多轮对话的篇章级抽象语义表示标注体系研究*

黄彤[1]，李斌[1]，闫培艺[1]，计婷婷[1]，曲维光[2]

1.南京师范大学　文学院，江苏　南京
2.南京师范大学　计算机科学与技术学院，江苏　南京
huangtong_njnu@126.com, libin.njnu@gmail.com, ypyheta@gmail.com
ting@163.com, wgqu@njnu.edu.cn

摘要

对话分析是智能客服、聊天机器人等自然语言对话应用的基础课题，而对话存在大量情感短语、省略、语序颠倒等现象，对句法和语义分析器的影响较大，对话自动分析的准确率相对书面语料一直不高。其主要原因在于缺乏严整的多轮对话形式化描写方式，不利于后续的分析计算。因此本文在梳理国内外针对对话的标注体系和语料库的基础上，提出了基于抽象语义表示的篇章级多轮对话标注体系，探讨了篇章级的语义结构标注方法，给出词语和概念关系的对齐方案，为称谓语和情感短语增加了相应的语义关系和概念，调整了表示主观情感词语的论元结构，并规定了对话中一些特殊现象，设计了人工标注平台，为大规模的多轮对话语料库标注与计算研究奠定基础。

关键词： 抽象语义表示；多轮对话；标注体系；语义计算；标注体系

Research on Discourse-level Abstract Meaning Representation Annotation framework in Multi-round Dialogue

Huang Tong[1], Li Bin[1], Yan Peiyi[1], Ji Tingting[1], Qu Weiguang[2]

1.School of Chinese Languageand Literature, Nanjing Normal University
Nanjing, Jiangsu, China
2.School of Computer Scienceand Technology, Nanjing Normal University
Nanjing, Jiangsu, China
huangtong_njnu@126.com, libin.njnu@gmail.com, ypyheta@gmail.com
ting@163.com, wgqu@njnu.edu.cn

Abstract

Dialogue analysis is the basic topic of natural language dialogue applications such as intelligent customer service and chat robots. The dialogue corpus is quite different from the regular written corpus. There are a number of complex phenomena such as vocatives, emotional phrases, omissions, word order reversal, redundancy, etc. Compared with the semantic analysis, the accuracy of automatic dialogue parser has been relatively low compared to the written corpus. The main reason is that the lack of rigorous formal description of multiple rounds of dialogue is not conducive to subsequent analysis and quantitative research. Therefore, we make a survey onthe tagging system and corpus for dialogues, then proposes a discourse-level multi-round dialogue tagging system based on abstract meaning representation. It specifically discusses the discourse-level semantic structure annotation method, gives the alignment scheme of

基金项目：国家社科基金项目(18BYY127)、国家自然科学基金(61772278)、江苏省高校哲学社会科学优秀创新团队建设项目。

words and concept relations, adds corresponding semantic relations and concepts for appellations and emotion phrases, adjusts the argument structure of subjective emotion words, and some special phenomena in the dialogue are stipulated, and a manual annotation platform is designed to lay the foundation for large-scale multi-round dialogue corpus annotation and quantitative research.

Keywords: abstract meaning representation , multi-round dialogue , annotation scheme , semantic computing , chinese information processing

1 引言

近年来，伴随着人工智能的浪潮，问答系统、智能助手、聊天机器人等成为了研究的热门，人们希望机器能够像人一样思考，与人类对话，这就要求机器要能够理解、处理人的对话内容，对话分析是自然语言对话应用的基础，口语对话的分析逐渐受到重视 (宗成庆,1999)。

