目录

第一章　话语与话语计算

本章提要

话语是指大于句子、具有一定交际功能的语义单位，话语语言学作为语言学的一个分支，从不同角度对话语进行研究，取得了丰硕的成果。在人工智能时代，人机对话、机器翻译、舆情监控等都对话语的计算分析提出更高要求，如何做到让计算机真正理解人类话语，实现从表层结构到深层语义的映射，是认知智能领域亟待解决的问题。

1.1　话语与话语研究

1.1.1　“话语”的概念

言语交际是人类活动的重要组成部分。近年来，人们对于话语的研究已由语言学领域延伸到整个人文社会科学领域，呈现出跨学科、多领域的发展态势，取得了丰硕的成果。由于不同学科对话语研究的内容和侧重点不同，在各种论著中，对话语的定义也各有不同。

Schiffrin,Tannen & Hamilton（2001）将话语归纳为：（1）话语是任何大于句子的语言单位；（2）话语是语言的使用；（3）话语是更广泛的社会实践。

Renkema（1993）认为，话语既可以用于口头交际，也用于书面交际，口语和书面书之间存在重要差异。

Halliday & Hasan（1976）认为，话语不是语言形式单位，而是意义单位，它没有大小，长短之分，表达一个整体意义，具有明显的话语特征（texture）。

国外研究者用“discourse”或者“text”表示话语，国内则倾向使用“话语”“语篇”或“篇章”等术语来表示。就其本质来说，都是对具有交际功能的语义单位进行的研究。

话语在交际过程中产生，意义在交际过程中建构。读者或听者接收的信息与自己在线心理表征出来的信息相结合，不断形成新的信息，话语意义就是一个建构与再建构的动态过程。一般来说，话语是大于句子、具有一定交际功能的语义单位，或口头或书面，或短或长，无论它的体现形式是什么，只要合乎语法、语义连贯，具有一个中心论题，能完成一定的交际功能，就是话语。

Mann & Thompson 认为，话语具有三个特征：功能、层次和关系（徐赳赳，2010）。功能特征是指话语不同层次的单位都不是任意堆砌的，都是为实现交际目的服务的，话语的整体性和内部联系都源于功能性。层次特征是指：两个小句之间的语义关系是最低层次的，然后几个小句和几个小句之间的关系组成高一层次，最后由更大的语言单位之间组成整个话语。关系特征是指小句和小句，语段和语段之间存在各种各样的语义关系，绝大部分语义关系是不对称的，构成“核心”+“辅围”关系。

本书将“话语”定义为“连续的话段或句子构成的语言整体，是一段有意义、传达一个完整信息、前后衔接、语义连贯，且具有一定交际目的和功能的言语作品”。[1] 本书主要讲述话语计算的理论与应用，其中的话语主要是指书面语，所选语料均选自书面语。

1.1.2 话语语言学

话语语言学（textlinguistics）作为一门独立的语言学分支出现在 20 世纪六七十年代。它把实际使用中大于句子的篇章作为研究对象，因此也叫篇章语言学或语篇分析（discourse analysis）。纵使名称不同、使用术语不同、研究重点不同，但是其研究实质是一致的：探索连贯话语内部的构成规律（王福祥，1994）。

话语语言学（Text linguistics）这个术语最早在 1967 年由德国语言学家 Weinrich 提出后，话语研究就成为语言学领域的重要议题。荷兰语言

1 在这里，“言语作品”这种定义更倾向于 text，是偏欧洲学派的定义。

学家 van Dijk（1979a）强调“话语语言学”在本质上不是指代单一的理论或者研究方法，相反，它指把话语作为主要研究目的的任何语言科学研究（Beaugrande，1981：18）。

话语语言学发源于布拉格学派关于语言功能和意义的探讨，明确区分了话语（utterance）和句子，从而明确了话语语言学的研究对象。英国语言学家弗思（J. R. Firth）强调了语境对于话语意义理解的重要性；而后，Halliday 系统功能语法理论的建立以及和 Hasan 对于衔接理论的全面阐释，为早期话语分析提供了一个坚实的理论基础和分析框架（李佐文，张天伟，2006）。南朝时期，著名文学理论家刘勰于公元 501—502 年完成的《文心雕龙》是我国第一部文章学专著，在其第三十四篇《章句》中指出：“夫人之立言，因字而生句，积句而成章，积章而成篇”（黄国文，1987）。可见我国学者对话语现象的研究由来已久。

