

文章编号: 1003-0077 (2011) 00-0000-00

面向阅读理解复杂问题的句子融合*

谭红叶^{1,2}, 赵红红¹, 李茹^{1,2}

(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: 阅读理解的研究是目前 NLP 领域的一个研究热点。阅读理解中好的复杂问题解答策略不仅要进行答案句的抽取, 还要对答案句进行融合、生成相应的答案, 但是目前的研究大多集中在前者。本文针对复杂问题解答中的句子融合进行研究, 提出了一种兼顾句子重要信息、问题关联度与句子流畅度的句子融合方法。该方法的主要思想为: 首先, 基于句子拆分和词重要度选择待融合部分; 然后, 基于词对齐进行句子相同信息的合并; 最后, 利用基于依存关系、二元语言模型及词重要度的整数线性规划优化生成句子。在历年高考阅读理解数据集上的测试结果表明, 该方法取得了 82.62% 的 F 值, 同时更好地保证了结果的可读性及信息量。

关键字: 阅读理解; 复杂问题; 句子融合; 文本生成

中图分类号: TP391

文献标识码: A

Sentence Fusion for Complex Problems in Reading Comprehension

TAN Hongye^{1,2}, ZHAO Honghong¹, Li Ru^{1,2}

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

Abstract: Reading comprehension system is a research focus in natural language processing. In these systems, both answer extraction and sentence fusion are necessary for answering complex problems. But current researches only concern the former. This paper studies the techniques of sentence fusion for complex problems, and presents a method which gives consideration to sentence importance, relevancy to queries and sentence readability. Its main idea is as follows: At first, the parts, which will be fused, are chosen based on sentence separation and word salient. Then, the repeated contents are merged by word alignments. Finally, the sentences are generated based on the integer linear optimization, which utilizes dependency relations, the language model and word salient. The experiments on reading comprehension datasets in college entrance examinations are: the F-measure is 82.62%. And the results show that this method better keeps the readability and informativity.

Key words: reading comprehension; complex problems; sentence fusion; text generation

1 引言

随着 1999 年文本检索会议 (TREC)^[1] 自动问答评测的开展, 自动问答及阅读理解技术的研究成为自然语言处理领域的热点, 也成为判别计算机理解人类语言的一个标准。从现有研究看, 问答系统和阅读理解涉及的问题可以分为事实型问题、列表型问题、定义型问题和复杂问题^[2]。

目前阅读理解方面的研究大多针对简单文本和简单问题, 如微软建立了一套面向儿童的开放域阅读理解数据集 MCTest 进行相关研究^[3]; Facebook 的 bAbI 项目围绕仿真生成的 20 个基本文本理解和推理任务进行研究^[4]。近几年, 随着日本国立情报学研究所 “Todai 机器

* 收稿日期: 定稿日期:

基金项目: 国家高技术研究发展计划 (863 计划) 项目 (2015AA015407); 国家自然科学基金资助项目 (61100138, 61403238); 山西省自然科学基金资助项目 (2011011016-2, 2012021012-1); 山西省回国留学人员科研项目 (2013-022); 山西省高校科技开发项目 (20121117); 山西省 2012 年度留学回国人员科技活动择优项目。

人”项目¹的推出，面向真实世界复杂问题的阅读理解研究成为热点，如 NTCIR 在 2013 年推出现实世界复杂问答的测评²；CLEF 继 QA4MRE 之后开展了高考阅读理解答题评测；2015 年我国也开展了“基于大数据的类人智能关键技术与系统”项目的研究³。这些测评及项目中涉及的复杂问题包含大量主观题，其特点是需要深层理解文本、综合各类信息才能正确回答，且答案可能不止一个版本。

目前复杂问题解答的研究大多集中在答案句的抽取^[5]，但是直接将抽取的答案句作为答案，会导致答案中存在冗余或与问题不相关的信息。因此，如果想要准确地回答问题，还必须对候选答案句进行句子融合、重新生成答案文本，如表 1 所示的问题解答示例。因此，在复杂问题的解答中有必要进行文本生成或句子融合的研究。

表 1：问题、答案句集及参考答案示例

<p>【问题】本文结尾写道：“梅花，几千年的书香缭绕得骨清魂香，几千年的诗心陶冶得如此美丽。”请紧扣“书香”与“诗心”，谈谈你对这句话的理解。 (本例选自2015年北京市语文高考试题)</p> <p>【候选答案句集合】</p> <p>①关于梅，人们谈论太多，种梅，赏梅，写梅，画梅，梅深入到人们生活的各个角落。 ②梅花是被中国人挂在墙上，捧在手上，供在心里的，是深入血液和灵魂的一种花。 ③梅花是人间尤物，人间与仙境的使者。 ④梅花是我们中国民族精神的象征，凌寒飘香，不屈不挠，自强不息，铁骨冰心。</p> <p>【参考答案】这句话的意思是，中国人爱梅，种梅，写梅，赏梅，梅深入到人们生活的各个角落，梅花是中华民族不屈不挠，自强不息，清高孤傲精神的象征，表达作者对梅花这种品格的赞美。</p>