但就目前而言，多轮对话的篇章分析仍存在问题：首先，目前对话语义分析往往以处理普通文本的方式分析对话，导致自动分析效果较差。语义分析大多仍处在相对规范的书面文本的层面上，口语对话不同于书面语，对话中存在更多省略、语法不规范等现象，并且分析口语对话不再局限于单句的分析，需要考虑上下文的信息，这些都给机器自动分析对话增加了难度。Adams (2017) 尝试使用不同模型对对话语料进行依存解析，得到的F值仅有85.7%和80.3%（带依存关系的评测方式LAS），而常规语料均能达到90%以上，存在一定差距，对话解析的效果不甚如意。其次，多轮对话缺乏整体的篇章表示体系和语料建设。目前的语料库资源大多是以书面语料为主，专门针对对话的语料较少，而面向对话的语料库和语料标注规范的研究主要集中在对话行为、篇章关系等特定领域，一般只标注说话人、话轮信息、词性标注或句法分析结构，而忽视话轮间应答关系、话轮内部小句的关系，以及省略恢复、指代消解等难题。对话在篇章层面上的语义结构、应答逻辑没有得到有效的研究和表述。因此需要高质量的口语对话资源以推动语义理解模型的发展 (郑桂东, 2018)。

本文提出了一种针对对话的语义表示方法——对话抽象语义表示（Dialogue Abstract Meaning Representation，DAMR），来解决篇章级多轮对话的语义表示问题。这个方法基于中文抽象语义表示（CAMR）改进而来。抽象语义表示（AMR）作为一种新兴的句子语义表示方法，采用单根有向无环图来表示句子的语义结构 (Bonial et al., 2013)，能够有效解决句子中的论元共享、省略、冗余、语序错乱等难题，并进行了多语言的理论和计算实践 (Oepen et al., 2019)，标注了上万句英文语料[1]和汉语语料[2]。不过，AMR 虽然已经能较好地表示句子语义，但由于对话语料和常规书面语存在较大差异，例如省略（省略主语、宾语）、独立的称呼语、情感短语（如“哈哈哈”）、冗余等，且目前CAMR仅针对单句进行了标注，而对话标注必然是篇章级别的，因此不能直接套用CAMR的规范来表示中文对话的语义，需要根据对话特点对CAMR的框架和规范进行调整、改进和扩充，使之能够表示多轮的对话语料。

因此，我们提出DAMR（Dialogue Abstract Meaning Representation，对话抽象语义表示）继承了CAMR的框架和理论，DAMR是一种针对中文对话的篇章级句子语义表示方法，DAMR从4个方面进行了改进：（1）改进概念关系对齐的语法，将篇章信息其融合到语料标注中；（2）针对对话特点，增加概念标签和关系标签；（3）调整了部分词语的论元结构；（4）对一些对话中的称呼语、情感短语特殊现象进行了规定。

全文结构如下：第2节总结了国内外对话语料的标注体系和方法，第3节介绍了数据来源和AMR标注体系，第4节介绍了DAMR针对对话做出的改进，第5节是结论和未来工作。

2 相关工作

专门针对对话的标注体系和语料库较少，由于主要面向智能机器人，因此多为限定领域（例如旅游行程制定、地图导航、智能音箱）的标注，且标注重点在于对话行为（反应说话者的意图、话语的结构）、篇章关系（句子之间的关系）等，而完整的对话标注体系还需要包括对话的基本信息（说话人编号、话轮等）、指代信息（共指和回指）、句法语义信息（词类、句法结构、语义）等。表1给出了下文提到的对话语料库的标注内容。

[1]https://catalog.ldc.upenn.edu/LDC2020T02
[2]https://catalog.ldc.upenn.edu/LDC2019T07

语料库	对话基本信息	指代信息	篇章结构	句法语义信息	对话行为
LUNA(2007)		✓		✓	✓
MATE(1999)		✓		✓	✓
Martinez(2002)	✓			✓	✓
ISO24617-2(2010)	✓				✓
Zhou(2010)	✓				✓
周小强(2018)	✓		✓		✓

表 1. 语料库标注信息

2.1 对话行为信息标注及语料

多层对话行为标注（Dialogue Act Markup in Several Layers，DAMSL）是应用最为广泛的一个面向任务的通用领域的标注体系，DAMSL在四个维度上对对话行为进行标注，包括：交际状态（Communicative Status）记录话语是否完整，信息层面（Information Level）标注话语的特征，向前功能（the Forward Looking Function）记录当前话语与之后话语的联系、向后功能（the Backward Looking Forward）记录当前话语与之前话语的联系 (ALLEN, 1994)。在DAMSL 提出之后，一部分学者使用DAMSL 标注体系对语料库进行标注，其中最为出名的是Switchboard（SWBD）电话语料库，其目的是进一步提高自动语音识别的语言模型 (Jurafsky et al., 1997)。MRDA （Meeting Recorder Project）对话标注体系则是在SWEB-DAMSL的基础上修改的标注体系，用于标注ICSI（International Computer Science Institute）项目的英语会议多人对话内容，形成了ISCI-MRDA 语料库 (方称宇，2013)。

