百家论坛

篇章语义分析：让机器读懂文章

张牧宇，刘 铭，朱海潮，秦 兵 / 哈尔滨工业大学

自然语言处理的研究从词汇、词典的研究起步，近年来一直把句子作为核心的研究对象，对篇章的语义分析多是语言学家从理论上进行探索，计算机科学家对篇章范围语义现象的关注有限。但是，很多语义问题必须在篇章层面上才能够得到根本性的解决，比如“共指消解”、“语义关系识别”和“事件融合与关系识别”等。同时，这些篇章级语义问题的解决对于词汇级和句子级的分析同样具有反哺性的指导意义。近年来，中文词汇、句子级自然语言处理技术的发展，特别是词义消歧、句法分析和语义角色标注等研究工作的进展，为篇章语义分析的研究创造了技术条件。同时，搜索引擎等重大互联网应用也向篇章语义分析提出了应用上的强烈需求。如果能够建立一套既具有理论深度，又具有现实可行性的篇章语义分析的理论和方法体系，对于自然语言处理学术和应用的发展无疑都将具有重要意义。本文概述了目前主流的篇章语义分析方法，并简要介绍了其应用前景。

什么是篇章语义分析

篇章（Discourse）是指由词和句子以复杂的关系链接而成，能够完成一定交际任务的完整连贯的语言单元。篇章语义分析（Discourse Analysis）是指在篇章层面上，将语言从表层的没有结构的文字序列转换为深层的有结构的机内表示，刻画篇章中的各部分内容的语义信息，并识别不同部分之间存在的语义关联，进而融合篇章内部信息和外部背景知识，更好地理解原文语义。篇章语义分析的研究建立在词汇级、句子级语义分析之上，融合篇章上下文的全局信息，分析跨句的词汇之间、句子与句子之间、段落与段落之间的语义关联，从而超越词汇和句子分析，达到对篇章等级更深层次的理解。

篇章语义分析方法概述

篇章语义分析主要有以下三个主流的研究方向。

以篇章结构为核心

此类研究工作的目标是识别不同文本块之间的语义关系，例如条件关系、对比关系等，亦称为修辞关系识别。根据是否需要将文本分割为一系列彼此不相交的覆盖序列，可以将本类方法进一步分成两大类：第一类以修辞结构理论（Rhetorical Structure Theory）和篇章图树库（Discourse GraphBank）为代表，要求先将文本切分为彼此不相交的语义单元，并随后分析各部分之间的语义关系及结构组成；第二类方法以宾州篇章树库理论（Penn Discourse TreeBank）为代表，不需要预先切分文本，而是直接识别篇章关系及其元素所在位置，并随后识别具体的语义关系类型。

修辞结构理论(RST，Rhetorical Structure Theory,)最早由 Mann 和 Thompson 在 1988 年发表的论文[1]中提出。除 Mann 和 Thompson 在该方向持续发表相关工作论文之外，Marcu 在其论文中对 RST 理论进行了分析，并持续探索该方向，提出两种基于 RST 理论分析的文本处理方法[2]：① 识别提示短语（CP，Cue Phrases），然后将整句打散成若干个子句；② 为无结构的文本建立一个有效的修辞结构树。RST 理论以文本结构为分析对象，从小单元之间的连接关系开始，逐步延伸到自然语言段落和完整的语篇。RST 在汉语中的跨语言可转移性有特殊的背景。可惜，虽然有不少对 RST 的中文介绍和初步应用计划等，但实质性的发

展应用很少。目前，山西大学李茹教授的团队正在尝试进行中文 RST 树库的构建工作，所产出的资源已经具有一定规模，非常值得期待，只是尚未有公开的成果发表。因此，RST 理论现在在中文尚没有一套完整可用的系统或理论，应用难度较高。