常见的生成文本的方法有基于压缩 (Compression-based) 的方法和基于生成 (Abstraction-based) 的方法。Compression-based 的方法通过删除句子中一些冗余的词或短语来改进生成的文本^[8-14]，但该方法主要针对单个句子，没有考虑多个句子间存在的冗余信息。Abstraction-based 的方法与人生成本文的过程较相似，对不同的句子使用删除、融合、改写等策略得到一个新句子。句子融合是典型的基于生成的方法，目前对句子融合的研究大多针对多文档摘要^[15]，基本思路是：首先根据相似度构造多个句子集，然后将每个句子集中的句子融合为一个新句子。其特点是：待融合的句子相似度高、冗余信息多；句子融合中词语重要性的度量主要基于句子本身来考虑；句子融合的流畅度通过语言模型或依存关系进行考虑。

本文面向阅读理解中的复杂问题进行句子融合研究，与上述句子融合的不同之处体现在：(1)待融合的句子大都是隐含意义丰富的复杂句，句子的冗余信息及重要信息都难以辨别，因此冗余信息的去除及重要信息的选择比多文档摘要难度更大；(2)需要结合问题来确定词语的重要程度。

此外，本文认为阅读理解问答中理想的句子融合应该满足：每个句子的重要信息都能保留，且与问题紧密相关，同时句子流畅度好。因此本文提出一种兼顾句子重要信息、问题关联度与句子流畅度的句子融合方法。该方法的主要思想为：(1)根据多个特征计算词重要度，在词重要度中考虑问题关联度；(2)基于句子拆分选择待融合部分。将句子集中的每个句子使用句法分析拆分成简单句，然后根据词的重要度估计每个简单句的重要度，将得分最高的简单句作为待融合部分；(3)基于词对齐考虑句子的关联，合并多个句子的相同信息，保留不同信息；(4)结合依存关系、二元语言模型及词重要度三方面的信息形成整数线性规划策略来优化简单融合后的句子，其中依存关系可以保证句子的主要成分没有缺失，语言模型可以保证句子的流畅度，词重要度可以保证重要信息的保留。通过在历年高考阅读理解数据集上进行测试，实验表明，本文所提句子融合策略取得了 82.62% 的 F 值。

¹ <http://21robot.org/>

² <http://research.nii.ac.jp/qalab/>

³ <http://network.chinabyte.com/12/13482012.shtml>

本文的贡献主要有：提出了阅读理解中词重要度的衡量指标，并将词重要度与问题相关联；提出用句子拆分的方法选择重要信息，去除阅读理解中难以辨别的冗余信息；将依存关系与语言模型相结合形成整数线性规划方案进行句子优化，更好地保证了句子的可读性及信息量。

2 相关研究

本文主要利用 Abstraction-based 的方法进行句子融合，因此着重对此类相关工作进行阐述。Abstraction-based 方法的基本思路是对多个句子采用一定的策略有效去除冗余信息，保留重要信息，然后得到一个新句子。基于这种思想的研究可以分为两类：词或短语到句子的生成、句子到句子的生成。

词或短语到句子的生成是以词或短语为基本生成单元，通过词的添加生成新句子的方法。Wan et al.提出一种根据一些输入的词生成一个新句子的全局修正的方法，他们把新句子的生成问题看做最大生成树的创建问题，用依存关系连接词或短语，生成过程中用了一系列硬性约束来限制语法合法性，但该方法的问题是语料中的依存关系有可能不符合当前语境^[18]。Cheung and Penn 提出了一种用格框架得到多文档摘要的方法，首先对句子进行依存分析，统计得到格框架，然后将具体句子与格框架进行模板匹配得到摘要^[19]。Bing and Li 所提方法采用的语言粒度比较粗，是通过句法分析抽取出其中的动词短语 VP、名词短语 NP 作为基本生成单元，之后通过最大化句子中的信息量选择并整合 NP、VP 短语，最大化过程中加入了一些约束，比如 NP 与 VP 是一对多的关系，代词必须删除、句子长度的限制等^[20]。