促使话语语言学形成的动力和原因主要有：（1）受语言研究中其他学科发展的影响，如计算机自然语言处理研究。这个领域的研究者发现，许多问题在句内无法解决，只有借助话语分析才能解决。（2）理论语言学自身发展的结果使人们把目光投向比句子更大的话语形式（徐赳赳，2010）。

与传统语法研究相比，话语语言学研究的对象是人们交际中使用的语言，所选语料都是自然发生的口头或书面话语。研究方法可以是定性或定量分析的方法，目的是探究话语表现的规律性和语言使用的倾向性，注重语境对话语生成和理解的影响、社会文化因素对话语方式和话语模式的制约作用等。

1.1.3　话语研究的角度

纵观国内外话语分析发展历程，从理论基础的构建到研究方法的融合，逐渐呈现出多视角跨学科的发展趋势，主要集中在以下几个角度。

（1）基于语料库的话语研究

语料库语言学是 20 世纪 50 年代 Chomsky 革命以来语言学领域在方法论方面最卓越的贡献之一（吕长竑，2010）。针对话语分析和语料库语言学均对现实生活中真实语言进行研究这一特点，研究者梳理语料库

话语分析产生背景、国内外现阶段发展，从而构建该分析方法的典型研究范式，即利用语料库对一些常见语法现象在语篇层面上的分布进行研究，分析其语篇功能（桂诗春等，2010）。语料库以自然发生的话语为内容，话语研究是以自然发生的交际话语为研究对象，二者间有天然的拟合性与兼容性。将语料库视角融入话语研究，一方面能为揭示话语意义提供丰富的语言例证和强大的分析方法；另一方面，语料库研究中的词语共现、语言特征共现等创新思路也为话语研究增添了理论维度（许家金，2019）。例如，以 Carter & McCathy（2006，1997，1995）和 Hughes（1998）为代表的英国诺丁汉学派利用 CANCODE 语料库对口语语法所做的系统的分析，对新兴语篇形式，如短信、电子邮件、网络聊天话语等进行的语料库语言学分析等都属于此类研究。

（2）批评话语分析

批评话语分析方法在 1979 年由 Fowler 等在其著作《语言与控制》（*Language and Control*，Routledge & K. Paul, 1979）中首次提出，旨在揭示语言形式和意识形态之间的相互作用，以及语言与社会结构和权力之间的关系（丁建新，2001）。法兰克福学派哲学家提出话语中主观力量所具有的变革性特点，认为具有理性的、逻辑的话语能够战胜扭曲的、含有晦涩意识形态意义，由此强调批评话语分析的重要性（田海龙，2008）。此类研究主要表现为采用法兰克福学派的批评理论和方法分析政治、媒体等语篇中的意识形态，解释其中的角色或身份定位等问题，如 van Dijk（1993）、Fairclough（2003)、Wodak（2006）等的研究属于该类型的研究。

（3）基于认知科学的话语分析

基于认知视角的话语研究主要表现为运用认知科学和认知语言学的理论成果对语篇进行分析。如：Langacker（2001）从认知语法的角度分析当前话语空间的推进模式，揭示语篇的认知操作过程。提出在运用认知语法理论进行语篇分析时，语篇层面的双极性、语言单位的连续性、动态分析和语篇期望以及突显原则（王寅，2003）。O'Halloran（2003）从连通论和关联理论角度探讨新闻语篇对读者的操控作用（支永碧，2007）。Lee（2001）把认知语言学中的“框架”“范畴”等基本概念运用到语篇分析中，证明认知语言学在篇章研究中的潜力（苗兴伟，2006）。

Ariel et al. 利用可及性理论、语篇世界和心理空间理论、向心理论等对语篇的理解、处理和指称照应等现象进行探讨。

（4）基于文化视角的话语研究

基于文化角度进行话语分析主要体现在国内研究立足于中国本土，放眼全球，研究中国话语的民族性特征，以提升中国话语在世界舞台的影响力。首先，以施旭为代表聚焦中国话语研究范式，探讨话语中所体现的中华文化，成为话语分析的文化转向（施旭，2008）。其次，从理论层面探讨中国话语体系的内涵和意义（张传民，2012；韩庆祥，2015）。再次，从多角度出发，论证中国特色话语体系的建构路径（孟威，2014；田鹏颖，2016）。最后，对中国话语体系建设过程中出现的难点进行深入探讨并提出相应的对策（杨鲜兰，2015）。但是现有对中国话语体系建设的研究多倾向于理论层面的论证，聚焦回归话语本身，进行实证的研究有待加强。