篇章图树库（Discourse GraphBank）最初由 Wolf et al[3] 提出。该理论认为，相比于树结构，篇章更适合于表示为图。在最初的文章中，他们详细讨论了图表示与树表示的差别，并构建了一个由 135 篇文档构成的篇章树库资源。他们提出，图表示允许将文章中的不同内容以更自由的形式表示出来，从而可以尽可能地获取丰富信息。关于 Discourse GraphBank 与 RST Discourse TreeBank 的区别可详见文献 [3]。

宾州篇章树库理论（Penn Discourse Treebank）是宾州大学的研究人员采用的一种以词汇为中心的方法 [4]，在句子级的 Penn TreeBank 树库的基础上，以篇章关联词语为核心，从语义角度出发构建了篇章关系树库 [5]。该研究检测同一篇章内两个文本单元（片段、分句、复句、句群、段落等）之间的逻辑语义关联（因果关系、转折关系等），将句内的语义分析结果扩展为篇章级别的语义信息，从而成为语义分析的重要解决途径之一。

根据文本单元间是否存在篇章连接词，可将篇章句间的关系分为包含关联词的显式篇章句间关系（Explicit Discourse Relation，简称显式关系）[6-7] 与不含关联词的隐式篇章句间关系（Implicit Discourse Relation，简称隐式关系）[8-11]。由于隐式篇章关系缺少关联词，无法直接推测语义关系类型，需要根据上下文进行推测，因此也更加难以识别。

目前采用 PDTB 标准构建的篇章语料主要面向英语 [5]，除此以外印度语 [12]、土耳其语 [13] 和阿拉伯语 [14] 上也有相应的研究和资源出现。在中文上，布兰迪斯大学的 Xue 教授最早尝试了中文关联词标注于分析工作 [15]，并尝试按照 PDTB 体系标注中文树库。除此以外，Huang et al[16] 也在相关工作上做了一定尝试。值得一提的是，哈尔滨工业大学社会计算与信息检索研究中心秦兵教授课题组，采用 PDTB 框架，历时数月，标注超过 20 000 个实例，构建了一份大规模的中文篇章语料库 [17]，并于 2014 年对学术界免费共享。

整体来说，以篇章结构为核心的篇章语义分析研究中，文本的语义信息首先被转换为文本块间的修辞结构，随后具体化为相应的语义关系类型（例如因果关系、转折关系等）。对于以修辞结构理论（RST）为代表的一类研究而言，文本块间的修辞结构应该满足一种树形结构；而对于以宾州树库理论（PDTB）和篇章图理论（Discourse GraphBank）而言，文本块间的修辞结构则倾向于线形结构，同时允许一定的交叉和跨越关系存在。这些研究兼有表现力和实用性，通过定义修辞结构和语义关系，这些方法可以获取一定程度的语义信息，并且采用超越了词汇级别的基本处理单元，表现力较强。缺点主要在于结构分析难度较大，无论 RST 还是 PDTB 都对篇章结构做了部分假设从而降低难度，提升操作性，但也损失了语义结构的完整性。更重要的是，在语义类型识别方面，由于语义问题本身的复杂性和歧义性，导致识别难度较大；而已有的相关研究主要关注篇章内部特征的挖掘和使用，对外部语义知识的使用不足，这也在一定程度上限制了最终的识别效果。

以词汇语义为核心

最典型的代表为词汇链理论（Lexical Chain Theory），其由 Morris et al[18] 于 1991 提出的。“词汇链”是指一个主题下的一系列相关的词共同组成的词序列。该算法的基本假设非常直观：用于描述特定主体的多个词语，在语义层面上应该是相关的，并且围绕特定主体展开构成一条相关词汇的链条。这样聚集起来的相关词汇的链条即称为“词汇链”，作为特定语言片段内部各个主题的指示。如果能够分析获知多个词汇链在文中的分布，那么对应的文章结构也就确定了，属于一种静态的语篇连贯研究方法。