句子融合又称为句子到句子的生成，一般通过删除句子中的词达到融合的目的。句子融合的概念最初由 Barzilay and McKeown 在 2005 年在多文档文摘研究中提出，他们将句子融合定义为得到一个简洁流畅的，能够反映所有句子共同信息的融合句，其目的是得到相关句子的公共信息，从而反映多文档摘要中的重要信息。他们所提句子融合的思想是首先通过对齐语法解析树得到需要融合的片段；然后选择包含这些片段最多的一个句子，删除其中的信息或添加其它句子中的信息，得到一个融合的网络；最后基于语言模型最优化，生成一个新句子^[15]。之后，Marsi and Krahmer 进一步提出句子融合不仅要得到句子的公共信息，同时还要保留待融合句子间的不同信息，他们也用语言模型得到最优句子，但该方法无法解决词序问题^[16]。Filippova and Struber 分析句子集中每个句子的依存树，对齐依存树得到一个依存图，然后用基于依存关系的整数线性规划压缩依存图，从而得到一个新的句子，通过 GermaNet 和维基百科检验语义兼容性^[17]。Ganesan 等融合冗余信息多的观点数据得到简洁的观点摘要，具体方法是首先通过将各评论中的词对齐构造一个图，然后根据冗余、子序列、可折叠三个属性选择图中的子路径，并对其打分生成候选摘要^[21]。

现有的句子融合的方法都是面向英文多文档摘要的，汉语中关于生成的研究大多仍是基于抽取的方法^[22,23]，即简单地将抽取到的相关句按照某种策略排序后得到的线性序列作为最终的生成文本。

3 句子融合

句子融合是将句子集中的多个句子整合为一个新句子的过程，其核心环节为重要词的保留、多余词的删除，其中需要考虑冗余的去除、词语的兼容性、句子的完整性、信息的丰富性多种因素。因为阅读理解中的句子复杂多变，隐含意义丰富，句子间的语义鸿沟较大，因此阅读理解中的句子融合难度较大，会产生以下问题：

(1)冗余信息难以识别。如表 1 所示，答案句集的句子①中前半句“种梅，赏梅，写梅，画梅”与后半句“梅深入到人们生活的各个角落”表达的是相同的意思，但是从字面上无法判别这两部分是冗余信息，句子②、③、④中同样存在冗余信息难以识别的问题。

(2)待融合部分的确定。同样以表 1 句子集中句子①为例，假设已经识别出前后两部分

属于冗余信息，那么将哪部分作为待融合部分同样是一个需要考虑的问题。

(3)多信息的融合。确定了待融合部分后，多个待融合部分间可能仍含有冗余信息，多个待融合部分的融合也是一个难点。

为了解决这些问题，本文提出面向阅读理解复杂问题的句子融合方法，其基本框架是：首先将一个复杂句拆分成多个简单句，并度量每个简单句的重要度，选出其中最重要的一个简单句作为待融合部分；然后用词对齐简单的把多个待融合部分合为一个句子；最后用基于依存关系、语言模型及词重要度的整数线性规划最优化，得到最终融合的句子。

3.1 选择待融合部分

句子融合中需要考虑冗余的去除及重要信息的选择，本文用句子拆分的方法，以拆分后的简单句为单位粗粒度的进行信息的删选，同时设定了词重要度的衡量指标，对句子拆分后的简单句进行信息的度量，选择最重要的简单句作为待融合部分。

3.1.1 句子拆分

由于中文句子很复杂，抽取出的句子中往往含有冗余或者不重要的信息，所以需要对这些信息进行删选。通过观察发现，这些冗余信息往往具有并列、平行、连谓等结构，其句法结构相似、表达的意思相同，因此本文提出从句法结构出发，用句子拆分的方法将一个复杂句拆分成多个简单句粗粒度的选择待融合部分。

句子拆分的具体过程为：首先使用 Stanford Parser⁴对一个复杂句进行分析，得到其句法分析树，然后对每个结点统计其孩子结点类型，如果其孩子结点中某一类型重复出现 n 次，则可以将该结点根据重复出现的结点类型分解为 n 个子句。分解动作只做一次，因为过多的分解会使拆分后的子句过于简单，造成信息的损失。