（5）基于生态语言学的话语分析

随着人类社会迅猛发展和生活水平的日益提高，人们越来越多地关注自身与自然、环境、资源等生态因素之间的关系，因此，生态学研究逐步和学科领域进行交叉，生态语言学就是生态学和语言学交叉而形成的新兴学科（黄国文，2016）。我国基于生态语言学的话语分析主要集中在对生态话语分析研究范式和理论建构的研究（何伟，魏榕，2018），鲜有研究从实证角度深入探索，这也是未来研究突破的方向。

1.2　AI 时期的话语研究

近年来，自然语言处理作为人工智能的重要领域之一，在词性标注、命名实体识别、句法分析等方面取得了丰硕的成果，但在话语层面的自动语义处理上还处于起步阶段，尚需加强。机器计算和存储能力大幅度提升，特别是以神经网络为基础的深度学习应用到自然语言处理，使计算机对语言的处理不再停留在词或句子层面，而是有能力处理话语层面的语义信息。随着互联网的飞速发展，以话语为呈现方式的海量信息全部依靠人工分析显然是不现实的，要实现大规模文本数据的自动语义处理，

必须搞清楚话语的语义特征、表征规律等等，因此面向自然语言处理的话语研究非常迫切。

早期的话语研究与古典修辞学、俄国的形式主义、法国的结构主义以及符号学有一定的渊源关系。20 世纪六七十年代，人文与社会科学的发展为话语分析提供了肥沃的土壤，民族学、结构主义与符号学、篇章语法、社会语言学、认知心理学、交际学等学科都开始关注话语问题。后来，心理学、心理语言学、人工智能等领域也都对语篇的记忆和理解过程进行探讨，形成了一些理论模型，为话语计算研究奠定了基础。人工智能迫切需要话语研究在以下几个方面有所进展。

首先，应该加强话语生成和理解过程中大脑的神经认知研究。话语是人们在社会交往过程中对语言系统的使用，它既是认知的对象又是认知的过程，表现为语言使用者如何感知、理解、记忆、评价语言单位，以及如何表达他们的交际意图。认知科学在解释话语的可计算性、话语意义表征、语篇知识和语境知识相互作用等方面做出了贡献。认知心理学在研究语言生成理解方面提出了很多模型，为话语计算奠定了基础。但是随着认知神经科学的发展，话语理解过程中的神经系统是如何操作的仍然是一个黑箱。深度学习过程中隐形层的神经元之间的函数计算是如何得到的仍不清楚。神经网络对于输入的信息会输出一个结果，但在高维空间的计算过程我们并不清楚。例如，人脑做决定是一种思维活动，至于决策是如何生成的以及决策过程是怎样进行的却难以解释。医生给病人看病，如果只给出诊断结果而说不清理由，病人是不敢接受治疗的。Alphago 能够在下围棋方面战胜人类，是因为它对围棋规则和过程非常清楚，根据情况不断变换和重组规则从而取得卓越的成效。让机器和人类一样用语言进行交流，提升它的智能水平，加强话语生成和理解过程的认知神经的研究非常重要。

其次，探寻话语层面语义表征的通用化模型是提高话语计算质量、机器理解人类自然话语的关键。近年来，搜索引擎、舆情监控、自动文摘等重大应用对话语的自动语义分析提出迫切需求。要想让计算机真正理解话语的意义，就必须研究话语的形式化结构。表层结构是深层语义结构的映现，体现着交际意图。建立一套既符合话语本质规律，又符合

自然语言处理应用的话语语义分析理论和方法无疑会更好地提升话语研究的应用价值。

话语是一个笼统的概念，其表现形式多种多样，将表层没有结构的文字序列转化成深层有结构的语义表征，刻画出各个部分之间的语义关联绝非易事。特别是探索适用于各种语类文体的通用型语义结构，并能使其融合话语内部信息和话外背景知识，进而更好地理解话语的主旨和意图，正是话语研究者在人工智能时代需要解决的重大问题。这个问题的解决，能使机器翻译更加有效地组织翻译结果，提升译文的连贯性和可读性，也对译文的质量评估大有益处。在自然语言理解方面，话语语义结构分析有助于加深理解各部分之间的语义关联，提高自然语言理解的全面性和准确性，因此深入探索面向语篇理解的结构化通用模型迫在眉睫。