与链状的词汇链不同，中心理论（Centering Theory）主要针对篇章结构中的焦点、指代表达式选择、话语一致性等进行研究。最初由 Grosz et al[19] 在 1995 年提出，通过跟踪句子的“中心”变化来描述篇章。“中心”指的是将当前句子与其他句子关联在一起的实体，如果一句话有了这种“中心”实体，那么它将不再是独立的句子，而是与上下文相关的语句。如此，他们将“句子 (Sentence)”与“语句 (Utterance)”区分开来，用“句子 (Sentence)”指代一个普通的词的序列；用“语句

（Utterance）”代指这种具有中心的、与上下文相关的句子。所以其认为，这些“中心”才是组成语篇结构的基础成分。

篇章连贯性理论（Discourse Coherence Evaluation）是篇章语义分析研究的另一典型代表。该研究最初始于 Grosz et al[19]1995 年提出的“中心定理”，通过对“中心”的刻画直接反映了篇章连贯信息。近年来，篇章连贯性分析研究获得了比较快的发展，出现了一些操作性较强的方法和研究。2005 年，Barzilay et al[20] 提出了经典的基于实体的连贯性评估方法，该方法分析各个实体在多个句子中是否出现及相应句法角色，将待评估的文章转化为 Entity-grid，并利用该 Entity-grid 抽取特征训练有指导模型来进行连贯性评估。2008 年，Elsner et al[21] 在经典的 Entity-grid 模型的基础上，对篇章实体进行了进一步细分，引入新实体的概念和实体间的共指信息，显著提升了系统性能。随后，他们进一步丰富了 Entity-grid 方法，向表格的项中添加了关于实体显著性的信息，以更加提升系统性能 [22]。

在上文介绍的以词汇语义为核心的篇章语义分析研究中，文本的语义信息通过词汇间的语义关联体现。具体来说，语义相关的词汇、实体在文档中的分布情况，也可以体现篇章的行文结构以及各部分之间的语义关联，此类研究中的不同理论与方法从不同的角度对篇章信息进行了刻画。具体来说，语义词汇链理论（Lexical Cohesion）通过分析普通词汇（包括名词、形容词等）的语义信息构建主题词汇链，利用词汇之间的分布和转移方式分析篇章语义。中心理论（Centering Theory）和连贯性分析则主要以实体为分析对象，利用实体（包括共指实体、相关实体等）的分布和重现刻画篇章信息。这一类的研究理论完善，操作性也比较强；但以词汇为分析对象，表现力比较有限，而且语义关系以关联为主，对具体的语义类型（例如因果关系、转折关系）没有进行更细致的区分。另外，此类方法通过词汇的衔接来反映篇章结构，不利于刻画复杂的篇章结构信息。

以背景知识为核心

此类研究工作需要借助语义词典作为背景知识，帮助分析篇章语义关系。经过国内外专家的努力，目前已经产生一些初具规模，并具有一定实用程度的语义词典资源。在国外有以描写词汇上下位、同义、反义等聚合关系为主的 WordNet[23]，以描写语言成分之间的各种组配关系为主的 FrameNet[24]。而国内比较知名的有知网（HowNet）[25]、清华大学开发的以语义组合关系为主的《现代汉语动词分类词典》[26]、北京大学基于 WordNet 框架开发的中文概念词典（CCD，Chinese Concept Dictionary）[27]、台湾中研院集成多资源的 SinicaBow（the Academia Sinica Bilingual Ontology WordNet）[28]、哈尔滨工业大学在同义词词林（Cilin）基础上开发的同义词词林（扩展版）等。

