基于汉语篇章框架语义分析的阅读理解问答研究

王智强[1] 李 茹[1),2)] 梁吉业[2)] 张旭华[1)] 武 娟[1)] 苏 娜[1)]

[1)](山西大学计算机与信息技术学院 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 太原 030006)

摘 要 答案句检索和答案抽取是阅读理解中的两个核心技术.针对汉语阅读理解,该文提出一种新的基于篇章框架语义分析的答案句检索和答案抽取方法.答案句检索是基于框架相似性、框架关系及篇章框架关系来实现.其中,基于框架相似性的方法是通过计算背景材料与问句之间语义场景(框架)的相似度来进行答案句检索;基于框架关系和篇章框架关系的方法可以从语义相关角度获得与问句语义相关的答案句.在答案抽取时,提出基于框架语义相似性、有定零形式线索及框架关系的答案抽取方法.基于框架语义相似性可以从语义相似的答案句中抽取出充当问句疑问角色的框架元素作为答案;有定零形式线索能够在篇章范围定位答案句中充当答案的缺失语义成分;框架关系则能够通过建立框架元素之间的关系,抽取相关度高的框架元素作为答案.针对 15 个领域的 552 个阅读理解问题,该方法在答案句检索时相比传统基于相似度的方法能够获得更好的答案句检索结果;相比基于框架相似性的 Baseline 实验,加入篇章框架关系、框架关系及有定零形式线索的篇章级框架语义特征,能够获得更优的答案句检索与答案抽取结果.

关键词 框架语义分析;阅读理解;框架关系;有定零形式;篇章框架关系;社会媒体;社交网络;自然语言处理

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2016.00795

Research on Question Answering for Reading Comprehension Based on Chinese Discourse Frame Semantic Parsing

WANG Zhi-Qiang[1)] LI Ru[1),2)] LIANG Ji-Ye[2)] ZHANG Xu-Hua[1)] WU Juan[1)] SU Na[1)]

[1)](*School of Computer & Information Technology, Shanxi University, Taiyuan* 030006)

[2)](*Key Laboratory of Computation Intelligence & Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006)

Abstract Answer-sentence retrieval and answer extraction are two core techniques in Reading Comprehension (RC). This paper proposed a new method of answer-sentence retrieval and answer extraction for Chinese RC based on discourse frame semantic parsing. At the stage of retrieving answer sentences, frame similarity, frame-frame relations and discourse frame-frame relations are employed in the new method. Specifically, frame similarity is used to compute the semantic scenarios (i.e. frames) similarities between questions and reading materials; frame-frame relations and discourse frame-frame relations are utilized to obtain answer sentences from the perspective of semantic relevance. After that, frame similarity, defined null instantiation, and frame-frame relations are applied in the process of answer extraction. Among them, frame similarity is used to extract the frame element which serves as question role from answer sentences with semantic similarities and the extracted frame element is treated as the final

收稿日期:2014-12-12;在线出版日期:2015-11-18.本课题得到国家自然科学基金(61373082,61432011,U1435212)、山西省科技基础条件平台建设项目基金(2014091004-0103)、山西省回国留学人员科研资助项目基金(2013-015)、国家"八六三"高技术研究发展计划项目基金(2015AA015407)和中国民航大学信息安全测评中心开放课题基金项目(CAAC-ISECCA-201402)资助.**王智强**,男,1987 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为中文信息处理、社会媒体数据挖掘.E-mail:zhiq.wang@163.com.**李 茹**(通信作者),女,1963 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为中文信息处理、信息检索.E-mail:liru@sxu.edu.cn.**梁吉业**,男,1962 年生,教授,中国计算机学会(CCF)杰出会员,主要研究领域为粒计算、数据挖掘与机器学习.**张旭华**,男,1990 年生,硕士研究生,主要研究方向为中文信息处理.**武 娟**,女,1991 年生,硕士研究生,中国计算机学会(CCF)会员,主要研究方向为中文信息处理.**苏 娜**,女,1989 年生,硕士研究生,主要研究方向为中文信息处理.

answer; defined null instantiation is utilized to locate the exact position of the missing semantic role in discourses and the missing semantic roles is considered to be the final answer; because frame-frame relations can establish the relation between frame elements, they are used to extract the highly relevant frame element and the highly relevant frame element is seen as the final answer. To evaluate the method, 552 RC questions in 15 different fields are utilized, and our method performs better than some traditional similarity-based methods in answer-sentence retrieval. Compared with the baseline experiment based on frame semantic similarity, the addition of discourse-level frame semantic features like frame-to-frame relations, discourse-frame-frame relations and defined null instantiation clues, gains better results in answer-sentences retrieval and answer extraction.

Keywords frame semantic parsing; reading comprehension; frame-frame relation; defined null instantiation; discourse-frame-frame relation; social media; social networks; natural language processing

1 引 言

阅读理解(Reading Comprehension,RC)技术作为问答(Question Answering,QA)技术研究的重要分支[1],2000 年以来一直受到国际与国内自然语言处理研究领域研究同行的关注.与传统问答有所不同,阅读理解任务的目标在于理解一篇文档并对提出的问题返回答案,它更注重问题与背景材料的语义分析以及答案的抽取.阅读理解对语言的分析技术有着更高要求,深度阅读理解技术将对推动问答智能具有重要意义.

阅读问答可以形式化描述为:给定问题 S_0 与背景材料(篇章)$D=\{S_1,S_2,\cdots,S_N\}$,在问题与篇章的分析基础上,首先定位问题答案所在的句子 $S_i\subseteq D$,进一步从语句 S_i 中抽取问题的准确答案 A,A 在这里指答案句 S_i 中的语块.例如:

给定问题:

S_0="蒙古族的祖先最早居住在哪儿?"

若从背景材料中检索到的答案句为:

S_i="他们早先居住在额尔古纳河一带,七世纪时西迁到鄂嫩河——石勒喀河和克鲁伦河流域,活动范围大至整个蒙西高原,由许多部落组成."

那么答案最终抽取结果为

A="额尔古纳河一带"

可以看到,例中答案 A 是从答案句 S_i 中抽取的,A 是答案句 S_i 的语块.

本文试图从框架语义分析角度,在对问题与背景材料进行篇章级框架语义分析的基础上,提出基于框架相似性、框架关系及篇章框架关系的答案句检索方法;进一步提出基于框架元素相似性、有定零形式线索及框架关系的答案抽取方法.为阅读问答的答案句检索与答案抽取提供一种新的解决途径.在本文中,框架、框架关系、篇章框架关系及有定零形式线索,统称为框架语义.

2 相关研究

答案句检索与答案抽取技术是阅读问答的两个关键技术.目前在答案句检索方面主要有:基于句子相似性的答案句检索、基于机器学习的答案句检索以及基于篇章分析的答案句检索;在答案抽取方面有基于模式匹配的答案抽取和基于机器学习的答案抽取.

2.1 答案句检索

基于相似性答案句检索的一个重要前提假设是:问题与句子间的相似性越高,那么当前句子越有可能成为答案句.相似度计算方法中有直接基于关键词匹配或关键词特征权重如词频(TF)、反文档频率(IDF)等来度量句子间相似度[2-4];针对词的同义或多义现象,文献[5]提出一种基于潜在语义分析(Latent Semantic Analysis,LSA)的问题和答案句子相似度计算方法,通过语义的潜在空间改善了词的同义和多义问题;文献[6]基于主题语言模型进行汉语问答系统检索,利用主题聚类与 Aspect Model,能够获得对句子语言模型更精确的描述;为了获得更深层的语义表示,有基于 HowNet[7]、Chinese FrameNet[8]、同义词词林[9]等语义资源的相似度计

算模型.

在基于机器学习的答案句检索方面，文献[10]将答案句检索看作机器学习的分类问题，提取问题和答案句的特征来训练答案句检索的最大熵模型. 文献[11]给出了一种 why 型问题答案句检索方法，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，然后检索得到多个答案句子或者段落，用排序模型对这些句子或者段落进行排序，将 Top-1 作为答案.