最后，加强语料库和语料资源建设，尤其是标记篇章层面宏观语义关系的语料库建设。深度学习在自然语言处理方面的应用解决了很多问题，如词汇的形态问题、句法结构问题等，但标记宏观语义关系语料资源相当短缺。基于神经网络的深度学习依赖于大规模有标注信息的语料，在训练过程中学习和掌握话语的结构性特征，如连贯关系、语篇结构特征等。语料的规模和丰富程度对于机器学习起到至关重要的作用，能否适应各种不同种类、不同体量的数据将直接影响到训练结果。由于语料资源的匮乏以及语篇关系分析任务本身的复杂性，迄今为止，汉语语篇关系和结构识别研究处于初级阶段，在一定程度上制约了文摘自动生成、新闻智能写作等领域的发展。

人工智能的快速发展对语言学研究提出新的要求。话语的计算研究既是语言学自身发展需要，也是人工智能对语言学工作者提出的必然要求，是两个领域的最佳融合点。话语的计算研究成为新时代话语研究的重要领域和方向。人们的话语意图、主要观点、情感态度、舆论立场等只有从话语整体层面才能得以准确地获取和分析。如何做到让计算机真正理解人类话语，实现从表层结构到深层语义的映射，是认知智能领域亟待解决的问题。

1.3 话语计算：话语研究的新领域

1.3.1 什么是话语计算

人工智能时代拓宽了话语语言学研究的路径和领域，对自然语言处理提出了更高和更新的要求。它不仅要解决词汇、句法、语义的问题，还要解决跨越句子层面的意义问题，使机器能够像人类一样理解语言。

话语计算（discourse computing），或称面向自然语言处理的话语研究，指用计算机来处理大于句子单位话语的意义，是自然语言处理的一个高级层面，也是话语分析的一个角度，从这个意义上来讲，话语计算也叫做计算话语学（Computational Textlinguistics）。从内容来看，话语计算包括话语生成和理解的模型设计、话语特征的选择和滤取，也包括话语层面语料库标注、设计和建设等，为深度学习提供素材。

计算话语学是一门研究如何在语言学理论框架内用可计算的形式抽象概括出话语意义操作模型的学科，是用话语形式特征实现语义计算的处理过程。它主要涉及话语语言学、认知语言学和计算语言学，是人工智能研究的重要内容（李佐文，严玲，2018）。例如：

> （1）chatGPT 是一种全新的聊天机器人模型，（2）它能够通过学习和理解人类的语言来进行对话，（3）还能根据聊天的上下文进行互动，并协助人类完成一系列任务。（4）作为一个人工智能模型，chatGPT 不具有感知能力，也不能产生情绪，（5）因此它很难评价自己“爆火”的情况。（6）但是，随着人工智能技术的发展和广泛应用，它必将对人类社会生活带来深远影响。

这段话由六个句子组成，（1）至（3）句介绍 chatGPT 的作用与功能，其中（2）、（3）句是对（1）句的详述；（4）、（5）句说明它的不足，两句之间是因果关系；（6）句是对上述内容的转折，说明它的意义，转折的内容属于全段核心信息，构成核心句。找出了核心句，就能提取出这段话的主要内容。这个过程就是通过句间的连贯关系来实现语义计算的过程。

目前已建立的对自然语言处理的各类形式模型和实操过程中出现的许多问题都与语篇层面的语义计算息息相关。由此，计算话语学在为适应 AI 时代面对自然语言处理技术不断革新的挑战中孕育而生，同时也是话语语言学历史发展的必然趋势。随着当今科学技术的不断发展以及计算机运算能力和存储能力的极大提升，特别是基于神经网路深度学习应用的拓展，进一步为计算机处理话语提供了坚实的技术基础。然而，擅长工程和算法的人工智能领域专家由于对话语语言学理论和规律缺乏了解，使得计算话语学研究进展较为缓慢。计算话语学将成为现代话语语言学的重要研究领域和方向（李佐文，严玲，2018）。