Bunt (2010)认为DAMSL的维度存在模糊性，提出了一种新的体系DIT++。DIT++细分若十个维度，如活动行为、交际管理、话轮管理等，并规定了每个维度下的交际功能，设计了两类标签：通用目的功能（general-purpose functions）和特定维度功能（dimension-specific functions）标记集，两个标记集下又分多层多个标签。DIT++体系已应用于多个语料库中，如DIAMOND人人对话库、面向任务的AMI人人对话库等 (方称宇，2013)。随着对话行为标注体系的不断发展，Bunt (2010)等人根据以DAMSL和DIT++等多个对话行为标注体系的特点，集各家所长提出多维度的对话行为标注国际标准：ISO24617-2，借鉴DIT++设定了九大维度，包括任务、自我反馈、启他反馈、话轮管理、时间管理、社会义务管理、自我交际管理、语篇构建等，各个维度下设计了相应的对话行为标签。除了通用领域的对话行为标注，还有部分针对特殊领域的标注语料库，如美国基于查询铁路交通的人机对话语料——TRIANS (Allen and Core, 1997)、查找路线的人人对话语料——英国HCRC语料 (HCRC group,1996)。这些标注体系都根据各自的语料特点设定了限于该领域的相应的标签。

汉语对话行为标注随着国外DA的发展，也开始受到重视，但对此的研究仍然有限。王珊等 (2016)建立了一个电视台访谈节目语料库，基于国外对话行为的研究，通过对语料库中的问答句子的分析，设计了汉语的单层级的对话行为的类别。周强 (2017)基于国外DAMSL、SWBD-DAMSL等标注体系，设计了五大标记集，各个标记集下面再分不同标记，并借鉴了ISO标准中的维度设计。

我们认为，在同一句话中对话行为可能包括多个，而说话人的意图有时也无法体现，对话行为也体现不出来，同时，AMR可根据原有的语义关系标签根据语义表示相应的意图或语用功能，例如语气“mode”标签可以表示说话人“祈使”、“询问”等意图，因此DAMR暂时不引入对话行为的标签，更注重使用原有体系表达说话者的实际语义。

2.2 篇章关系信息标注及语料

对话中篇章关系标注主要沿用宾州篇章树库（Penn Discouse TreeBank，PDTB）、修辞结构篇章树库（Rhetorical Structure Theory Discourse Treebank，RST）两大体系。

PDTB仅考虑相毗邻的句子之间的关系，借鉴了谓词论元结构，以连接词（connective）为核心分别定义了两个论元arg1和arg2，连接关系包括显性关系（Explicit）、隐性关系（Inplicit）、替代关系（AltLex）、实体关系（EntRel）、无关系（NoRel），如果没有显性的连接词，标注人员要根据自己的判断表示出其连接关系，同时设定了多层多类语义关系标

签 (PDTB-Group, 2009)。Sara (2010)等人将PDTB体系用于LUNA口语对话语料库中，针对对话语料的特征对意义标签进行了调整。Xue等 (2016)也同样将PDTB体系的用于标注SMS短信息对话，根据信息对话的特点对标签进行增删。

修辞结构理论（RST）将篇章关系称为修辞（rhetorical）关系，设定了两种修辞关系：单核心和多核心，修辞关系所连接的篇章单位如果存在主次区别，那么就是单核心关系，反之就是多核心关系。RST 与PDTB最大区别在于其篇章结构树有层次，每个修辞关系都可以连接两个或多个篇章单位，这些篇章单位又可以组成大的篇章单位和其他篇章单位形成修辞关系，最终一个篇章形成一个有层次的篇章结构树 (Carlson et al., 2001)。Stent (2000)首次将RST 用于标注面向任务的对话语料中，针对对话或是标注领域特有的特点，新增了一些修辞关系（如问答关系），并为某些范围过于广泛的标签设置了更具体的下级标签。