基于篇章分析的答案句检索中，文献[12]针对阅读理解问答中 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[13]认为篇章中有关 why 型问题答案的句子，往往不是一个名词短语形式，而是跨越了多个句子并且这些句子之间具有某种篇章关系，最终基于篇章结构树的索引方法进行答案句检索. 文献[14]基于修辞结构理论分析了篇章结构在确定复杂答案方面能起到较大作用.

2.2　答案抽取

在基于模式匹配的答案抽取研究方面，文献[15]采取模式匹配、语言学特征匹配、词汇语义关联推理以及上下文辅助策略实现答案句的获取，提高 RC 系统性能. 文献[16]提出了一种基于模式学习的中文答案句子模式获取方法，利用搜索引擎与人工标注建立问答训练语料，并通过统计学习提取答案句模式. 文献[17]提出一个基于语义模板的用户交互问答系统，语义模板用于标识句子中各成分之间的语义关联关系. 文献[18]提出了一种基于表面文本模式匹配的答案提取算法，利用广义后缀树算法提取句子公共字符串，对公共字符串经过过滤后生成答案模板.

在基于机器学习的答案抽取方面，文献[19]提出基于句法结构特征分析及分类技术的答案提取算法. 该方法将答案提取问题看成是候选答案的分类问题，通过简单特征和句法结构特征训练分类器. 文献[20]提出了一种基于无监督学习的问答模式抽取技术. 该算法只需用户提供每种提问类型两个或以上的提问实例，就可以通过 Web 检索、主题划分、模式提取、垂直聚类和水平聚类等步骤完成该类型提问的答案模式学习.

3　汉语篇章框架语义分析

汉语篇章框架语义分析是一种以汉语框架网(Chinese FrameNet，CFN)[21]为依据，在句子级框架语义分析基础上[8]，利用框架关系、有定零形式线索和篇章框架关系来进行篇章级语义分析的方法. 其中句子级框架语义分析是在 CFN 的语义表示体系下，对句子进行目标词识别、框架排歧、框架语义角色标注来获得句子的框架语义结构. 句子级框架语义分析是篇章级框架语义分析的基础，可参考文献[22-23].

框架关系是 FrameNet 构建的框架与框架之间的语义关系，能够从语义场景角度为篇章框架语义单元间建立关联，为篇章语义理解提供了一种框架语义所特有的方式；有定零形式线索则是为篇章中语义场景缺失的语义参与者在篇章范围内找回相应的框架元素，有定零形式线索能够将句子级局部的语义参与者(框架元素)扩展至篇章范围，是进行篇章语义分析的重要线索，类似的任务在国际上曾于 SemEval-2010 语义评测[24]中被提出；篇章框架关系则更加关注篇章的语义逻辑与连贯性. 本文的篇章级框架语义分析就是建立在这 3 种不同语义形式下的语义分析.

3.1　框架关系

FrameNet[25]的框架关系揭示了框架(语义场景)间的相关关系，用三元组 $N(F,E,R)$ 表示框架关系网. 其中，F 表示框架关系网中所有框架 $f_i(i\in[1,N])$ 的集合；E 表示框架关系网中框架间边的集合，如果框架 f_i 与 f_j 之间存在连边，那么有 $\langle f_i,f_j\rangle\in E$；$R$ 则表示框架关系网中框架间的关系类型 r_k 的集合. 目前 FrameNet 中已有的框架关系类型包括继承、视角、使用、总分、因果、起始和参照共 7 种($k\in[1,7]$). 图 1 为从 FrameNet 框架关系网中截取的部分框架关系示例.

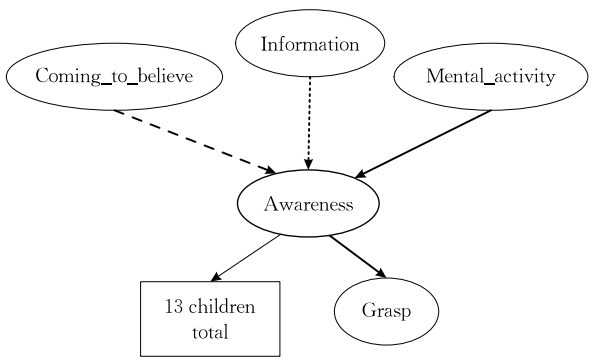


图 1　FrameNet 框架关系示例

图 1 中椭圆表示框架，不同线型区分不同的框架关系.“13 children total”表示 Awareness 框架还

包括 13 个子框架.

框架关系是语言学家基于框架语义学[24]建立的语义场景之间的关系,不同的框架关系类型具有不同的语义含义.

继承关系. 也称父子关系,表示子框架的语义角色是从父框架继承而来;

视角关系. 一些框架从不同角度理解会形成不同视角下的语义,比如对于购买框架,从消费者和销售者的角度形成了购买框架和销售框架;

使用关系. 在父子框架中,子框架的框架元素属于父框架的一部分,在构建子框架时会使用父框架中的框架元素等,从而形成使用关系;

总分关系. 对于复杂的语义场景,又可分为多个小语义场景,在复杂场景和小语义场景之间构成总分关系;

因果关系. 一般指行为框架之间的因果关系,其中一个行为框架是另一个行为框架的原因,反过来为结果;

起始关系. 针对状态框架而言,若一个状态是另一个状态的开始,那么另一个状态就是这个状态的结束;

参照关系. 对于框架多义现象,为了区分其语义关系,建立了一种参照关系.

不同的框架关系具有特有的性质,如框架"给予"与"提供"有"使用"关系,它们之间的框架元素具有映射关系,如图 2.

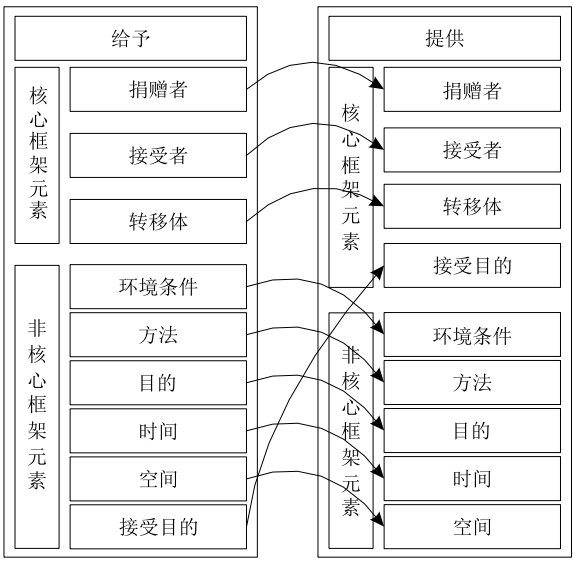


图 2 框架"给予"与"提供"框架元素映射

这种由框架关系建立的框架元素之间的映射,是本文进行篇章级框架语义分析的重要依据. 本文所用框架关系来源于英文 FrameNet. 针对 CFN 现有的 234 个框架,课题组构建了 268 个关系.

3.2 有定零形式线索

给定句子 S,对其进行句子级框架语义分析[21],即目标词识别、框架排歧及框架语义角色标注. 对于 S 中任意目标词 w,它激起的框架为 F,CFN 中一个框架 F 所包含的框架元素集合为 $E=\{e_1,e_2,\cdots,e_M\}$,若其中只有部分框架元素 $SubE\subset E$ 在句子 S 中被实现,那么未被实现的框架元素 $E\text{-}SubE$ 为句子中缺失的框架元素,称之为零形式. 而在这些零形式中,一些框架元素通过篇章范围可以被填充,这种能够被填充的框架元素称为有定零形式(Definite Null Instantiation,DNI)框架元素[26].

图 3 为一个篇章片段,其中句 S_2 中框架 F_2 的框架元素存在缺失,即出现 DNI 线索,而 DNI 必须通过篇章上下文才能够被理解. 此例中用句 S_1 的框架元素"天葬卫星"作为 DNI 的填充对象.