1.3.2　计算话语学与其他学科的关系

计算话语学是在自然语言处理技术不断提升，话语语言学研究不断深入，人工智能从计算智能、感知智能向认知智能发展的背景下提出的概念，具有明显的跨学科特征。

计算话语学是计算语言学（自然语言处理）的一个研究领域。它将大于句子的语篇作为处理对象，将其规律性特征模型化，然后形成算法，来分析处理语篇的意义或语篇现象，从而达到机器模拟人类话语理解能力的目的。比如，语段是话语中的基本表达单位，它由连贯的语句组成，表达相对完整的意义，如何让计算机识别语段的边界，分析它的语义结构，如何生成一个语段来表达确定的意思等就是计算话语学研究的内容，当然，解决这些问题的挑战性还是很大的。

俞士汶认为应多层次认识计算语言学的处理对象（俞士汶，2003）。在计算语言学的研究中，通常把“语言”这个大对象分解为一些相对独立的“部分”或“层次”来“分而治之”。按照处理的语言对象的不同，又可以区分出不同的研究领域，比如以语音为处理对象的领域称为语音识别或语音合成，以词为处理对象的领域叫作词法分析，以语句为处理对象的领域叫作句法分析，以语篇为处理对象的领域叫做计算话语学，包括语篇理解和文本生成。针对不同的处理对象，计算机所采取的算法和策略也不尽相同，这些都可以作为相对独立的研究内容。不过，对于大单位的语言成分进行处理，是建立在对小单位的语言成分进行处理的

基础之上的，任何一个相对独立的研究部分都离不开也不应该离开将语言作为一个系统来考察这一整体背景（俞士汶，2019）。计算语言学在语音合成、语音识别、字识别、拼写检查和语法检查等应用研究领域取得了骄人的成就；建立了基于词汇语法、概率统计以及语义语用处理的语言形式模型（冯志伟，2011）。以上研究成果为计算话语学的产生与发展奠定了坚实的基础。

此外，计算语言学使用的一些算法如 n 元语法、依存语法、统计方法、深度学习等也可以用在话语计算过程当中。

计算话语学是话语语言学或话语分析研究的一个重要方向。它是面向自然语言处理的话语研究，要求对话语的结构规律、意义表征等以精确的、形式化的、可计算的方式呈现出来。话语的结构研究始终是广受关注的重要课题，伴随着话语语言学的发展历史。语篇的线性特征、连贯特征、层次特征、主题特征构成话语的可计算特征。国外的语言学家如 van Dijk 的宏观结构理论、Danes 提出的语篇推进模式、Mann & Thompson 的修辞结构理论、Rumellhart 的故事语法等极大推动了话语计算理论的发展。国内学者也一直注重篇章结构的研究，如郑文贞的《段落的组织》、廖秋忠的《篇章中的论证结构》、娄开阳的《现代汉语新闻语篇的结构研究》等，在现代汉语篇章结构研究方面取得了开创性的成果，为进一步深入研究打下基础。 但是还有很多问题需要探索，比如要计算现代汉语中语篇的连贯，到底需要多少种连贯关系？各类体裁的语篇中是否存在通用的语义结构？语段在形式和结构上有什么特征？语段和自然段落是什么关系，等等。寻找面向自然语言处理的话语本质性、普遍性规律，还有很多工作要做。

真正意义上的人工智能应该是基于语义的、可解释的智能。研究话语意义的表征规律是话语语言学的任务，也是对话语计算的贡献。话语意义是言语交际所传递的主要信息，它不是词语意义和句子意义的简单叠加，而是超越句子的语言复合体所展现出来的发话者的意图，以前学者们的研究大多把话语意义看作是心理的属性，在其形式化表征方面的研究相对薄弱，是制约自然语言处理、制约人机对话发展的瓶颈之一。

计算话语学以认知科学和认知心理学为基础。认知科学为话语的可计算性提供理论依据。语言使用的过程是一种认知的过程，表现在话语使用者感知、理解、记忆、计划、表达他们的交际意图上，从这个意义上来讲，话语既是认知的对象，又是认知的过程。从认知的角度研究话语，更能透过语言的表层现象，揭示出语义操作的本质性规律。西蒙（H. A. Simon）曾经指出，人所有的智能活动都可以用符号和计算来实现。 在认知科学中，计算是一个广义的概念，是“依赖形式特征对语义进行主动控制”的过程。话语理解的过程就是一种计算过程，涉及计算这个概念的三个主要核心要素：话语形式、话语意义、主动控制。基于可计算理论的认知主义话语观将话语理解看作是一种可控制的符号系统及其操作的过程，为话语的可计算性提供了理论基础。