中文AMR中规定了10种篇章关系，我们将沿用这些关系来标注对话中的篇章关系，因为对话话题较为分散，因此存在篇章关系的两个或多个句子不仅局限于相毗邻的两个句子中，同时也会根据对话的实际特点增加相应的篇章关系标签。

2.3 综合信息标注语料库

综合标注的对话语料库指标注了多种信息的语料库，包括上文提到的对话行为、篇章结构，还有语义信息、共指等信息。

LUNA语料库是一个跨语言（意大利语、波兰语、法语）、跨领域的人人、人机对话语料库，采用了层标注，第一层为语义标注，第二层为领域属性标注，以及非必须的其他层，包括谓词结构框架、对话行为、指代信息等 (Raymond, 2007)。如图1，领域属性标注层标注句子中的语义块所属的领域及其属性，以“属性-值”对构成，语义块来自第一层的语义标注；谓词结构框架借鉴了FrameNet 框架标注语义结构，为预先设定好的领域设定框架；再填入相应的元素。对话行为沿用DAMSL 体系标签；同时，LUNA 语料库标注了共指信息，将可标记共指的元素标记为given或new，如果标为given，则找出最近发生的对象并增加指针指向它。

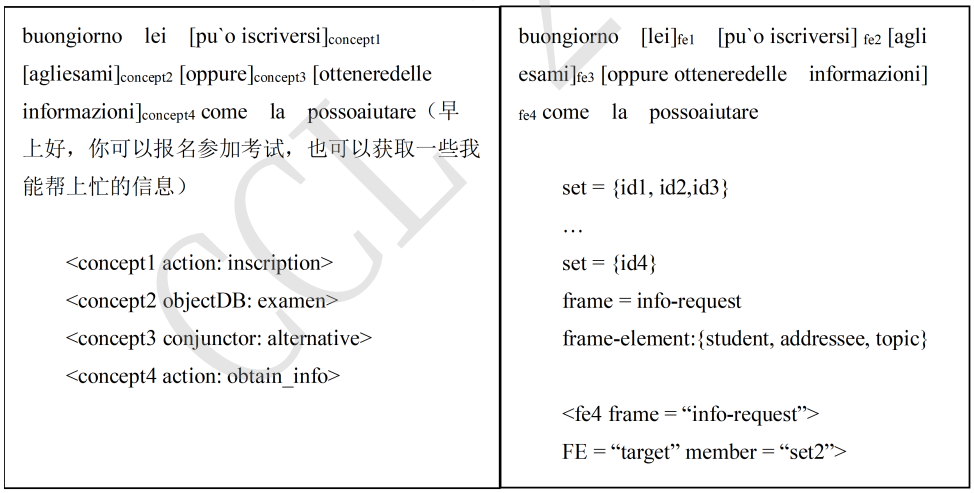

图 1. LUNA标注方法示例

其他还有较为知名的语料库还有Martinez (2000)在铁路信息系统的对话语料上标注了三层标签，分别为对话行为、框架（Frames）和实例（case），对话行为基于TRAINS体系的标签进行了调整；框架借用FrameNet的思想，为具体任务设置相应框架；实例则用来填充框架的槽；MATE 语料库标注了语义信息、对话行为、共指信息 (Poesio et al.,1999)；Zhou (2010)建立了一个汉语的旅游领域的语料库，共标注了十三层信息，包括话轮、主题、说话者信息、分词词性信息、拼音、语音转录、语音边界、句子重音、音量、非语言信息、基于ICSI-MARA体系的对话行为、形式错误信息、情绪。

LUNA、MATE、Martinez建立的语料库都是面向任务的语料库和标注方法，因此其意义标注仍是从对话行为出发，更注重抽取出说话者所要实现的功能意图，再根据意图设定论元结构，无法完整地表示句子的语义。在指代标注上，LUNA等其他标注共指的语料库都只涉及名词，将上下文中共指的元素用同样的id连接依赖，忽略了指向一个完整事件的代词，因此不利