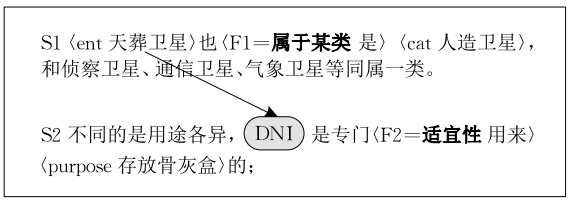


图 3 有定零形式线索

3.3 篇章框架关系

给定篇章 $D=\{S_1,S_2,\cdots,S_N\}$,其中 $S_i(i\in[1,N])$为篇章 D 中的第 i 个句子,对篇章的所有句子进行框架语义分析,篇章中相邻框架会按一定的语义关系结合形成最基本的结构,然后语义上存在关系的相邻结构继续结合,最终通过语义关系形成一棵有层次的篇章框架语义结构树. 其中的语义关系我们称之为框架之间的篇章框架关系,这里的关系不同于 FrameNet 中的框架关系,它表示框架单元在篇章中的语义逻辑性与连贯性.

本文所用篇章语义关系主要参考黄国荣和廖序东的《现代汉语》[27],并结合 CFN 框架语义表示特点,确定了 3 层篇章框架关系结构. 第 1 层根据篇章框架单元间语义是否平等,将篇章框架关系划分为联合与偏正两大类. 其中,联合关系揭示了篇章单元间语义是平等的,偏正关系指篇章单元间语义不平等. 第 2 层篇章框架关系中,联合关系又细分并列关系、承接关系、递进关系、选择关系、解说关系;偏正关系分为条件关系、假设关系、因果关系、目的关系、转折关系和属于关系. 第 3 层级篇章框架关系是根据前后篇章框架单元的语义发展顺序以及逻辑关系

进一步细分为 22 类：平列关系、对比关系、时间关系、空间关系、事理关系、连贯关系、一般递进关系、衬托递进关系、未定选择关系、已定选择关系、解释关系、总分关系、有条件关系、无条件关系、背景条件关系、环境条件关系、一致假设关系、让步假设关系、说明因果关系、推理因果关系、得到关系和避免关系.

在篇章框架语义结构树中，叶子结点是从每个篇章单元中抽取出来的框架，结构树的内部结点可以用一个二元组 $T(F,R)$ 来表示. 其中，F 表示篇章中所有框架 $f_i(i\in[1,N])$ 的集合；R 表示篇章框架关系类型 R_k 的集合. 图 4 为给定篇章示例的篇章结构树.

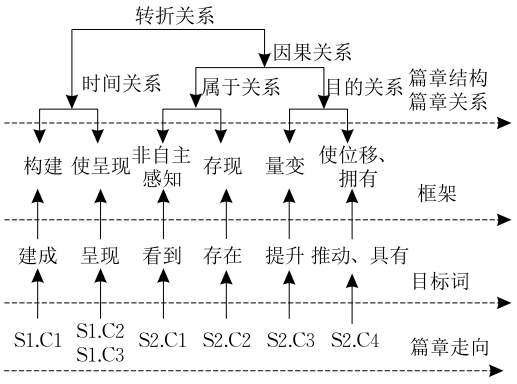


图 4　篇章框架语义结构树

给定基于框架语义分析的语篇示例：

S1. C1 自 2010 年中国东盟自贸区〈tgt＝构建 建成〉以来，S1. C2 中国云南和广西多个与越南、缅甸和老挝接壤的边境口岸，S1. C3 贸易量均〈tgt＝使呈现 呈现〉井喷状态. S2. C1 但本报记者在调查中〈tgt＝非自主感知 看到〉，S2. C2 东南亚的公路建设仍〈tgt＝存现 存在〉一些“硬件”和“软件”问题，S2. C3 加速〈tgt＝量变 提升〉这一地区的公路网，S2. C4 对〈tgt＝使位移 推动〉各国经济发展和中国东盟合作〈tgt＝拥有 具有〉重要作用.

建立其篇章框架语义结构树，如图 4.

篇章框架语义结构树提供了篇章的语义结构信息，在阅读理解的答案句检索时，可以基于结构树，借助篇章框架关系抽取与问题相关的答案句.

4　基于篇章框架语义分析的阅读问答

4.1　答案句检索

答案句检索是本文阅读理解问答的第 1 阶段. 以下将介绍基于篇章框架语义分析的 3 种答案句检索方法：基于框架相似性的答案句检索；基于框架关系的答案句检索；基于篇章框架关系的答案句检索.

4.1.1　基于框架相似性的答案句检索

基于框架相似性的答案句检索是在对问题与篇章进行框架语义分析的基础上，直接对问句与候选答案进行框架相似度计算，通过相似度高低来判断答案句. 给定篇章 $D=\{S_1,S_2,\cdots,S_N\}$ 与问题 S_0，对问句与背景材料的所有句子进行句子级框架语义分析，获得篇章 D 中每个句子 $S_i(i\in[1,N])$ 的框架语义分析表示 $F_i(f_i,t_i,E_i,P_i)$ 以及问句 S_0 的框架语义分析表示 $F_0(f_0,t_0,E_0,P_0)$. 其中，f_i 代表句子 S_i 中目标词 t_i 所激起的框架；E_i 表示句子 S_i 中的框架元素集合；P_i 则表示 E_i 中框架元素对应的框架元素类型. 则问句 S_0 与篇章 D 中任意句子 S_i 之间的框架相似度定义为

$$Sim(S_0,S_i)=\frac{1}{|E_0|}\sum_{e_{0l}\in(E_0-e_{0k}^*),e_{il}\in E_j}Sim(e_{0l},e_{il}) \tag{1}$$

其中，e_{0l} 与 e_{il} 分别表示 S_0 与 S_i 中框架元素类型为 $p_{0l}(p_{0l}\in P_0)$ 与 $p_{il}(p_{il}\in P_i)$ 的框架元素，且有 $p_{0l}=p_{il}$；需要注意问句 S_0 的框架元素集合 E_0 中包含一种特殊的框架元素，即充当疑问语块的框架元素，记为 $e_{0k}^*(e_{0k}^*\in E_0)$，框架元素间的相似度计算详见文献[8]. 因此式(1)在计算问句与篇章句子间的框架相似度时排除了问句中充当疑问语块的框架元素. 框架相似度的意义在于，问句 S_0 与篇章 D 中任意句子 S_i 之间的相似度 $Sim(S_0,S_i)$ 是由两个句子框架元素之间的相似度所决定，框架元素越相似，两个句子越相似.

框架相似性是通过对语义场景之间的各种参与者(框架元素)进行相似度计算综合得到框架间的相似度，从而能够对问题中与背景材料中所涉及到的语义场景进行框架相似度计算来实现答案句检索. 不足之处在于只能获得语义相似的答案句，但并不意味着一定是答案句.

4.1.2　基于框架关系的答案句检索

基于框架关系的答案句获取是在对问题与背景材料进行框架语义分析的基础上，借助 FrameNet 中的框架关系网筛选出与问句 S_0 的框架 f_0^* 直接相连的框架集合 F^{temp}，并进一步剔除掉 F^{temp} 中没有出现在背景材料篇章 D 中的框架，将筛选后剩余框架所对应的句子作为答案句. 框架关系能够通过框架与框架间的相关性，从语义相关度角度从背景材料中获得与问题紧密相关的语义场景，弥补了基于框架相似性方法的不足. 基于框架关系的答案句检索

如算法 1.

算法 1. 基于框架关系的答案句检索.