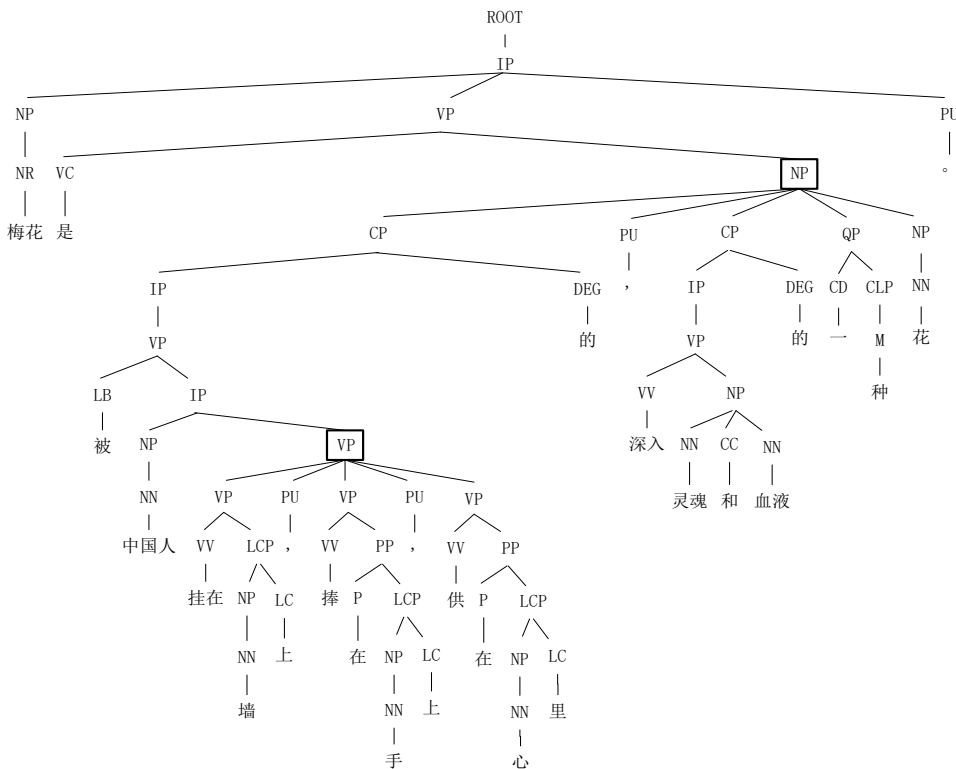


图 1: 复杂句的句法分析树

以图 1 为例，第四层中的 NP 结点的孩子结点类型 CP 重复出现 2 次，因此可以将句子根据这一信息拆分为以下 2 个句子，“梅花是被中国人挂在墙上，捧在手上，供在心里的一种花”、“梅花是深入灵魂和血液的一种花”，可以发现，两个句子的语义都是完整的。而更

⁴ <http://nlp.stanford.edu/software/lex-parser.shtml>, 参见文献[24]

底层的 VP 结点同样也有这一特征，但是根据前面的约束，对其不做任何操作。

3.1.2 待融合部分的重要度评价

对复杂句拆分之后，需要选择合适、重要的简单句作为待融合部分，其重要度的度量是通过词重要度的综合得到的。

本文根据以下特征衡量词的重要度。

(1)tfidf 特征：该特征是将句子作为短文档得到词的 tfidf 值，将其值记为 $\alpha(w_i)$ ，具体按照公式(1)计算：

$$\alpha(w_i) = \frac{w_i \text{在文档中出现的次数}}{\text{文档中总词数}} \times \log \frac{\text{文档总句子数}}{\text{包含 } w_i \text{ 的句子数}} \quad (1)$$

(2)情感词特征：散文等文学作品很多时候是借景抒情、托物言志，通过描写景物来传达作者的情感，因此情感词尤为重要。比如：“梅花是中国最美的花”，其中“美”是一个情感词，表达了作者对梅花的喜爱和赞美之情。因此本文认为情感词的重要度的权重需要加大。如果词 w_i 是情感词，则词重要度值增加 $\beta(w_i)$ 。情感词的判断是根据 hownet⁵中词的 S_C 属性。

(3)问题关联度特征：一般情况下，词与问题的关联度越大，该词越重要。本文中词 w_i 与问题 Q 的关联度由词 w_i 与 Q 中除停用词之外的每个词 q_j 的相似度衡量，使用 word2vec⁶将词表示成词向量的形式，用向量的余弦相似度计算词的相似度。 w_i 与 q_j 的相似度记为 $\cos(w_i, q_j)$ ，取其最高值作为词 w_i 与问题 Q 的关联度，记为 $\gamma(w_i, Q)$ ，具体如公式(2)所示：

$$\gamma(w_i, Q) = \max_{q_j \in Q} \{ \cos(w_i, q_j) \} \quad (2)$$

基于上述特征，在特定问题 Q 下，词 w_i 的重要度按照公式(3)计算：

$$I(w_i, Q) = a\alpha(w_i) + b\beta(w_i) + c\gamma(w_i, Q) \quad (3)$$

待融合部分 s 的重要度按照公式(4)计算：

$$I(s, Q) = \sum_{w_i \in s} I(w_i, Q) \quad (4)$$

3.2 基于词对齐的句子融合

多个待融合部分的简单融合可以通过词对齐实现。词是句子最基本的构成单元，通过词对齐可以将冗余信息减少到最低，同时，其它有用信息也可以完整的保留下来。

词对齐的过程仍然需要用到句法分析的结果，将词表示为带句法信息的词，句法信息即词的句法路径，比如图 2 中的“梅花”，其句法信息表示为“ROOT-IP-NP-NR”。然后词的对齐可以根据以下规定进行：

- (1)两个词完全相同，并且词的句法信息相同；
- (2)两个词是近义词，并且词的句法信息相同。

其中，近义词根据“哈工大社会计算与信息检索研究中心同义词词林扩展版”的第五级进行判断。融合中需要注意的一个词是“的”，由于“的”之后限定的词可能会不同，所以不能将满足规定的“的”进行融合，否则，会使句子混乱。