于判断指示词和先行语之间的关系和指代消解的实现。另外，这些语料库的重点仍然是单句的语义标注，没有将有相应问答或其他对应关系的句子表示出来。

周小强 (2017)等人设计了一个交互式问答语料的关系结构标注体系。除标注了对话行为类别外，还标注了问答中的语义匹配关系和语义补充关系。其对应关系只限于句子和句子之间的关系，但在实际语料中情况更为复杂，有补充和匹配关系的不一定为整个句子，可能只是句子中的一部分，因此这种方法存在问答点对应不明确的问题。

3 数据来源及AMR体系介绍

3.1 数据来源

我们在改进的中文抽象语义表示标注平台上试标部分中文短信息SMS对话语料[3]以分析对话标注存在的问题。该语料总共有15000篇对话，我们从中选取了10篇对话、475个句子进行试标注，语料信息包括话语编号、说话人编号、时间信息。我们对其进行预处理，增加了话轮编号信息。选其作试标语料主要因为：短信对话保留了日常对话的基本要素和特征，同时避免了肢体语言或现实环境语境对录音转写语料内容的影响。

3.2 AMR体系

抽象语义表示（Abstract Meaning Representation, AMR）是一种新兴的语义表示方法，它用单根有向无环图来表示句子语义，将句子中的实词抽象为概念节点，实词之间的关系则抽象为带有非核心语义关系标签的有向弧，忽略了虚词和一些较虚的语义（冠词、时态、单复数），允许增加、删除或修改概念 (Bonial et al., 2013)。

在这个基础上，O'Gorman (2018)等人提出了标注多句AMR（Multi-sentence AMR, MS-AMR），即将AMR拓展到篇章层面，但只关注了篇章中的共指现象，标注了名词、动词、代词、隐形角色的共指关系，MS-AMR设定了三种共指关系：一致关系、部分-整体关系、成员-集合关系。Bonial等人 (2020)针对人机对话语料对AMR 进行了改进，构建了对话AMR体系，主要有以下几点扩充：在AMR的最上层设定了36个对话行为标签；增加了时、体标签；针对该人机对话语料的用途设定了空间参数。

李斌 (2017)在AMR体系的基础上提出融合概念对齐的一体化标注方案，针对汉语特有现象进行了改进，形成了中文抽象语义表示方法（Chinese Abstract Meaning Representation）。CAMR的改进如下：为量词、时、体新增了语义关系标签；还原重叠式，如"试试"还原为"试"；组合离合式，如把"睡一会觉"合成概念"睡觉"；为复句关系增加了关系概念标签。

使用CAMR能够更完整、合理地标注对话：

第一，AMR关注的并非句子中的具体词语，而是句中抽象的概念和关系，允许增加、删除或修改概念，利用这个特点，可以在一定程度上解决对话中的倒序、冗余等情况，也可以对对话中省略的概念进行恢复（如图2中的"整治"），将话语中的语义合理地表示出来。

第二，CAMR进行了对齐改进，采用句中的序列进行编号，实现了概念与句中单词的对齐，有利于合理地表示指代、省略等情况，也有助于照应语和先行语之间关系的标注。

第三，CAMR为复句关系添加了并列、因果、让步、条件、转折、解释说明、选择、目的、递进、时序等10个标签，如图2中的并列复句关系"and"。DAMR可用这10 个复句关系标签表示对话篇章结构关系。

第四，CAMR新增的dcopy和refer用来标注两个概念之间的关系，有助于省略和指代照应的标注。

但CAMR标注对话存在一些问题，比如目前CAMR仅标注单句，复句关系仅限于同一句中的标注，一些在句中没有充当明确成分的词语无法标注（称呼，叹词)等。因此，我们在CAMR的基础上进行了改进，具体标注方法在第4节中说明。