输入:篇章 $D=\{S_1^D,S_2^D,\cdots,S_N^D\}$;对应的框架 $F^D=\{f_1^D,f_2^D,\cdots,f_{D_n}^D\}$;问句 S_0 对应的框架 f_0^*;框架关系网 $N(F,E,R)$,其中 $F=\{f_1,f_2,\cdots,f_N\}$

输出:答案句集 $S^{answers}$,相应的框架集 $F^{answers}$

1. 初始化 $S^{answers}=\varnothing$,$F^{answers}=\varnothing$,$F^{temp}=\varnothing$;
   //集合 $S^{answers}$ 中存储答案候选句,集合 $F^{answers}$ 存储相应的框架,集合 F^{temp} 作为临时框架存储集合
2. FOR f_i IN F
3. IF $path(f_i,f_0^*)=1$
   //寻找框架关系网中与问句框架 f_0^* 具有直接框架关系的框架 f_i
4. 将 f_i 添加到 F^{temp} 中;
5. END IF
6. END FOR
7. FOR f_k^D IN F^D
8. IF f_k^D IN F^{temp}
   //判断篇章 D 中哪些框架与 f_0^* 具有直接的框架关系
9. 将 f_k^D 添加到 $F^{answers}$ 中;
10. 将 S_k^D 添加到 $S^{answers}$ 中;
11. END IF
12. END FOR

返回 $S^{answers}$,$F^{answers}$.

4.1.3 基于篇章框架关系的答案句检索

基于篇章框架关系的答案句检索,是借助篇章框架语义结构树,利用篇章框架关系来实现.首先基于问句框架定位至篇章框架语义结构树中相应的目标框架,然后从目标框架出发基于篇章框架关系进行搜索,搜索到的局部框架将作为答案候选来源.

给定语篇:

S1 迎春花又名黄素馨,原产我国云南.S2 喜温暖湿润和充足阳光,怕严寒和积水,稍耐阴,较耐旱,以排水良好、肥沃的酸性沙壤土最好.S3 以扦插为主,也可用压条、分株繁殖.S4 扦插,春、夏、秋三季均可进行,剪取半木质化的枝条 15 厘米长,插入沙土中,保持湿润,约 15 天生根.S5 压条,将较长的枝条浅埋于沙土中,不必刻伤,40～50 天后生根,翌年春季与母株分离移栽.S6 分株,可在春季芽萌动时进行.S7 春季移植时地上枝干截除一部分,需带宿土.S8 在生长过程中,注意土壤不能积水和过分干旱,开花前后适当施肥 2～3 次.S9 秋、冬季应修剪整形,保持株新花多.S10 病虫害常发生叶斑病和枯枝病,可用退菌特可湿性粉剂倍液喷洒.S11 虫害有蚜虫和大蓑蛾危害,用辛硫磷乳油倍液喷杀.S12 迎春枝条长而柔弱,下垂或攀援,碧叶黄花,可于堤岸、台地和阶前边缘栽植,特别适用于宾馆、大厦顶棚布置,也可盆栽观赏.

建立其篇章框架语义结构树,如图 5.

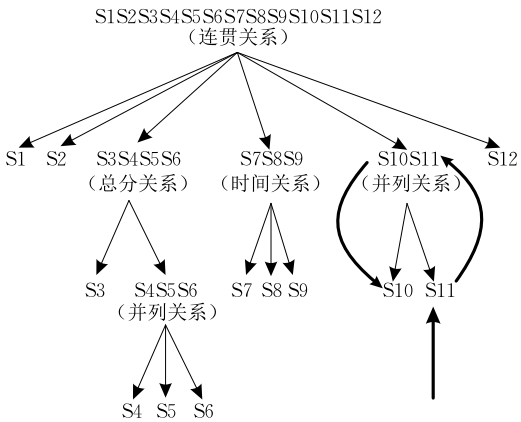


图 5 基于篇章框架关系的答案句检索

图 5 中,基于框架相似度或框架关系将答案句首先定位于篇章框架语义结构树的目标框架单元 S11,然后利用篇章框架关系如"并列关系",通过基于树的搜索获得答案的答案句检索.篇章框架关系则是从篇章语义逻辑连贯角度,通过篇章级关系获得与问题语义逻辑相关的答案句,不足之处在于容易带来干扰句.基于篇章框架关系的答案句检索如算法 2.

算法 2. 基于篇章框架关系的答案句检索.

输入:篇章 $D=\{S_1^D,S_2^D,\cdots,S_N^D\}$;对应的框架 $F^D=\{f_1^D,f_2^D,\cdots,f_{D_n}^D\}$;问句 S_0 对应的框架 f_0^*;篇章 D 对应的篇章结构树 $T^D(F^D,R^D)$,其中 $R^D=(r_1^D,r_2^D,\cdots,r_M^D)$ 是篇章框架关系集

输出:答案句集 $S^{answers}$,相应的框架集 $F^{answers}$

1. 利用框架相似度方法初始化问句 S_0 对应的种子答案句子集 $S^{answers}$ 和种子框架集 $F^{answers}$,集合 S^{temp} 作为与句子集 $S^{answers}$ 有篇章框架关系存在的临时句子存储集.
2. FOR $S_i^{answers}$ IN answers
3. IF $S_i^{answers}.siblings()$
   //寻找在篇章框架语义结构树中与答案句 $S_i^{answers}$ 具有直接篇章框架关系的兄弟节点句子
4. 将 $S_i^{answers}$ 的兄弟节点句子 $S_i^{answers}.getSiblings()$ 添加到 S^{temp} 中;
5. END IF
6. END FOR
7. FOR $S_i^{answers}$ IN S^{temp}
8. 将 $S_i^{answers}$ 添加到 $S^{answers}$ 中
9. 将 $F_i^{answers}$ 的添加到 $F^{answers}$ 中
10. END FOR

返回 $S^{answers}$,$F^{answers}$.

4.2 答案抽取

答案抽取阶段是在获得答案句的基础上进一步抽取答案句中能够充当答案的语块. 框架语义分析的特点在于它能够提供语块级的语义分析结果. 本文提出了基于篇章级框架语义分析，利用框架相似性、篇章有定零形式线索和框架关系等 3 种答案抽取方法.

4.2.1 基于框架相似性的答案抽取

基于框架相似性的答案抽取同样存在一个假设前提：问句与答案句之间框架相似性越大，答案句中存在答案的可能性越大，进而可以从最相似的答案句中得到最有可能作为问题答案的语块. 答案语块抽取任务可以转化为从答案句中抽取出最有可能充当问句中疑问角色的框架元素 e_{0k}^{*}.

首先，依据框架相似度高低，在答案句集 S^{answers} 中获得最相似的答案句：

$$S_i^{*}=\arg\max_{S_i\in S^{\text{answers}}} Sim(S_0,S_i)=$$

$$\arg\max_{S_i\in S^{\text{answers}}}\frac{1}{|E_0|}\sum_{e_{0l}\in(E_0-e_{0k}^{*}),e_{jl}\in E_j} Sim(e_{0l},e_{jl}) \quad (2)$$

然后，答案句 S_i^{*} 中同 e_{0k}^{*} 的框架元素类型相对应的框架元素 e_{ik}^{*} 作为抽取的答案结果.

图 6 中，与问句 S_0 最相似的答案句为 S_1，问句 S_0 中充当疑问角色的框架元素为“什么时代”，相应的框架元素类型为“time”，答案句 S_1 中与框架元素类型“time”对应的框架元素为“东汉”，即作为问题的答案.

基于框架相似性的答案抽取是指利用框架中框架元素之间的相似性，能够从相似性角度为问题提供语块级的答案抽取结果. 不足之处在于：(1) 无法处理相关的语义成分；(2) 当答案句中充当答案的框架元素缺失，需要从上下文中抽取时，无法获得答案.

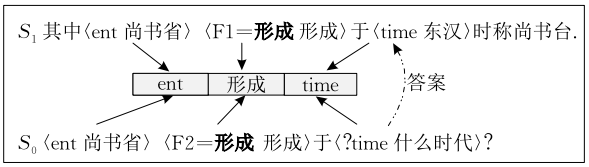


图 6　基于框架相似性的答案抽取

4.2.2 基于篇章有定零形式线索的答案抽取

零形式是语言中存在的一种普遍现象，零形式出现时，由于语义成分缺失，基于框架相似性的方法无法直接进行答案抽取. 有定零形式线索则能够在答案抽取阶段处理充当答案的框架元素缺失问题. 基于篇章有定零形式线索的答案抽取就是通过在篇章范围找到句子中充当答案的缺失语义成分. 在 4.2.1 节基于框架相似性的答案抽取基础上，若获得的框架元素为 $e_{ik}^{*}=$“DNI”，则判定为有定零形式，若有定零形式 e_{ik}^{*} 被篇章 $D=\{S_1,S_2,\cdots,S_N\}$ 的第 j 个句子 S_j 中的框架元素 e_{jk_j} 填充，则基于篇章有定零形式线索的答案抽取结果为 e_{jk_j}.