随着 Web 2.0 的发展，用户产生内容使得互联网上的信息量爆增。以 Wikipedia（维基百科）为代表的，使用群体智慧构建的在线百科就是其中的典型代表。Wikipedia 是一种在线协作式编辑的多语言百科知识库，它以概念（concept）为单位维护一个独立的页面，其中包含对该概念的全面丰富的内容介绍（content）。Wikipedia 具有开放式的分类，不局限于特定的层次分类。每个概念根据不同角度可以归入不同的类别，即每个概念可以属于一个或多个分类（category）。Strube et al [29] 最早提出基于 Wikipedia 的语义相关度计算方案——WikiRelate。他们使用 Wikipedia 的分类节点为代表词，计算节点之间的最短路径衡量词的相关程度，达到了与 WordNet 相当的效果。Gabrilovich et al[30] 提出了显式语义分析（ESA，Explicit Semantic Analysis) 模型，他们首先将文本表示成高维 Wikipedia 概念向量，通过计算向量余弦相似度等得到文本之间的相关程度。这种将文本表示成概念集合的方式易于理解，且语义表示能力较强。Witten et al [31] 在前人工作基础上，提出了 WLM（the Wikipedia Link-based Measure）度量方法，主要使用 Wikipedia 概念中包含的大量超链接，而非分类和概念文章内容，反映文本的语义信息。类似的方法还有文献 [32-33]。

由于 Wikipedia 蕴含着丰富的语义知识，已有工作大都采用词匹配或检索方法将文本映射到 Wikipedia 的概念网络，并以此作为对文本的补充。然而，由于 Wikipedia 页面中的信息过多，引入整个页面较易导致噪音问题。此外，中文维基百科的质量远不及英文，也会限制中文相关的研究工作。

哈尔滨工业大学的张牧宇博士根据认知心理学中的联想主义理论将背景知识（例如 Wikipedia）表示为统一的三元组结构后，将其引入到篇章语义分析中，并将分析结果用于检测篇章语义的连贯性，以衡量联想背景知识的效果 [34]。

框架语义学（Frame Semantic）是由 Fillmore et al[35] 在格语法基础上，进一步提出的研究词语意义和句法结构意义的语义学理论。该理论认为，词汇的语义必须跟具体的认知结构相联系，同一个词语在不同的结构中可能具有不同的语义，而这里所说的认知结构即为“框架”。框架语义学认为，词语的意义通常与人脑中预先存在的概念结构相互联系，而这些概念结构又与个体所处的具体情境有关，涉及到实体属性、社会制度、行为模式等语义框架的约束。因此，人们可以根据自己的经验刻画不同的背景框架，并进而对同一个框架下的各个词语定义具体的框架元素。该项目最早起源于美国加州大学伯克利分校于 1997 年开始的一个以框架语义学为理论基础，以真实语料为事实依据的计算机词典编撰工程，且至今仍在进行。目前为止，FrameNet V1.5 已构建了 960 个语义框架，覆盖 11 600 个词汇，其中超过 6 800 个词汇被完全标注，已标注 15 万多个例句，并仍然在不断扩充。

从整体上来说，以背景知识为核心的篇章语义分析研究中，文本语义信息通过人工构建的背景知识资源体现，分析过程也围绕相应资源来展开。根据知识源的特点，分析过程和侧重点也各不相同。具体而言，语义词典（Dictionary）和在线百科（Online Encyclopedia）相对宽泛，适用于多种语义信息需求以及丰富的应用场景；框架语义学（FrameNet）以动词为核心，通过构建“语义框架”将语义知识转化为计算机词典，用词义间的关联反映语义，此方法信息丰富，对语义的刻画相对完整，便于计算机使用，所提供的语义信息可以用于各种应用，价值很高。缺点在于严重依赖于背景知识资源的覆盖率，对资源质量要求很高；而此类资源又大都专业性较强，构建过程耗时耗力，很难形成规模，难以穷尽现实场景，从而限制了实用性。基于在线百科的资源又存在噪音较大，信息不够精确等问题。

篇章语义分析的应用

由于篇章语义分析以篇章结构和语义信息为分析目标，因此对机器翻译研究（MT，Machine Translation）的促进作用最为直接。在已有的工作中，研究人员利用篇章语义分析技术从很多角度辅助机器翻译系统的性能提升。首先，篇章语义分析研究结果能够刻画 MT 系统的输入文本块之间的语义关系，这对 MT 系统更合理地组织翻译结果无疑是有益的 [36]。此外，篇章语义分析对关联词、文档结构都进行了比较深入的分析，这些信息有助于提升翻译文本的连贯性，生成可读性更好的翻译结果 [37-38]。另外，篇章级别的机器翻译评价始终是一个难题，通过引入篇章语义分析研究结果，可以在篇章层面上利用核函数捕捉结构信息，有助于更好地进行翻译质量评估 [39-40]。