词对齐之后，两个对齐词的中间信息按照线性序列排列即可。如图 2(a)、图 2(b)中，能对齐的词只有“梅花”及标点符号“。”，中间信息按照线性序列排列，得到对齐后的句子为“梅花是深入灵魂和血液的一种花，有几千年的栽培史。”，其中标点符号“，”是对齐过程

⁵ <http://www.keenage.com/>

⁶ <http://Code.google.com/p/word2vec/>

中加入的。

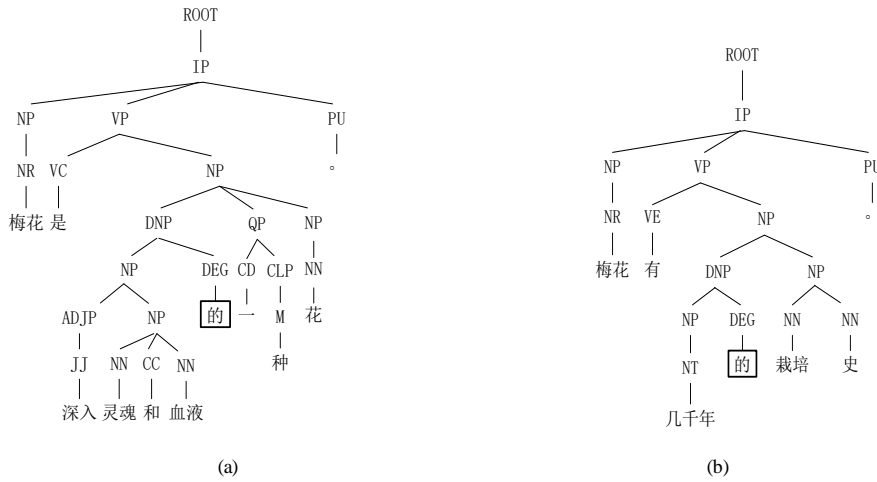


图 2: 简单句的句法分析树

3.3 句子生成

词对齐融合后的句子有时存在多余信息，同时流畅度不够，因此需要进一步处理。本文使用整数线性规划思想，将句子生成问题看做优化问题，即优化目标是使句子可读性及重要度最大化。目标函数中综合了依存关系、语言模型及词重要度三种因素，其中句子可读性通过词的依存关系概率及二元语言模型概率进行决策，句子重要度通过词重要度进行决策。

目标函数如公式(5)所示：

$$z = \max \left\{ \sum_{i=1}^n \delta(w_i) \cdot P(d_i/h_i) \cdot I(w_i, Q) + \sum_{i=0}^n \sum_{j=i+1}^{n+1} \eta(w_i, w_j) \cdot P(w_j/w_i) [I(w_i, Q) + I(w_j, Q)] \right\} \quad (5)$$

其中， w_i 表示第 i 个词， $P(d_i/h_i)$ 表示依存关系概率，其中 h_i 表示依存关系中 w_i 的父亲节点词， d_i 表示 h_i 与 w_i 的依存关系类型； $P(w_j/w_i)$ 表示二元语言模型。

$\delta(w_i)$ 、 $\eta(w_i, w_j)$ 分别由依存关系、语言模型决定，其取值如(6)(7)所示：

$$\delta(w_i) = \begin{cases} 1 & \text{如果词 } w_i \text{ 保留} \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in [1..n] \quad (6)$$

$$\eta(w_i, w_j) = \begin{cases} 1 & \text{如果词 } w_i, w_j \text{ 都保留} \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in [0..n], \forall j \in [i+1..n+1] \quad (7)$$

其中， $i=0$ 时 w_i 表示句子的开始标志， $j=n+1$ 时 w_j 表示句子的结束标志。

整数线性规划的约束如下：

(1)句子起始约束：以下两个约束保证一个句子中必须有一个词作为开始词，同时，必须有一个词作为结束词。

$$\sum_{j=1}^n \eta(w_0, w_j) = 1 \quad (8)$$

$$\sum_{i=1}^n \eta(w_i, w_{n+1}) = 1 \quad (9)$$

(2)语言模型连通性约束：下式保证语言模型中每个词都有后继词，同时每个词都有前驱词。

$$\sum_{i=0}^{k-1} \eta(w_i, w_k) = \sum_{j=k+1}^{n+1} \eta(w_k, w_j), \forall k \in [1..n] \quad (10)$$

(3)依存关系的连通性约束：下式表示词 w_i 如果保留，则其父亲结点词 h_i 也保留，以此保证依存关系的连通性。

$$\delta(w_i) - \delta(h_i) \leq 0, \forall i \in [1..n] \quad (11)$$

(4)语言模型、依存关系一致性约束：语言模型词的保留与依存关系中词的保留一致。

$$\sum_{i=0}^{k-1} \eta(w_i, w_k) = \delta(w_k), \forall k \in [1..n] \quad (12)$$