图 7 中，与问句 S_0 最相似的答案句为 S_2，两句共同具有的框架为“适宜性”，而在 S_2 中作为答案抽取对象的框架元素“Evaluee”缺失了，记为 DNI，则需要利用篇章范围的框架元素进行填充. 在此例中，DNI 的填充成分来源为句 S_1 中的框架元素“天葬卫星”，也作为答案的抽取结果.

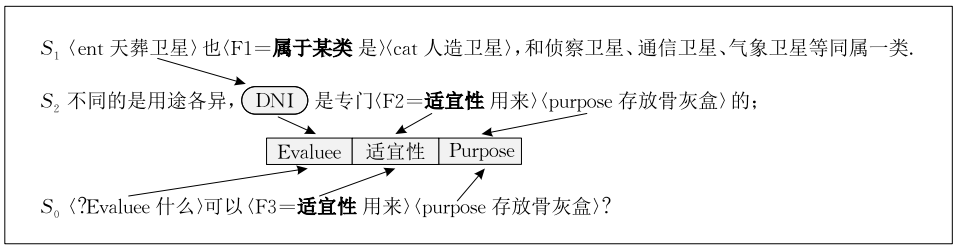


图 7　基于有定零形式线索的答案抽取

4.2.3 基于框架关系的答案抽取

基于框架关系进行答案抽取的思路不同于基于框架相似性的基本假设，问题答案不只是来源于语义相似性较高的答案句，也可能来源于相关度高的答案句. 在框架关系网中直接相连的两个框架的框架元素之间具有高度相关性，如“3.1 节的框架关系”介绍的框架元素间映射关系，这种映射可以为本文答案抽取提供语块级的相关性信息. 基于框架关系的答案抽取建立在 4.1.2 节基于框架关系的答案句检索基础上.

图 8 问句 S_0 的目标词“帮助”激起框架“协助”，语篇中句子 S_i 的目标词“进行”激起框架“有意行为”，在 FrameNet 的框架关系网中框架“帮助”与“有意行为”具有“使用”框架关系，而具有“使用”关系的框架之间存在特定的框架元素映射. 依据这种映射关系，图 8 中句 S_i 的框架元素类型“dom(领域)”对应的框架元素“在矿业、能源制造业、电信、农业等领域”即为问题 S_0 的答案映射.

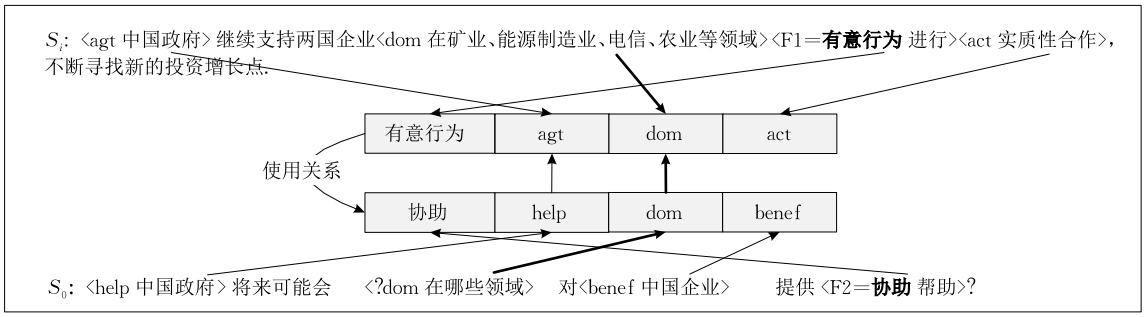


图 8 基于框架关系的答案抽取

框架关系能够在答案抽取阶段弥补框架相似性的不足，能够通过框架关系建立起框架元素之间的相关性，从而获得相关度高的答案.

5 实验及结果分析

5.1 实验数据和相关工具

实验所用的阅读理解问答语料是山西大学 CFN 研究室构建的阅读语篇与问题集，共包含问题 552 条，阅读理解语篇 167 篇，涉及地理、环保、历史等 15 个领域，语料中问题的答案都进行了人工标注与校对. 依据不同的实验要求，语料进行了分词、词性标注、目标词识别、框架选择、框架元素标注的预处理，并对篇章级的有定零形式线索、篇章框架关系进行了标注. 实验中用到的框架关系来源于 FrameNet 项目组[25]提供的框架关系资源，词性标注工具使用了中国科学院计算技术研究所的 ICTCLAS，还使用了知网[28]平台提供的词汇语义相似度计算工具. 由于目前汉语框架语义自动分析性能还偏低[29-30]，因此在进行目标词识别、框架选择、框架元素标注、有定零形式线索标注及篇章框架关系标注时进行了人工参标注与矫正.

实验中使用 Top-k 的准确率 P 来评价答案句检索和答案抽取结果：

(1) 答案句检索的 Top-k 准确率：

$$P_1=\frac{CorrectSentence(k)}{k}\times 100\%.$$

$CorrectSentence(k)$ 表示针对所有 552 个问题的前 k 个答案句中正确的答案句数目.

(2) 答案抽取的 Top-k 准确率：

$$P_2=\frac{CorrectAnswer(k)}{k}\times 100\%.$$

$CorrectAnswer(k)$ 表示针对所有 552 个问题的前 k 个答案抽取结果中正确的答案数目.

5.2 实验设置

为了验证本文基于篇章框架语义分析汉语阅读理解问答方法的有效性，实验设置了答案句检索和答案抽取两个阶段.

答案句检索阶段，实验中设置基于框架相似度的答案句检索作为基线 1 方法(Baseline1)，并与分别加入框架关系(Baseline1＋FR)、篇章框架关系(Baseline1＋DFR)及全部框架语义(Baseline1＋ALL1)的答案句检索结果进行比较，以验证篇章级框架语义特征对答案句检索的有效性. 进一步，为了验证本文框架语义在答案句检索时的优势，实验设置了相关文献中较为常用的比较方法，包括：

(1) 基于知网语义(HowNet Semantic，HNS)的句子相似度计算

对于问题 S_0 与句子 S_1 中的任意词 A_i 与 B_j，基于 HowNet 计算其相似度 $S(A_i,B_j)$，取 $a_i=\max\{S(A_i,B_1),S(A_i,B_2),\cdots,S(A_i,B_n)\}$，$b_i=\max\{S(B_i,A_1),S(B_i,A_2),\cdots,S(B_i,A_n)\}$，则目标句子 S_0 与 S_1 之间的相似度为

$$Sim(S_0,S_1)=\left(\frac{\sum_{i=1}^{m}a_i}{m}+\frac{\sum_{i=1}^{n}b_i}{n}\right)\Big/2 \qquad (3)$$

(2) 基于向量空间模型[31](Vector Space Model，VSM)的句子相似度计算

对于问题 $\boldsymbol{S}_0$ 与句子 $\boldsymbol{S}_1$，计算其中词的 TF-IDF 值，得到句子对应的向量分别为 $\boldsymbol{S}_0=(w_1,w_2,\cdots,w_n)$ 和 $\boldsymbol{S}_1=(w'_1,w'_2,\cdots,w'_n)$，句子间的相似度利用两个向量之间的夹角余弦值来表示：

$$Sim(\boldsymbol{S}_0,\boldsymbol{S}_1)=\frac{\sum_{i=1}^{n}(w_i\times w'_i)}{\sqrt{\sum_{i=1}^{n}w_i^2\times\sum_{i=1}^{n}w_i'^2}} \qquad (4)$$