4 对话AMR标注体系

我们在改进版的中文抽象语义表示标注平台上试标注了500句中文短信息SMS对话语料，尽可能在现有CAMR体系的基础上标注，同时根据标注对话遇到的问题对其进行调整，以期扩充CAMR的兼容性。对话中会有一些特有的成分，例如称呼、表示情绪的成分，存在指代照应

[3]https://catalog.ldc.upenn.edu/LDC2016T13

运河[1] 的[2] 整治[3] 改善[4] 了[5] 该[6] 县[7] 的[8] 投资[9] 环境[10]，[11] 吸引[12] 了[13] 外商[14] 投资[15] 。[16]

x18/and
  :op1() x4/改善-01
    :arg0( ) x3/整治-01
      :arg1(x2/的) x1/运河
    :aspect( ) x5/了
    :arg1( ) x10/环境
      :mod( ) x9/投资
      :poss(x8/的) x7/县
        :mod( ) x6/该
  :op2( ) x12/吸引-01
    :arg0( ) x3/整治-01
    :arg1( ) x15/投资-01
      :arg0( ) x14/外商
    :aspect( ) x13/了

图 2. CAMR示例

的距离较远的现象，话轮间问答不直接对应，出现大量的省略，诸如说话人/听话人人称代词的省略。因此针对对话的特点在以下几点对CAMR做出了改进：实现双层标签概念对齐、增加若干个标签、修改部分词语的论元结构、并规定了一些对话中的特殊现象的标注。

4.1 概念对齐

DAMR的每个句子都包含以下字段：语篇编号、话轮编号、句子编号、说话人编号、问答位置信息（见表2）。

语篇编号	话轮编号	句子编号	说话人编号	句子	问答位置信息
3	48	83	151460	过[1] 一会[2] 和[3] 你[4] 说[5]	-
3	49	84	131525	好[1]	-

表 2. DAMR语料字段

前5个字段根据语料顺序自动分配，问答位置信息由人工标注，标注答句所对应的问句的位置信息，如果非问答对应则不标注(本文其他例句如不涉及问答则省略该字段)。

CAMR的概念标签采用xn的形式，n是根据输入的已分词的原始句子序列分配的有序编号。人工补充的概念则由标注系统分配随机编号。目前的CAMR的编号仅适用于单句，无法跨句子进行标注，因此为了实现篇章级别的标注，DAMR 采用了双层编号，即用sn_xn来对齐句中的概念，sn根据输入的句子序列分配，xn 则仍旧根据词语在句中的序列分配。样例见图3。

s83_x5/说-01
  :arg3(x3) x4/你
  :time() x1/过-01
  :arg1() x2/一会
 	
s84_x1/confirm
  :arg1() s83_x5/说
	x5/说-01
	  :arg3(x3) x4/你
	  :time() x1/过-01
	  :arg1() x2/一会
	
	x1/好-01
	  :arg0() x3/说

图 3. DAMR/CAMR概念对齐示例

为减轻标注人员的操作量，当句子只出现当前句子的概念，则仅使用xn标签，当出现其他句子的概念时，才完整表示sn_xn。

4.2 新增标签

DAMR沿用了CAMR的5个核心语义关系标签、44个非核心语义关系标签和109个专名概念。argx （x∈[0,4]）表示核心语义角色关系，每个谓词的每个义项都有自己的核心语义角色框架。非核心语义关系是指核心语义关系之外的语义角色关系，CAMR规定了在AMR的基础上规定了目的、处所、时间等44种对所有谓词通用的非核心语义关系。根据对话语料特点，DAMR新增了4个概念标签和2个非核心语义关系标签以兼容对话中会出现的语义关系。

语篇编号	话轮编号	句子编号	说话人编号	句子
16	400	709	131525	太[1] 土[2] 了[3]

s709_x5/speak
:arg0() x4/speaker
:arg2() x6/hearer
:arg1() x2/土-01
:degree() x1/太
:aspect() x3/了

图 4. speak概念

4.2.1 说话speak

DAMR为对话新增了speak、speaker和hearer概念，对话中的每一个句子的根节点都为概念speak，概念speak规定了三个论元，分别为：arg0: speaker（说话人）；arg1: thing speak（说话内容）；arg2: hearer（听话人）。说话人speaker和听话人hearer 为新增概念，标注时需根据实际语义标注出话语的说话人和听话人。如图4。