(5)语义完整性约束：保证句子主要成分，即主、谓、宾，必须保留。其中， $d_i=HED$ 时表示词 w_i 是依存关系中的核心词， $d_i=SVB$ 时表示词 w_i 是核心词的主语， $d_i=VOB$ 时表示词 w_i 是核心词的谓语。

$$\delta(w_i) = 1, \forall d_i \in \{HED, SVB, VOB\} \quad (13)$$

本文具体实现时采用了开源整数线性规划代码⁷，利用哈工大 LTP⁸进行依存分析。

4 实验结果

4.1 实验过程

实验过程中 word2vec 的训练及 $P(d_i|h_i)$ 、 $P(w_j|w_i)$ 的估计都是基于网络爬取的大规模文本数据。由于本文面向的是文学作品的阅读理解，且 word2vec 训练得到的词向量对训练语料的依赖性很强，因此我们下载了近 7 万篇文学作品，规模为 416MB，来进行相关参数的估计。词重要度中的参数在实验中经过多次测试后，最终设定为 $\beta(w_i)=0.1$ ， $a=0.8$ ， $b=0.1$ ， $c=0.1$ 。

本文的目标是复杂问题的解答，所以选择北京市近 11 年的语文高考试题进行测试，其中涉及到句子融合的复杂问题有二十道，在该数据上进行的实验记作实验一。该实验中的句子融合是基于正确的答案句子集进行。

由于实验一中的相关数据及问题比较少，为了验证本文方法的有效性，我们设计了另一组实验，记作实验二。一般来说，文章中每个自然段是一个独立的短文本，段中各句互相关联，且其主题与文章题目相关。基于该观察，实验二为：将每个文章中的每个自然段落作为一个句子集，文章题目作为问题，然后对每段中的句子做句子融合。共收集了 135 篇文章，每篇文章平均有 10 个段落，相当于对 1350 个句子集进行句子融合。

4.2 结果及分析

目前国际上还没有专门针对句子融合自动评测方法。文献[17]中使用人工评测的方法，从可读性及信息量两个方面评价句子融合结果。参照文献[17]，本文采用人工评测的方法，从可读性及信息量两个方面对句子融合结果进行打分，分值为 1-5。这里的信息量表示与问题相关的程度。实验中由三名标注者从这两个评测方面对句子融合结果打分，最后取平均值作为评测结果。另外，本文还借鉴了文献[14]中句子压缩的评测方法：首先人工对句子集中的每个词打标签，1 表示该词应该保留，否则表示删除，然后将人工标注结果与句子融合结果相对比，最后基于准确率、召回率计算得到 F 值来评价实验结果。标注者对 7267 个词标注的结果显示，两两标注的一致性最高为 84.07%，说明句子融合任务较困难。

⁷ <https://sourceforge.net/projects/lpsolve/>

⁸ <http://www.ltp-cloud.com/>

本文实现了一个类似文献[17]的方法作为 Baseline1，即首先将句子集表示成依存图，然后用基于依存句法的整数线性规划压缩依存图，生成新的句子，其中考虑了依存关系连通性约束及语义完整性约束，该方法是目前所查文献中最具代表性的句子融合方法。其中词的重要度通过 tfidf 值计算，因为在单篇文章中，tfidf 值更能显示出词之间的差异性。另外，为了去除词重要度的影响，我们将 Baseline1 中的词重要度计算方法替换为本文的词重要度计算方法，作为 Baseline2。Baseline3 是参考文献[15]中的方法，基于语言模型的整数线性规划对句子进行优化，但词重要度计算方法与本文所提方法一致。Baseline4 是参考文献[20]中的方法，该方法以句法分析后的名词短语 NP 及动词短语 VP 为基本生成单元，然后通过最大化句子重要度生成新句子。本文方法与 Baseline 的实验结果如表 2 所示，其中实验二是从句子融合结果中随机抽取 200 句进行人工评测得到的。

表 2： 本文方法与 Baseline 对比实验结果

方法	实验一（复杂问题解答中的句子融合）				实验二（针对自然段落的句子融合）			
	F值（%）	可读性	信息量	平均长度	F值（%）	可读性	信息量	平均长度
本文方法	82.62%	3.78	3.92	36.63	82.21%	3.81	3.98	33.48
Baseline1	67.42%	3.11	3.04	27.05	63.43%	3.11	3.16	26.68
Baseline2	77.49%	3.58	3.75	32.53	71.64%	3.59	3.82	28.50
Baseline3	56.77%	1.33	1.84	17.63	56.56%	1.67	2.16	15.41
Baseline4	54.56%	2.12	2.06	23.39	53.01%	2.23	2.09	21.21