心理学，尤其是认知心理学为话语计算提供理论模型。要想让机器像人类那样去理解话语，必须以人的认知方式作为基础，研究人在各种情景下运用话语进行交际时的言语表现，弄清楚理解话语时所采用的策略和步骤，以及如何进行推理等等，在总结这些规律的基础上，提出心理模型，然后用计算机进行模拟。话语生成和理解是一个复杂的心理过程，没有心理学家的反复试验，很难提出可操作、可解释的计算模型。有些心理学家如奎林（M. R. Quillian），他们的研究为人工智能做出了很大贡献，也为话语计算提供了可借鉴的模型。奎林认为，人们说话或理解别人的话语，都离不开记忆，存储记忆之中的一些概念和概念之间的关系。这些概念及其关系构成一种复杂网络，就是“语义网络”，在语义网络中，概念用节点表示，节点之间的连线表示概念间的关系，形象地表征了语义记忆的结构和工作原理。语义网络是作为人的记忆的心理学模型提出来的，却为计算机模拟提供了基础。若干年来，人们对语义网络理论一直感兴趣，几经修改和补充，现已成为研究计算机理解自然语言的重要途径之一（陈永明，2013）。

可以看出，计算话语学是基于认知、面向计算的话语研究，是可解释的智能不可逾越的重要一步。它是根植于话语语言学、计算语言学和认知科学基础上的二代交叉学科，如图 1-1 所示。

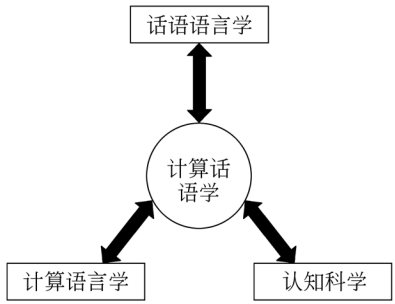

图 1-1 计算话语学与相关学科的关系

它们都关注话语的生成与理解的过程。例如：指称的确认、句子间的连贯关系、作者的观点和态度、话题的确定等。话语是人们对语言符号的实际使用，涉及语境的问题、背景知识、社会文化等因素，语言本身也是灵活多变的，一个词语在不同场合意思完全不同。另外，人们对语言理解的机制缺乏深刻的认识，致使话语计算这一领域具有很大的挑战性，很少有所触及。随着认知智能时代的到来，话语计算成为可解释、可推理智能的重要环节，在自然语言处理中发挥重要作用。

1.3.3 话语计算的应用

话语计算是对大于句子的语言单位进行语义的计算和处理。在当今互联网时代，人们通过短信、邮件、微博、电子报刊等话语形式实现信息的沟通交流，它们是承载信息的主要形式，且体量庞大。要想在短时间内获取想要的信息，或对信息进行过滤和筛选，用人工的方法显然是不现实的，因此对话语进行自动的计算处理显得尤为重要。话语计算可用于以下几个方面。

（1）**自动文摘** 自动文摘就是利用计算机自动地从原始文本中提取全面准确反应该文档中心内容的、简单连贯的短文。计算话语学不是用简单统计的方法或提取关键词的方法来组构文摘，而是根据文本内部语句的连贯关系确定核心句，依据权重按照要求提取中心内容，整个过程是基于语义的计算，有依有据，可解释性强。

（2）**事件抽取** 自然语言处理中的事件抽取也是需要在语篇层面进行处理的一项工作。事件是发生在某个特定的时间点或时间段，在某个

特定的地域范围内，由一个或者多个角色参与的一个或者多个动作组成的行为状态。事件抽取指从自然文本中抽取出感兴趣的事件信息并以结构化的形式呈现出来，如什么人、什么时间、在什么地方、做了什么事。

涉及内容包括事件描述（event mention）、事件触发词（event trigger）、事件元素（event argument）、元素角色（argument role）等。语言学家应该对事件的特征以及表征事件的语言特征进行总结和归纳，建立事件或事件句的模板并进行形式化描述，明确构成目标信息的上下文约束环境等。