本文其他例句会省略speak、speaker、hearer概念以使例子更清晰。

语篇编号	话轮编号	句子编号	说话人编号	句子
2	85	144	135882	555[1] ，[2] 复习[3] 好[4] 痛苦[5]
2	85	145	135882	……
2	86	146	138459	嗯[1]

s144_x5/痛苦-01
:arg0() x6/speaker
:arg1() x3/复习
:degree() x4/好

s146_x1/confirm
:arg1() s144_x5/痛苦

图 5. confirm概念

4.2.2 肯定confirm

对话是交互的，听话人会对说话人的话语表达态度，最常见的是对上一句的肯定（是的、嗯嗯等）,针对这种情况，DAMR新增了一个confirm概念。具体示例如图5，句146 的根节点是肯定概念confirm，“嗯”是对“复习好痛苦”的肯定，标注时将“嗯”抽象为概念“confirm”，不再单独表示出来。

4.2.3 情感:feeling

对话中说话人会用多种形式表达自己的心情，如“哈哈”“呵呵”“呜呜”等，以及在线上对话文本中会出现的表情包，甚至是单纯的标点符号，如“。。。”、“...”，DAMR新增了非核心语义关系标签“feeling”来表示这种语义。

如图5所示，“555”表示说话人的心情，将其置于根节点“痛苦”的下层。由于心情的表示形式太过复杂，例如“呵呵”可表示偏正向的愉悦情绪，而现在网络上的新兴用法也将“呵呵”表负面情绪，因此为了避免标注的不统一，目前DAMR只使用“feeling”标签，不区分具体的情绪类别。另外，一部分表情并非表示心情，而是表示“再见”、“你好”等概念，对于这部分表情，DAMR要求对其进行语义转写，将其真正语义表示出来。

4.2.4 称呼:naming

存在称呼语是对话中突出的特点，称呼本身不存在于谓词概念的论元结构中，为了更好地表示对话中的称呼现象，DAMR引入非核心语义关系标签“naming”。注意与原标签“name”区分，称呼并不等同于实际名字“name”，称呼是动态的，“name”则是静态的。

如图6所示，根节点speak支配的论元分别为arg0说话人、arg1说话内容、arg2听话人，因为说话人称呼听话人为“王老师”，因此naming在arg2节点的下层。如果称呼就是听话人的名字，则“naming”和“name”同时出现，都在arg2节点的下层。

语篇编号	话轮编号	句子编号	说话人编号	句子
20	549	967	151461	王老师[1] 你[2] 是否[3] 可以[4] 给[5] 出[6] 更[7] 多[8] 的[9] proposition[10]

```
s967_x12/speak
    :arg0() x12/speaker
    :arg2() x4/hearer
        :naming() x1/王老师
    :arg1() x4/可以-01
        :arg0() x5/给-01
            :arg0() x2/你
            :arg1() x10/proposition
                :quant(x9/的) x8/多
                    :degree() x7/更
            :direction() x6/出
        :mode() x3/interrogative
```

图 6. naming关系标签

语篇编号	话轮编号	句子编号	说话人编号	句子
20	567	997	138158	今天[2] 的[3] 电影[4] 好看[5] 么[6] ？[7]

```
x5/好看-01
    :arg0() x1/电影
        :time(x3/的) x2/今天
    :argh() x4/hearer
    :mode() x6_x7/interrogative
```

图 7. ”好看”的论元结构

4.3 论元结构

对话中的话语常常带有说话人或者听话人的主观态度，在CAMR中没有将这种态度表示出来，因此DAMR 对一部分谓词的论元结构进行了修改，增加了一个论元：argh （态度持有人）。目前修改的谓词是含有主观态度的形容词或短语，如“好看”“丑”“很有主意”，这部分词语原来只有一个论元：arg0（entity describe）。