本文方法与 Baseline1 对比可以看出，在复杂问题中本文句子融合方法的效果明显优于 Baseline1；与 Baseline2、Baseline3 对比可以看出，基于依存关系及语言模型的整数线性规划函数比单纯基于依存关系或语言模型的整数线性规划函数效果好；Baseline4 的语言粒度较粗，而本文面向的句子都是复杂句，因此该方法效果较差，容易产生冗余无法去除、融合效果差、语法不合理等问题。

为了评测复杂问题中词的问题关联度的影响，我们设计了一组去除问题关联度特征的对比实验，实验结果如表 3 所示，可以看出，问题关联度特征对问题的回答有重要的指示作用。

表 3： 去除问题关联度的对比实验结果

方法	实验一（复杂问题解答中的句子融合）				实验二（针对自然段落的句子融合）			
	F值（%）	可读性	信息量	平均长度	F值（%）	可读性	信息量	平均长度
本文方法	82.62%	3.78	3.92	36.63	82.21%	3.81	3.98	33.48
去除问题关联度	67.86%	3.42	3.21	29.21	68.13%	3.51	3.27	29.21

为了评价本文所提词重要度表示方法中各项参数对实验结果的影响，我们设计了 3 组对比实验。本文方法涉及到的参数分别为：词重要度中的 tfidf 特征(TI)，词重要度中的情感词特征(SW)，词重要度中的问题关联度特征(QR)。对比实验结果如表 4 所示：

表 4： 各参数影响

参数	实验一（复杂问题解答中的句子融合）				实验二（针对自然段落的句子融合）			
	F值（%）	可读性	信息量	平均长度	F值（%）	可读性	信息量	平均长度
① TI+ SW+ QR	82.62%	3.78	3.92	36.63	82.21%	3.81	3.98	33.48
② TI+ SW	67.86%	3.42	3.21	29.21	68.13%	3.51	3.27	29.21
③ TI+ QR	80.67%	3.69	3.86	36.21	74.09%	3.76	3.83	30.88
④ TI	67.06%	3.29	3.14	28.79	62.02%	3.43	3.20	28.55

由实验①、②、③、④可以看出，整数线性规划目标函数相同的情况下，词重要度的度量对实验结果有一定影响，包含 TI、SW、QR 三个特征的词重要度量比其它度量指标效果好。

综上所述，本文方法中词重要度的度量指标对信息量的衡量有明显作用，同时基于依存

关系及语言模型的线性规划也很大程度的提高了句子的流畅度，保留了句子中的重要信息。

本文实验中存在的问题有以下两点：(1)实验严重依赖于句法分析及依存分析工具，句法分析结果对句子拆分、词对齐有重要影响，依存分析结果对整数线性规划最优化有重要影响，错误的分析结果最终会产生不合理的句子；(2)由于阅读理解中语言现象比较复杂，经常会出现很多不常见的搭配词汇，因此不可避免的整数线性规划部分 $P(d_i|h_i)$ 、 $P(w_j|w_i)$ 的统计结果会有数据稀疏的问题，对实验结果也有一定的影响。表 5 是对表 1 中句子集融合时，各个处理环节结果的一个展示。

表 5: 句子融合示例

<p>【问题及待融合句子集】见表1中的【问题】、【候选答案句集合】</p> <p>【待融合部分选择结果】</p> <ul style="list-style-type: none">①关于梅，人们谈论太多，种梅，赏梅，写梅，画梅。②我说梅花是深入血液和灵魂的一种花。③梅花是人间与仙境的使者。④梅花是我们中国民族精神的象征。 <p>【词对齐融合结果】关于梅，人们谈论太多，种梅，赏梅，写梅，画梅，我说梅花是深入血液和灵魂的一种花，梅花是人间与仙境的使者，我们中国民族精神的象征。</p> <p>【句子优化生成结果】关于梅，人们谈论多，种梅，赏梅，写梅，画梅，我说是深入血液的一种花，是人间使者，我们中国民族精神象征。</p>

5 总结

阅读理解中复杂问答的研究，除了答案句抽取之外，必不可少的一项是答案的生成，针对这一难题，本文提出了句子融合的生成方法。在句子融合时，首先考虑词语重要度的衡量指标；然后拆分待融合句，根据词重要度从拆分得到的简单句中选择最重要的简单句作为待融合部分；再用简单的词对齐策略融合多个句子的关联信息；最后进行句子流畅度的考量，使用基于依存关系、语言模型及词重要度的整数线性规划得到最优句。最终的实验结果表明这种句子融合的方法效果较好。在未来的工作中，首先需要考虑具体问题的答案句集的构建；其次在句子融合过程中，除词对齐的策略外，应该再考虑其它融合策略。