自动问答系统（QA，Question Answering）是另一个从篇章语义分析研究中受益的重量级应用。通常情况下问答系统包括问句理解和答案抽取两个模块。在问句理解部分，篇章语义分析有助于理解题干各部分之间的语义关系，从而加深对问题的理解 [41]；在答案抽取方面，篇章语义分析可以用来更精确地分析答案所在文本 [42]，进行候选答案的重排序，有助于更准确的回答问题 [43]。除了传统 QA 研究之外，近年来阅读理解研究也受到了越来越多的关注。阅读理解的任务是对于给定的一篇自然语言文章和给定与文章相关的问题，计算机根据词语特征等语义信息来自动选择与问句相关的候选答案句。在阅读理解任务中，文章主题的广泛性要求对语料库进行深度加工和处理，才能得到比较好的结果。阅读理解研究可以直接应用到许多的社会领域，它不但是自然语言处理的一个重要的研究方向，而且可以对自然语言处理技术的成熟有很大促进作用。事实上，组成篇章结构的语句、片段之间有着明显的语义关系，这些关系可以加深对问题的理解 [41]。在文献 [44-45] 中，已经证明了句法关系对阅读理解答案抽取有促进作用，但是其性能的提升并不明显。目前已有的基于概率和机器学习的答案抽取方法中，都是将篇章中的各个句子看作是相互没有语义关联的独立信息描述单位。但在实际上，篇章中的不同句子之间存在者紧密的逻辑语义关系，全部句子结合之后来才能完成对篇章主题的全面描述。因此，通过在篇章中逐一判别每个句子和用户问题之间逻辑匹配度的方法来选择答案句，就无法正确回答用户的所有问题。基于此，即有了

结合篇章语义分析的阅读理解方法，与传统 QA 类似，该研究也得益于篇章语义分析 [41,46-47] 而获得了性能的显著提升 [48]。

挑战与机遇

目前主流的篇章语义分析方法以有指导的分析方法为基础，其依赖于带标注信息的语料资源，而此类资源严重匮乏。其次，由于语料资源的匮乏以及篇章关系分析任务本身的复杂性，目前为止，中英文篇章关系的识别的研究均处于初期阶段，限制了篇章语义分析研究的继续深入。事实上，按照张牧宇博士的论文 [34]，原文之外的相关背景知识能够有效地帮助挖掘原文内容中的语义信息。因此，原文并不能独立于背景知识而存在，缺少背景知识必然会影响对原文的分析与理解。但是，目前缺少一种合适的背景知识表示方法，并且也缺少一种有效的将背景知识和原文进行连接的方法。这些问题限制了篇章语义分析性能的提升。

作为一个新兴的研究热点，篇章语义分析方面的研究还远远不够，无论是背景知识获取还是原文语义分析都有更进一步发展的空间。除本文介绍的应用之外，融入背景知识的篇章语义分析还可以应用在其他很多领域，例如，篇章语义分析结果有助于生成更好的文摘结果 [49]；篇章语义分析还可用于文本可读性分析，即通过篇章语义分析判定文本结构是否合理、语义是否连贯，进而评估文章的可读性 [50]。总之，无论从理论研究的角度，还是从应用需求的角度，篇章语义分析都已经成为一个非常重要的研究方向。随着研究工作的不断深入和相关方法技术的逐渐成熟，篇章语义分析研究定会向更深入、更全面、更完善的目标前进，并促进机器翻译、自动问答、自动文摘，以及自然语言生成等相关研究的发展。

参考文献

[1] Mann W. C., Thompson S. A. Rhetorical structure theory: Toward a functional theory of text organization [J]// Text, 1988, 8(3): 243-281.