（3）**体裁识别**　文本按体裁自动分类属于形式分类的范畴，对于信息分类检索具有重要意义。体裁识别的首要问题是分类特征的选取，这些特征不仅与文本内容有关，也与话语的社会功能有关，是话语的功能变体和社会规约，如何识别、描述、计算这些体裁特征确实是一项具有挑战性的工作，需要话语语言学和计算语言学学者的通力合作。

（4）**主题计算**　话语的主题计算是在话语整体层面上进行的，是对概念意义的抽象和概括，是计算话语学的主要内容之一。主题意义是从整合话语的各个组成部分中提取出来的能够涵盖整个语篇内容的主旨，而不是各个句子或段落意义的简单叠加。对于主题意义的计算目前人工智能的主要方法是通过词汇频率统计的方式，或 TF-IDF 的方法先确定话语中的关键词语，这种方法简便快捷，但回答不了“为什么”这样的问题。通过计算它的连贯关系找出权重大的句子，再汇集成连贯的语句的计算方法，可以增强问题的可解释性。

（5）**情感分析（sentiment analysis）**　也叫作观点挖掘（opinion mining），是从自然话语中提取观点和情感信息的研究领域。随着信息社会的发展，网络社交媒体，如论坛、微博等在人们的生活中发挥着越来越重要的作用，他们表达的意见和观点在现代社会中已经成为重要的决策参考信息。自 2000 年以来，情感分析已成为自然语言处理领域最为活跃的研究问题之一。从话语语言学的角度探讨观点和态度问题，可以更清晰地了解这一问题的基本结构和常用于表达观点和情感的语言方式。然而目前的计算机还没有和人一样的理解能力，很多的语言学知识还不适用于计算机处理。计算话语学就是面向机器的实际需求，探讨从自然

话语有关观点、情感及相关概念的系统。

此外，话语计算还可以应用于信息检索、学生作文自动评测、热点话题追踪等领域。

思考与讨论

1. 计算话语学和计算语言学的区别与联系。
2. 根据话语的特点，谈一谈话语计算的复杂性，它的难点在哪里？
3. 人工智能时代的话语研究涉及哪些方面？
4. 话语计算作为人工智能的重要领域，可以应用到哪些方面？

参考文献

[1] Ariel, M. et al. Accessibility marking: Discourse functions, discourse profiles, and processing Cues [J]. *Discourse Processes*. 2004, 37 (2): 91-116.

[2] Beaugrande R., W. Dressler. *Introduction to Text Linguistics* [M]. London: Longman, 1981.

[3] Carter, R. Orders of reality: CANCODE, communication, and culture [J]. *ELT Journal*, 1995, 52 (1): 43-56.

[4] Carter, R. & M. McCarthy. *Exploring Spoken English* [M]. Cambridge: Cambridge University Press, 1997.

[5] Carter, R. & M. McCarthy. Grammar and the spoken language [J]. *Applied Linguistics*, 1995, 16 (2): 141-158.

[6] Fairclough, N. *Analysing Discourse: Textual Analysis for Social Research* [M]. London: Routledge. 2003.

[7] Halliday, M. A. K. & H. Ruqaiya. *Cohesion in English* [M]. London: Longman, 1976: 1.

[8] Hughes, R. & M. McCarthy. From sentence to discourse: Discourse grammar and English language teaching [J]. *TESOL Quarterly,* 1998, 28 (2): 263-287.

[9] Renkema, Jan. *Discourse Studies: An Introductory Textbook* [M]. Amsterdam: John Benjamins Publishing Company, 1993.

[10] Schiffrin, Tannen & Hamilton (ed.). *The Handbook of Discourse Analysis* [M]. London: Blackwell, 2001.

[11] van Dijk, T. A. Principles of critical discourse analysis [J]. *Discourse & Society.* 1993, 4: 249-283.

[12] Walker, M. A., A. K. Joshi & E. F. Prince. *Centering Theory in Discourse* [M]. Oxford: Clarendon Press, 1998: 1-28.

[13] Werth, P. *Text Worlds: Representing Conceptual Space in Discourse* [M]. London: Longman. 1999.

[14] Wodak, R. Dilemmas of discourse (analysis) [J]. *Language in Society*. 2006, 35: 595-611.

[15] 陈永明. 言语与智能 [M]. 北京：北京师范大学出版社，2013.

[16] 陈忠华，杨春苑，赵明炜. 批评性话语分析述评 [J]. 外语学刊，2002（01）：182-86.