参考文献

- [1] Voorhees E M, Tice D M. Building a question answering test collection[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2000:200-207.
- [2] 张志昌, 张宇, 刘挺, 等. 开放域问答技术研究进展[J]. 电子学报, 2009, 37(5):1058-1069.
- [3] Matthew Richardson, 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.
- [4] Jason Weston, Antoine Bordes, Sumit Chopra, et al. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks[J]. Computer Science, 2015.
- [5] Lynette Hirschman, Marc Light, Eric Breck, et al. Deep Read: A reading comprehension system[C]// Meeting of the Association for Computational Linguistics, 2002:325-332.
- [6] 张志昌, 张宇, 刘挺, 等. 基于话题和修辞识别的阅读理解 why型问题回答 [J]. 计算机研究与发展, 2011, 48(2):216-223.
- [7] Jawad Sadek, Fairouz Chakkour, Farid Meziane. Arabic Rhetorical Relations Extraction for Answering "Why" and "How to" Questions[C]// International Conference on Applications of Natural Language Processing and Information Systems, 2012:385-390.
- [8] Kevin Knight, Daniel Marcu. Statistics-Based Summarization - Step One: Sentence Compression[C]// Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. AAAI Press, 2000:703-710.
- [9] Nitin Madnani, Jimmy Lin, Bonnie Dorr. TREC 2007 ciQA Task: University of Maryland[C]// Sixteenth Text Retrieval Conference, Trec 2007, 2007:214-220.

- [10] K. Knight, D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression[J]. *Artificial Intelligence*, 2002, 139(1):91-107.
- [11] J. Turner and E. Charniak. Supervised and unsupervised learning for sentence compression[C]// *Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005:140-141.
- [12] RT McDonald. Discriminative Sentence Compression with Soft Syntactic Evidence.[C]// *Eacl 2006, Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 2006.
- [13] Wanxiang Che, Yanyan Zhao, Honglei Guo, et al. Sentence compression for aspect-based sentiment analysis[J]. *Audio Speech & Language Processing IEEE/ACM Transactions on*, 2015, 23(12):2111-2124.
- [14] Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, et al. Sentence Compression by Deletion with LSTMs[C]// *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015: 360-368.
- [15] Barzilay, Regina, Kathleen R. McKeown. Sentence Fusion for Multidocument News Summarization[J]. *Computational Linguistics*, 2005, 31(3):297-328.
- [16] Marsi, Erwin, Emiel Kraemer. Explorations in sentence fusion[C]// *Proceedings of the 10th European Workshop on Natural Language Generation*, 2010:109-117.
- [17] Katja Filippova, Michael Strube. Sentence fusion via dependency graph compression[C]// *Conference on Empirical Methods in Natural Language Processing*, 2008: 177-185.
- [18] Stephen Wan, Robert Dale, Mark Dras, et al. Global revision in summarization: Generating novel sentences with Prim's algorithm[C]// *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 2007: 26-235.
- [19] Jackie Chi Kit Cheung, Gerald Penn. Towards robust abstractive multi-document summarization: A caseframe analysis of centrality and domain[C]// *Meeting of the Association for Computational Linguistics*, 2013: 1233-1242.
- [20] Lidong Bing, Piji Li, Yi Liao, et al. Abstractive Multi-Document Summarization via Phrase Selection and Merging[C]// *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015: 1587-1597.
- [21] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: A Graph-based Approach to Abstractive Summarization of Highly Redundant Opinions[C]// *COLING 2010 -, International Conference on Computational Linguistics, Proceedings of the Conference*, 2010:340-348.
- [22] 王红玲, 张明慧, 周国栋. 主题信息的中文多文档自动文摘系统[J]. *计算机工程与应用*, 2012, 48(25):132-136.
- [23] 刘江鸣, 徐金安, 张玉洁. 基于隐主题马尔科夫模型的多特征自动文摘[J]. *北京大学学报: 自然科学版*, 2014, 50(1):187-193.
- [24] Marie-Catherine de Marneffe, Bill MacCartney, Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses[J]. *Lrec*, 2006:449-454.



谭红叶 (1971——), 女, 副教授, 博士, 硕士生导师, 主要研究方向为中文信息处理、信息检索。

Email: tanhongye@sxu.edu.cn



赵红红（1992——），女，硕士研究生，主要研究方向为中文信息处理。

Email: 1325046270@qq.com



李 茹（1963——），女，教授，博士，博士生导师，主要研究方向为中文信息处理、信息检索。

Email: liru@sxu.edu.cn

作者联系方式：谭红叶，山西省太原市小店区坞城路 92 号 山西大学计算机与信息技术学院，030006，13623514258，tanhongye@sxu.edu.cn；赵红红，山西省太原市小店区坞城路 92 号 山西大学计算机与信息技术学院，030006，18734819125，1325046270@qq.com；李茹，山西省太原市小店区坞城路 92 号 山西大学计算机与信息技术学院，030006，13073538012，liru@sxu.edu.cn。