(3) 基于词袋模型[31-33](Bag of Word，BOW)的句子相似度计算

对于问题 $\boldsymbol{S}_0$ 与句子 $\boldsymbol{S}_1$，构建一个词表，分别计算问题和句子中关键词的出现次数，得到对应的向量分别为 $\boldsymbol{S}_0=(f_1,f_2,\cdots,f_n)$ 和 $\boldsymbol{S}_1=(f'_1,f'_2,\cdots,f'_n)$，句子间的相似度利用两个向量之间的夹角余弦值来表示：

$$Sim(\boldsymbol{S}_0,\boldsymbol{S}_1)=\frac{\sum_{i=1}^{n}(f_i\times f'_i)}{\sqrt{\sum_{i=1}^{n}f_i^2\times\sum_{i=1}^{n}f'^2_i}} \tag{5}$$

(4) 融合方法

对于问题 S_0 与句子 S_1，设利用本文框架语义方法获得相似度为 $Sim_1(S_0,S_1)$，用其它方法获得相似度为 $Sim_2(S_0,S_1)$，那么融合的结果如下：

$$Sim(S_0,S_1)=\alpha\times\frac{Sim_1(S_0,S_1)}{\sum_{1\le i\le n}Sim_1(S_0,S_i)}+(1-\alpha)\times\frac{Sim_2(S_0,S_1)}{\sum_{1\le j\le n}Sim_1(S_0,S_j)} \tag{6}$$

其中，$0<\alpha<1$.

在答案抽取阶段，实验中设置基于框架相似度的答案抽取作为基线 2(Baseline2)，分别对加入框架关系(Baseline2＋FR)、有定零形式线索(Baseline2＋DNI)及其融合(Baseline2＋ALL2)的方法进行比较，用来验证框架语义对答案抽取的有效性，其中 ALL2＝(FR＋DNI).

5.3　实验结果及分析

5.3.1　答案句检索

(1) 不同篇章级框架语义特征加入对答案句检索的影响如表 1.

表 1　不同篇章级框架语义特征加入时的答案句检索结果比较　(单位：%)

实验方法	Top-1	Top-2	Top-3	Top-4	Top-5
Baseline1	39.49	50.18	52.54	52.72	52.72
Baseline1＋FR	**44.75**	**54.35**	**60.69**	**65.40**	**69.75**
Baseline1＋DFR	**43.12**	46.74	47.64	47.83	47.83
Baseline1＋ALL1	43.84	50.18	52.54	53.99	**55.43**

表 1 中，本文基于框架相似度的答案句检索 Bseline1 方法，获得了(Top-4，Top-5)52.7%的抽取结果，它仅是基于框架相似度的答案句检索结果. 基于“Baseline1＋FR”的结果从 Top-1～5 都要高于 Baseline1，且获得了最高 69.75%的结果，这是因为框架关系能够考虑框架语义的相关性，从而抽取出与问题相关的答案句，能够在 Baseline1 的基础上提高答案句的抽取结果. “Baseline1＋DFR”在 Top-1 高于基于 VSM、HNS、BOW 相似度的结果，而在 Top-2～5 反而不如这些比较方法，这是因为篇章框架关系揭示了篇章中上下文语义的相关性，其优势在于能够抽取出与问题相关度很高但不一定具有较高相似度的句子，而在实验语料中，大多问题都属于需要通过语义相似来获取答案的. 因此，利用这种篇章框架关系在 Top-2～5 检索出了更多干扰句，最终影响实验结果，但 Top-1 的结果要高于基于 VSM、HNS、BOW 相似度的方法，这说明利用篇章框架关系能够找出无法用相似度检索到的答案句. 并不是所有的关系类型都能够用来抽取答案句，如因果关系可以解决 why-型的问题，属于关系可以解决意图和意图所有者之间的关系；条件关系和假设关系可以解决一些答案发生的前提等一些问题；并列关系、承接关系、递进关系、选择关系、解说关系和转折关系可以用于解决一个问题存在多个答案的情况. 本文问答语料中答案大多数只源于一个句子，因此许多篇章框架关系在本实验中较少用到. 例如，运用“因果关系”进行答案句检索实例：

给定语篇：

你只要打开中学语文课本，就会发现：几册课本的内容，绝大多数是历代的散文；所选的小说片段，也具备散文的特质. 于是，“散文是文学的正宗”，就有了不容置疑的佐证.

问题：

为何说“散文是文学的正宗”?

经相似度计算，“于是，“散文是文学的正宗”，就有了不容置疑的佐证.”一句作为答案句检索出来，而正确答案是跟这句具有因果关系的“你只要打开中学语文课本，就会发现：几册课本的内容，绝大多数是历代的散文；所选的小说片段，也具备散文的特质.”

表 1 中“Baseline1＋All1”方法也并没有达到最好，其原因与“Baseline1＋DFR”相同.

(2) 基于篇章框架语义分析的答案句检索(Baseline1＋FR)、传统基于相似度方法及其融合方法的比较如表 2.

表 2　“Baseline1＋FR”、传统基于相似性及其融合方法的答案句检索结果比较　(单位：%)

实验方法	Top-1	Top-2	Top-3	Top-4	Top-5
VSM	18.15	39.38	51.38	58.62	63.08
HNS	42.92	53.85	59.54	62.46	64.62
BOW	34.05	47.77	55.16	58.71	62.56
Baseline1＋FR	**44.75**	**54.35**	**60.69**	**65.40**	**69.75**
(Baseline1＋FR)＋VSM	38.76	53.93	**62.17**	**65.92**	**70.04**
(Baseline1＋FR)＋HNS	35.39	**55.62**	**65.92**	**71.16**	**75.66**
(Baseline1＋FR)＋BOW	38.95	**56.74**	**65.36**	**70.60**	**76.40**

表 2 显示，在答案句检索时，“Baseline1＋FR”的方法相比传统基于向量空间模型“VSM”、基于知网词汇语义“HNS”及基于词袋模型“BOW”的相似度方法能够获得最好的答案句检索结果，说明框架语义要优于一般的相似度方法. 在融合后的结果中（我们选取实验结果最好的参数 α，其中融合方法（Baseline1＋FR）＋VSM 中 $\alpha=0.5$；（Baseline1＋FR）＋HNS 中 $\alpha=0.8$；（Baseline1＋FR）＋BOW 中 $\alpha=0.7$），从 TOP-3～5 都要高出单独基于 VSM、BOW、HNS 相似度方法的结果，这也验证了加入框架关系对于答案句检索是有效的；而在 Top-1 中，3 种融合结果都要低于“Baseline1＋FR”的结果，这说明当在准确定位问题的唯一答案句时，融合 VSM、BOW、HNS 相似度会带来干扰句. 例如：

给定语篇：

蒙古族的祖先是室韦族，最早出现在五世纪的北魏文献记载中. 他们早先居住在额尔古纳河一带，七世纪时西迁到鄂嫩河——石勒喀河和克鲁伦河流域，活动范围大至整个蒙西高原，由许多部落组成. “蒙古”是一个部落的名称，1206 年由铁木真统一了各部落建立了蒙古国，加快了血缘关系向地缘关系的转化，逐渐融合为一个新民族共同体，“蒙古”也由原来的一个部落名称变为整个民族的名称.

问题：

蒙古族的祖先最早栖身在哪儿？

基于 HNS、VSM 及 BOW 相似度的 Top-1 答案句检索结果：

蒙古族的祖先是室韦族，最早出现在五世纪的北魏文献记载中.

基于框架相似性与框架关系的 Top-1 答案句检索结果：

他们早先居住在额尔古纳河一带，七世纪时西迁到鄂嫩河——石勒喀河和克鲁伦河流域，活动范围大至整个蒙西高原，由许多部落组成.