[2] Marcu D. The rhetorical parsing of natural language texts [C]// In Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, 1997: 96-103.

[3] Wolf F., Gibson E. Representing discourse coherence: A corpus-based study [J]// Computational Linguistics, 2005, 31(2): 249-287.

[4] Webber B. L., Joshi A. K. Anchoring a lexicalized tree-adjoining grammar for discourse [C]// In Coling/ACL Workshop on Discourse Relations and Discourse Markers, 2002: 86-92.

[5] Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A., Webber B. The Penn Discourse TreeBank 2.0 [C]// In Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008.

[6] Pitler E., Raghupathy M., Mehta H., Nenkova A., Lee A., Joshi A. K. Easily identifiable discourse relations [R]// Technical Reports (CIS), 2008.

[7] Pitler E., Louis A., Nenkova A. Automatic sense prediction for implicit discourse relations in text [C]// In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009: 683-691.

[8] Marcu D., Echihabi A. An unsupervised approach to recognizing discourse relations [C]// In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002: 368-375.

[9] Lin Z., Kan M. Y., Ng H. T. Recognizing implicit discourse relations in the Penn Discourse Treebank [C]// In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: 343-351.

[10] Wang W., Su J., Tan C. L. Kernel based discourse relation recognition with temporal ordering information [C]// In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 710-719.

[11] Louis A., Joshi A., Prasad R., Nenkova A. Using entity features to classify implicit discourse relations [C]// In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2010: 59-62.

[12] Prasad R., Husain S., Sharma D. M., Joshi A. Towards an annotated corpus of discourse relations in Hindi [C]// In Proceedings of the 3th International Joint Conference on Natural Language Processing, 2008: 73-80.

[13] Zeyrek D., Webber B. L. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus [C]// In Proceedings of the 3th International Joint Conference on Natural Language Processing, 2008: 65-72.
[14] Alsaif A., Markert K. Modeling discourse relations for Arabic [C]// In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011: 736-747.
[15] Xue N. Annotating discourse connectives in the Chinese Treebank [C]// In Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky, 2005: 84-91.
[16] Huang H. H., Chen H. H. Chinese discourse relation recognition [C]// In Proceedings of the 5th International Joint Conference on Natural Language Processing, 2011: 1442-1446.
[17] 张牧宇 , 秦兵 , 刘挺 . 中文篇章级句间语义关系体系及标注 [J]// 中文信息学报 , 2014, 28(2): 28-36.
[18] Morris J., Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text [J]// Computational Linguistics, 1991, 17(1): 21-48.
[19] Grosz B. J., Weinstein S., Joshi A. K. Centering: A framework for modeling the local coherence of discourse [J]// Computational linguistics, 1995, 21(2): 203-225.
[20] Barzilay R., Lapata M. Modeling local coherence: An entity-based approach [C]// In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005: 141-148.
[21] Elsner M., Charniak E. Coreference-inspired coherence modeling [C]// In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, 2008: 41-44.
[22] Elsner M., Charniak E. Extending the entity grid with entity-specific features [C]// In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011: 125-129.
[23] Fellbaum C. WordNet [M]// Wiley Online Library, 1998.
[24] Fillmore C. Frame semantics [M]// Linguistics in the Morning Calm, 1982: 111-137.
[25] Dong Z. D., Dong Q. HowNet and the computation of meaning [M]// World Scientific, 2006.
[26] 林杏光 , 王玲玲 , 孙德金 . 现代汉语述语动词机器词典 [J]// 北京语言学院出版社 , 1994.
[27] Yu J., Yu S., Liu Y., Zhang H. Introduction to Chinese concept dictionary [C]// In Proceedings of the 2001 International Conference on Chinese Computing, 2001: 361-367.
[28] Huang C. R., Chang R. Y., Lee H. P. Sinica BOW (Bilingual Ontological WordNet): Integration of bilingual WordNet and SUMO [C]// Lecture Notes in Computer Science, 2004.
[29] Strube M., Ponzetto S. P. WikiRelate! Computing semantic relatedness using Wikipedia [C]// In Proceedings of the 21st National Conference on Artificial Intelligence, 2006: 1419-1424.
[30] Gabrilovich E., Markovitch S. Computing semantic relatedness using Wikipedia based explicit semantic analysis [C]// In Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007: 1606-1611.
[31] Witten I., Milne D. An e ective, low-cost measure of semantic relatedness obtained from Wikipedia links [C]// In Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy, 2008: 25-30.
[32] Milne D., Witten I. H. Learning to link with Wikipedia [C]// In Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008: 509-518.
[33] Yeh E., Ramage D., Manning C. D., Agirre E., Soroa A. WikiWalk: Random walks on Wikipedia for semantic relatedness [C]// In Proceedings of the 2009 Workshop on Graph- based Methods for Natural Language Processing, 2009: 41-49.
[34] 张牧宇 . 融入背景知识的篇章语义分析方法研究 [D]// 哈尔滨工业大学 , 2016.
[35] Fillmore C. J., Johnson C. R., Petruck M. R. Background to framenet [J]// International Journal of Lexicography, 2003, 16(3): 235-250.
[36] Li J. J., Carpuat M., Nenkova A. Assessing the discourse factors that influence the quality of machine translation [C]// In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 283-288.
[37] Tu M., Zhou Y., Zong C. Enhancing grammatical cohesion: Generating transitional expressions for SMT [C]// In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 850-860.

