. Machine Comprehension with Syntax, Frames, and Semantics[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015.
- [13] Riloff E, Thelen M. Rule-based Question Answering System for Reading Comprehension Tests[J]. Proceedings of Workshop on Reading Comprehension, NAACL/ANLP-2000, 2000.
- [14] Hao X, Chang X, Liu K. A Rule-based Chinese Question Answering System for Reading Comprehension Tests[C]// International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE, 2007:325-329.
- [15] Akour M, Abufardeh S, Magel K, et al. QArabPro: A Rule Based Question Answering System for Reading Comprehension Tests in Arabic.[J]. American Journal of Applied Sciences, 2011, 8(6):652-661.
- [16] 由丽萍. 构建现代汉语框架语义知识库技术研究[D]. 上海师范大学, 2006.
- [17] 郝晓燕, 刘伟, 李茹, 等. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报, 2007, 21(5):96-100.
- [18] 郝晓燕. 汉语框架网络工程构建及应用[M]. 电子工业出版社, 2011.
- [19] Zhou, Dengyong, Weston, et al. Ranking on Data Manifolds[J]. Advances in Neural Information Processing Systems, 2003:169-176.
- [20] 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.
- [21] 刘敏. 基于词向量的句子相似度计算及其在基于实例的机器翻译中的应用[D]. 北京理工大学, 2015.



李国臣（1963—），男，教授，硕士生导师，主要研究方向为中文信息处理。E-mail:lgc1017@163.com



刘姝林（1990—），女，硕士研究生，主要研究方向为中文信息处理。E-mail:shulin\_sxu@sina.com



杨陟卓（1983—），男，讲师，博士，主要研究方向为计算语言学、词义消歧。E-mail:yangzhizhuo\_662@163.com



李茹（1963—），女，教授，博士生导师，主要研究方向为中文信息处理、数据智能预测。E-mail: liru@sxu.edu.cn



张虎（1979—），男，讲师，博士，主要研究方向为计算语言学。E-mail:Zhanghu@sxu.edu.cn



钱捍丽（1977—），女，博士，副教授，硕士生导师，主要研究方向为自然语言处理。E-mail:qyl@sxu.edu.cn