基于“Baseline1＋FR”的答案抽取结果：

额尔古纳河一带

例中问题“蒙古族的祖先最早栖身在哪儿？”的主要关键词为“蒙古族、祖先、最早、栖身”，基于 HNS、VSM、BOW 相似度方法检索到的 Top-1 答案句是“蒙古族的祖先是室韦族，最早出现在五世纪的北魏文献记载中.”，它成为正确答案句“他们早先居住在额尔古纳河一带，七世纪时西迁到鄂嫩河……”的干扰句. 相比之下，本文方法能够将问题中被关键词“栖身”激起的框架“居住/Residense”直接定位至背景材料中关键词“居住”激起的框架“居住/Residense”，从而准确检索到答案句. 进一步再在答案抽取阶段，通过抽取问题框架“居住/Residense”中承担问题答案的框架元素“额尔古纳河一带”，作为最终答案.

5.3.2 答案抽取

（1）不同篇章级框架语义特征对最终答案语块抽取的影响如表 3.

表 3 Baseline2 加入不同篇章级框架语义特征时的答案抽取结果比较 （单位：%）

实验方法	Top-1
Baseline2	37.14
Baseline2＋FR	39.74
Baseline2＋DNI	39.56
Baseline2＋ALL2	42.03

答案抽取阶段，基于框架元素语义相似度的 Baseline2 方法获得了 37.14%的答案抽取结果. 相比 Baseline2，基于“Baseline2＋FR”的答案抽取结果有一定提升，这是因为框架关系能够提供框架间框架元素的映射，使得获取答案语块的途径不仅从框架相似度角度，而且能够从框架间的语义相关度来获取. 基于“Baseline2＋DNI”的答案抽取方法则是将答案的抽取从基于相似度的句子级语块范围扩展至了篇章级语块范围，可以看到此方法也在 Baseline2 的基础上有所提升. 最终将 Baseline2 与框架关系、有定零形式线索融合，在“Baseline2＋ALL2”的方法中获得了最高 42.03%的答案抽取结果.

（2）答案句检索对答案抽取的影响如图 9.

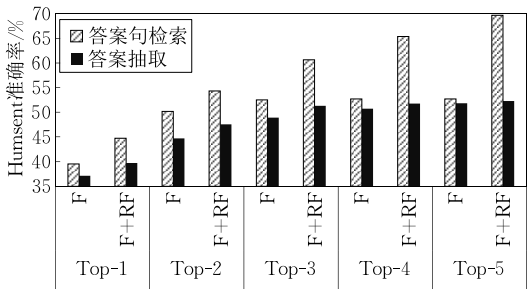


图 9 答案句检索对答案抽取的影响

图 9 中，F 对应的斜线柱体代表基于框架相似性的答案句检索结果，黑色柱体代表基于框架相似性的答案抽取结果. F＋FR 对应的斜线柱体则代表框架相似性与框架关系融合的答案句检索结果，黑色柱体代表答案抽取结果. 可以看出，无论是单纯基于框架相似性（F）的实验还是基于“框架＋框架关系”（F＋FR）的实验，都是随着答案句检索结果提高，答案抽取结果也随之提升.

6 结论与展望

针对汉语阅读理解问答的答案句检索和答案抽取关键技术，本文提出了一种新的基于汉语篇章框架语义分析的答案句检索与答案抽取方法. 在答案句检索阶段，基于框架语义相性、框架关系及篇章框架关系对答案句进行检索；在答案抽取阶段，利用框架元素相似性、框架关系及有定零形式线索进行答案抽取. 实验表明：本文基于篇章级框架语义分析的阅读理解问答方法能够从相似性与相关性两方面获得更好的答案句检索与答案抽取结果. 相比 FrameNet，CFN 的框架语义资源覆盖率低，使得该方法目前只能处理 CFN 框架资源库中有限范围的问答数据.

下一步，将在资源构建方面不断提高现有 CFN 框架语义资源的覆盖率，使基于篇章框架语义分析的阅读理解问答技术向更多领域拓展.

致　谢　本文实验用到知网平台提供的词汇语义相似度计算工具；中国科学院计算技术研究所的 ICTCLAS 分词工具；FrameNet 项目组的框架关系网资源. 在此表示感谢！

参　考　文　献

[1] Hirschman L, Light M, Breck E, et al. Deep read: A reading comprehension system//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Maryland, USA, 1999: 325-332

[2] Zheng Shi-Fu, Liu Ting, Qin Bing, Li Sheng. Overview of question answering. Journal of Chinese Information Processing, 2002, 16(6): 46-52 (in Chinese)
(郑实福，刘挺，秦兵，李生. 自动问答综述. 中文信息学报，2002，16(6)：46-52)

[3] Zhang Yong-Kui, Zhao Zhe-Qian, Bai Li-Jun, Chen Xin-Qing. Internet based Chinese question-answering system. Computer Engineering, 2003, 29(15): 84-86(in Chinese)
(张永奎，赵辄谦，白丽君，陈鑫卿. 基于互联网的中文问答系统. 计算机工程，2003，29(15)：84-86)

[4] Cui Huan, Cai Dong-Feng, Miao Xue-Lei. Research on Web based Chinese question answering system and answer extraction. Journal of Chinese Information Processing, 2004, 18(3): 24-31(in Chinese)
(崔桓，蔡东风，苗雪雷. 基于网络的中文问答系统及信息抽取算法研究. 中文信息学报，2004，18(3)：24-31)

[5] Yu Zheng-Tao, Fan Xiao-Zhong, Guo Jian-Yi, Geng Zeng-Min. Answer extracting for Chinese question answering system based on latent semantic analysis. Chinese Journal of Computers, 2006, 29(10): 1889-1893(in Chinese)
(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)

[6] Wu You-Zheng, Zhao Jun, Xu Bo. Sentence retrieval with a topic-based language model. Journal of Computer Research and Development, 2007, 44(2): 288-295(in Chinese)
(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)

[7] Li Liang-Fu, Fan Xiao-Zhong, Li Hong-Qiao. Domain specific QA driven by computation of semantic similarity. Journal of Beijing Institute of Technology, 2005, 50(11): 21-25(in Chinese)
(李良富，樊孝忠，李宏乔. 语义相似计算驱动领域自动问答. 北京理工大学学报，2005，50(11)：21-25)

[8] Li Ru, Wang Zhi-Qiang, Li Shuang-Hong, et al. Chinese sentence similarity computing based on frame semantic parsing. Journal of Computer Research and Development, 2013, 50(8): 1728-1736(in Chinese)
(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. 计算机研究与发展，2013，50(8)：1728-1736)

[9] Tian Jiu-Le, Zhao Wei. Words similarity based on tongyici cilin in semantic web adaptive learning system. Journal of Jilin University, 2010, 28(6): 602-608(in Chinese)
(田久乐，赵蔚. 基于同义词词林的词语相似度计算方法. 吉林大学学报，2010，28(6)：602-608)

[10] Sun Ang, Jiang Ming-Hu, Ma Yan-Jun. An instance-based approach for pinpointing answers in Chinese question answering //Proceedings of the 8th International Conference on Signal Processing. Beijing, China, 2006: 1620-1623

[11] Higashinaka R, Isozaki H. Corpus-based question answering for why-questions//Proceedings of the 3th International Joint Conference on Natural Language Processing. Hyderabad, India, 2008: 418-425

[12] Zhang Zhi-Chang, Zhang Yu, Liu Ting, Li Sheng. Why-questions answering for reading comprehension based on topic and rhetorical identification. Journal of Computer Research and Development, 2011, 54(2): 216-223(in Chinese)
(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)

[13] Verberne S, Boves L. Evaluating discourse-based answer extraction for why-question answering//Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, Netherlands, 2007: 735-736

[14] Santhosh S, Ali S. Discourse based advancement on question answering system. International Journal on Soft Computing, Artificial Intelligence and Applications, 2012, 1(2): 1-12

[15] Du Yong-Ping, Huang Xuan-Jing, Wu Li-De. Using pattern and linguistic features to improve reading comprehension performance. Journal of Computer Research and Development, 2008, 51(2): 293-299(in Chinese)
(杜永萍，黄萱菁，吴立德. 利用模式及语言学特征提高阅读理解性能. 计算机研究与发展，2008，51(2)：293-299)