[38] Biran O., McKeown K. Discourse planning with an n-gram model of relations [C]// In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1973-1977.
[39] Guzmán F., Joty S., Arquez L., Nakov P. Using discourse structure improves machine translation evaluation [C]// In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 687-698.
[40] Zhai K., Williams J. D. Discovering latent structure in task-oriented dialogues [C]// In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 36-46.
[41] Narasimhan K., Barzilay R. Machine comprehension with discourse relations [C]// In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 1253-1262.
[42] Friedrich A., Pinkal M. Discourse-sensitive automatic identification of generic expressions [C]// In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 1272-1281.
[43] Jansen P., Surdeanu M., Clark P. Discourse complements lexical semantics for non-factoid answer reranking [C]// In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 977-986.
[44] Ng H. T., Teo L. H., Lai J., Kwan P. A machine learning approach to answering questions for reading comprehension tests [C]// In Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 2000: 124-132.
[45] Xu K., Meng H. Using verb dependency matching in a reading comprehension system [C]// In Proceedings of the 1st Asia Information Retrieval Symposium, 2004: 190-201.
[46] Song W., Fu R., Liu L., Liu T. Discourse element identification in student essays based on global and local cohesion [C]// In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 2255-2261.
[47] Cohan A., Goharian N. Scientific article summarization using citation-context and article's discourse structure [C]// In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 390-400.
[48] Afantenos S., Kow E., Asher N., Perret J. Discourse parsing for multi-party chat dialogues [C]// In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 928-937.
[49] Louis A., Joshi A., Nenkova A. Discourse indicators for content selection in summarization [C]// In Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2010: 147-156.
[50] Pitler E., Nenkova A. Revisiting readability: A unified framework for predicting text quality [C]// In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008: 186-1.

张牧宇

哈尔滨工业大学计算机学院博士研究生（已毕业）。主要研究方向为自然语言处理、篇章语义分析。

刘 铭

哈尔滨工业大学计算机科学与技术学院副教授，硕士生导师。主要研究方向为自然语言处理、文本挖掘、篇章语义分析。

朱海潮

哈尔滨工业大学计算机学院社会计算与信息检索研究中心在读硕士研究生。主要研究方向为自然语言处理、篇章语义分析。

秦 兵

哈尔滨工业大学计算机学院教授，博士生导师，社会计算与信息检索研究中心副主任。主要研究方向为自然语言处理、情感分析、信息抽取、篇章语义分析。