[17] 单胜江. 新闻语篇的批评性话语分析 [J]. 外语学刊，2011（06）：78-81.

[18] 丁建新，廖益清. 批评话语分析述评 [J]. 当代语言学，2001（04）：305-310.

[19] 冯志伟. 计算语言学的历史回顾与现状分析 [J]. 外国语（上海外国语大学学报），2011，34（01）：9-17.

[20] 桂诗春，冯志伟，杨惠中，何安平，卫乃兴，李文中，梁茂成. 语料库语言学与中国外语教学 [J]. 现代外语，2010，33（04）：419-426.

[21] 韩庆祥. 中国话语体系的八个层次 [J]. 社会科学战线，2015（03）：1.

[22] 何安平. 基于语料库的英语教师话语分析 [J]. 现代外语，2003（02）：161-170.

[23] 何伟，魏榕. 话语分析范式与生态话语分析的理论基础 [J]. 当代修辞学，2018（05）：63-73.

[24] 胡江. 意义单位与批评话语分析——基于语料库的西方媒体涉华军事报道意识形态分析 [J]. 解放军外国语学院学报，2016，39（05）：73-81.

[25] 黄国文. 生态语言学的兴起与发展 [J]. 中国外语，2016，13（01）：1+9-12.

[26] 黄国文. 语篇分析概要 [M]. 长沙：湖南教育出版社，1987：6.

[27] 李佐文，李楠. 新闻话语的可计算特征 [J]. 现代传播（中国传媒大学学报），2018，40（12）：41-44.

[28] 李佐文，严玲. 什么是计算话语学 [J]. 山东外语教学，2018，39（06）：24-32.

[29] 李佐文，张天伟. 话语语言学是语言学历史发展的必然——中国话语语言学研究会成立大会暨首届全国话语语言学学术研讨会会议综述 [J]. 外语研究，2006（03）：16-19.

[30] 吕长竑. 语篇的语料库研究范式评介 [J]. 外国语（上海外国语大学学报），2010，33（02）：35-43.

[31] 孟威. 改进对外传播 构建“中国话语体系”[J]. 新闻战线，2014（07）：82-85.

[32] 苗兴伟. 语篇分析的进展与前沿 [J]. 外语学刊，2006（01）：44-49.

[33] 邵斌，回志明. 西方媒体视野里的“中国梦”——一项基于语料库的批评话语分析 [J]. 外语研究，2014（06）：28-33.

[34] Simon H. A. 认知科学的新进展 [J]. 心理学报，1991，23（02）.

[35] 施旭. 话语分析的文化转向：试论建立当代中国话语研究范式的动因、目标和策略 [J]. 浙江大学学报（人文社会科学版），2008（01）：131-140.

[36] 唐青叶，史晓云. 国外媒体“一带一路”话语表征对比研究——一项基于报刊语料库的话语政治分析 [J]. 外语教学，2018，39（05）：31-35.

[37] 田海龙. 语篇研究的批评视角 [J]. 外语教学与研究，2008（05）：339-344+400.

[38] 田鹏颖. 在解构“西方话语”中建构中国话语体系 [J]. 马克思主义研究，2016（06）：138-145.

[39] 王福祥. 话语语言学的兴起与发展 [J]. 外语与外语教学，1994（04）：3-10.
[40] 王寅. 认知语言学与语篇分析——Langacker 的语篇分析观 [J]. 外语教学与研究，2003（02）：83-88.
[41] 徐赳赳. 现代汉语篇章语言学 [M]. 北京：商务印书馆，2010.
[42] 许家金. 语料库与话语研究 [M]. 北京：外语教学与研究出版社，2019.
[43] 杨鲜兰. 构建当代中国话语体系的难点与对策 [J]. 马克思主义研究，2015（02）：59-65+159.
[44] 余凯，贾磊，陈雨强，徐伟. 深度学习的昨天、今天和明天 [J]. 计算机研究与发展，2013，50（09）：1799-1804.
[45] 俞士汶. 计算语言学概论 [M]. 北京：商务印书馆，2003.
[46] 张传民. 文化自觉、理论自觉与中国话语体系的建构 [J]. 山东社会科学，2012（10）：183-187.
[47] 支永碧. 批评话语分析研究新动态 [J]. 外语与外语教学，2007（03）：27-32.