[16] Liang Zheng-Ping, Ji Zhen, Liu Xiao-Li. Research on semantic pattern based question answering system. Journal of Shenzhen University (Science and Engineering), 2007, 24(3): 281-285 (in Chinese)
(梁正平, 纪震, 刘小丽. 基于语义模板的问答系统研究. 深圳大学学报(理工版), 2007, 24(3): 281-285)

[17] Yu Zheng-Tao, Mao Cun-Li, Deng Jin-Hui, et al. Answer extraction scheme for Chinese question answering system based on pattern learning. Journal of Jilin University (Engineering and Technology Edition), 2008, 52(1): 142-147(in Chinese)
(余正涛, 毛存礼, 邓锦辉等. 基于模式学习的中文问答系统答案抽取方法. 吉林大学学报(工学版), 2008, 52(1): 142-147)

[18] Ravichandran D, Hovy E. Learning surface text patterns for a question answering system//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA, 2002: 41-47

[19] Hu Bao-Shun, Wang Da-Ling, Yu Ge, Ma Ting. An answer extraction algorithm based on syntax structure feature parsing and classification. Chinese Journal of Computers, 2008, 31(4): 662-676(in Chinese)
(胡宝顺, 王大玲, 于戈, 马婷. 基于句法结构特征分析及分类技术的答案提取算法. 计算机学报, 2008, 31(4): 662-676)

[20] Wu You-Zheng, Zhao Jun, Xu Bo. Unsupervised answer pattern acquisition. Journal of Chinese Information Processing, 2007, 21(2): 69-76(in Chinese)
(吴友政, 赵军, 徐波. 基于无监督学习的问答模式抽取技术. 中文信息学报, 2007, 21(2): 69-76)

[21] Hao Xiao-Yan, Li Ru, Liu Kai-Ying. Description systems of the Chinese FrameNet database and software tools. Journal of Chinese Information Processing, 2007, 21(5): 96-100, 138 (in Chinese)
(郝晓燕, 李茹, 刘开瑛. 汉语框架语义知识库及软件描述体系. 中文信息学报, 2007, 21(5): 96-100, 138)

[22] Li Ru. Research on Frame Semantic Structure Analysis Technology for Chinese Sentences [Ph. D. dissertation]. Shanxi University, Taiyuan, 2012(in Chinese)
(李茹. 汉语句子框架语义结构分析技术研究[博士学位论文]. 山西大学, 太原, 2012)

[23] Das D, Chen D, Martins A, et al. Frame-semantic parsing. Computational Linguistics, 2014, 40(1): 9-56

[24] Ruppenhofer J, Sporleder C, Morante R, et al. SemEval-2010 task 10: Linking events and their participants in discourse//Proceedings of the 5th International Workshop on Semantic Evaluation. Stroudsburg, USA, 2010: 45-50

[25] Baker C, Fillmore C, Lowe J. The Berkeley FrameNet project//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Montreal, Canada, 1998: 86-90

[26] Wang Ning, Li Ru, Lei Zhang-Zhang, et al. Document oriented gap filling of definite null instantiation in FrameNet. Lecture Notes in Artificial Intelligence Volume, 2013, 8202: 85-96

[27] Huang Bo-Rong, Liao Xu-Dong. Modern Chinese Language. Beijing: Higher Education Press, 2011(in Chinese)
(黄伯荣, 廖序东. 现代汉语. 北京: 高等教育出版社, 2011)

[28] Liu Qun, Li Su-Jian. Word similarity computing based on HowNet. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76(in Chinese)
(刘群, 李素建. 基于《知网》的词汇语义相似度. 中文计算语言学, 2002, 7(2): 59-76)

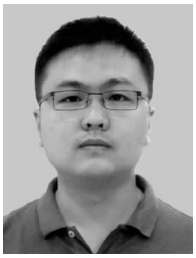
[29] Li Ru, Liu Hai-Jing, Li Shuang-Hong. Chinese frame identification using T-CRF model//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China, 2010: 674-682

[30] Li Ji-Hong, Wang Rui-Bo, Wang Wei-Lin, Li Guo-Chen. Automatic labeling of semantic roles on Chinese FrameNet. Journal of Software, 2010, 21(4): 597-611(in Chinese)
(李济洪, 王瑞波, 王蔚林, 李国臣. 汉语框架语义角色自动标注. 软件学报, 2010, 21(4): 597-611)

[31] Zhang Zhi-Hui. Answer Extraction Based on Logic Representation and Reasoning for Reading Comprehension [M. S. dissertation]. Harbin Institute of Technology, Harbin, 2008 (in Chinese)
(张志辉. 基于逻辑表示与推理的阅读理解答案抽取[硕士学位论文]. 哈尔滨工业大学, 哈尔滨, 2008)

[32] Du Yong-Ping. Research on Key Techniques of Pattern Knowledge Based Question Answering [Ph. D. dissertation]. Fudan University, Shanghai, 2005(in Chinese)
(杜永萍. 基于模式知识库的问题回答关键技术研究[博士学位论文]. 复旦大学, 上海, 2005)

[33] Zhang Zhi-Chang, Zhang Yu, Liu Ting, et al. Answer sentence extractionof reading comprehension based on shallow semantic tree kernel. Journal of Chinese Information Processing, 2008, 22(1): 80-86(in Chinese)
(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)



WANG Zhi-Qiang, born in 1987, Ph. D. candidate. His main research interests include Chinese information processing and social media data mining.

LI Ru, born in 1963, professor, Ph. D. supervisor. Her research interests include Chinese information processing and information retrieval.

LIANG Ji-Ye, born in 1962, professor, Ph. D. supervisor. His main research interests include granular computing, data mining, and machine learning.

ZHANG Xu-Hua, born in 1990, M. S. candidate. His

main research interest is Chinese information processing.

WU Juan, born in 1991, M. S. candidate. Her main research interest is Chinese information processing.

SU Na, born in 1989, M. S. candidate. Her main research interest is Chinese information processing.

Background

This paper proposed a method for Answer-sentence Retrieval and Answer Extraction in Chinese Reading Comprehension (RC), based on Discourse-level Frame Semantic Parsing. Generally speaking, RC is one of the most important and challenging problems in the fields of Natural Language Processing (NLP) and Question Answering (QA). It is very hard to achieve a higher intelligence-level without the technology of deep semantic parsing. With the progress of NLP technology, research on Answer-sentence Retrieval and Answer Extraction for RC has gained some achievements, but it is still far from a real intelligence.

This paper provided a discourse-level frame semantic parsing method to solve the Answer-sentence Retrieval and Answer Extraction for the Chinese RC, and we put forward a new approach to Chinese discourse-level semantic parsing based on the theory of Frame Semantic. As far as we know, this is the first time we applied Frame Semantic theory to discourse semantic parsing.

Our work belongs to "Automatic Extraction and Semantic Reasoning of Chinese Discourse Frame Semantic Relations Network" of National Natural Science Foundation of China (No. 61373082). This project focuses on the deep semantic parsing from the perspective of frame semantic, aiming to develop a new way for the Chinese semantic parsing, and establish a frame semantic resource database to support the frame semantic parsing of Chinese and further provide a new semantic parsing solution for QA.

Many researchers have made great efforts in the study of RC, such as Answer-sentence Retrieval based on similarity of sentences, machine learning and discourse parsing, and Answer Extraction based on pattern and machine learning. Previous research has laid a good foundation for RC technology.

This research mainly explored discourse frame semantic parsing and its application in RC. This work is funded by the National Natural Science Foundation of China (Nos. 61373082, 61432011, and U1435212) and the National High Technology Research and Development Program (863 Program) (No. 2015AA015407). My thanks also goes to the support of Shanxi Platform Project (2014091004-0103) and Scholarship Council (2013-015), and Open Project Foundation of Information Security Evaluation Center of Civil Aviation, Civil Aviation University of China (No. CAAC-ISECCA-201402).