如上图所示，“好看”除了arg0（电影）外，还增加了态度持有人argh，根据语义，“好看”的态度持有者是听话人hearer。

4.4 对话中的特殊现象

4.4.1 问答句的对应

问答是对话中常见的形式，但是在对话中说话人可能同时进行两个或多个话题，问句和答句不一定为相邻句，问答的语义对应较为分散，并且答句不一定正面回答问句，为体现问句和答句之间的语义联系，DAMR在每个句子上增加一个字段，可以将答句与问句联系起来。

语篇编号	话轮编号	句子编号	说话人编号	句子	**问答位置信息**
4	149	257	138375	那[1] 个[2] socio[3] 难[4] 不[5] 难[6] 整[7] ？[8] ？[9] ？[10]	
4	149	258	138375	嗯[1] 是[2] 的[3] 小[4] 野兽[5] 最近[6] 也[7] 满[8] 虚弱[9] 的[10]	
4	150	259	138194	哎[1] 我[2] 觉得[3] 我[4] 重新[5] 上[6] 了[7] 一[8] 次[9] socio[10] 似的[11]	**s257_x12**

表 3. DAMR语料字段

```
s257_x12/or
    :op1() x4/难-01
        :arg0() x7/整-01
            :arg1() x3/socio
                :mod() x1/那
                :cunit() x2/个
    :op2() x6/难-01
        :polarity() x5/-
        :arg0() x7/整-01
    :mode() x7_x8_x9/interrogative
```

图 8. 疑问句

问句位置信息有两个维度。第一个维度为答句所对应的问句编号，第二个维度为所对应问句的根节点。如上图，问句257的根节点为x12，因此答句259 的问句位置信息为s257_x12。

4.4.2 问句的省略

由于对话双方处于同一个语境中，完成对话理解所需的背景知识是两者共享的，因此对话中的一方提问常常会省略很多成分，在标注时，需要根据句子实际语义将省略的问句成分表示出来。如图9，句2仅用一个问号表示说话者的疑问，其完整语义为：为什么说atac疯了，在标注时需将实际语义标注出来。

4.4.3 人称省略

在对话中说话人常常会省略自己或听话人的人称，标注时要将省略的说话人或听话人补出来。如图10，“弄完”的施事为听话人，在标注时我们将省略的hearer补充出来。

4.5 小结

我们改进了原CAMR标注平台，加入了篇章对话信息（语篇编号、话轮编号、说话人编号），通过对476 句对话语料的标注，针对对话特点新增了标签，处理了称呼语、情感短语等对话特有现象，规定了省略、话轮间应答关系的标注，使CAMR 体系从单句拓展到篇章级别。

语篇编号	话轮编号	句子编号	说话人编号	句子
1	1	1	151430	atac[1] 疯[2] 了[3]
1	2	2	131525	？[1]

s1_x3/疯-01
:arg0() x1/atca
:aspect() x3/了

s2_x2/amr-unknown
:cause-of() s1_x3/疯
:mode() x1/interrogative

图 9. 省略问句

语篇编号	话轮编号	句子编号	说话人编号	句子
7	143	232	131525	嗯[1] 不过[2] 先[3] 把[4] 小说[5] 弄完[6] 吧[7]

s232_x18/contrast
:arg1() x10/confirm
:arg1() s231_x16/causation
:arg2(x6/不过) x10/弄完-01
:arg0() x24/hearer
:arg1(x4/把) x5/小说
:time() x3/先
:mode() x7/imperative

图 10. 人称省略

5 结论及未来工作

近年来对话系统的发展越来越受到重视，对话语义的形式化表示的作用愈发凸显，但国内目前还没有较完整的标注体系表示对话的语义。本文梳理了国内外对话标注体系和语料库的发展，在CAMR体系的基础上进行改进扩充：实现跨单句层面的概念对齐，新增适用于对话语料的概念标签和非核心语义关系标签，修改词语的论元结构，规定了问答句对应、省略等对话中特有现象的标注，形成了对话标注体系DAMR。这些改进有利于解决对话中的省略和跨句子指代等问题，使问答点的对应更明确，更完整地表达说话者的语义，对合理表示对话语义，为对话的自动理解与分析有较大价值。

在今后的工作中，第一，我们将加强对对话语义特点的研究，尝试标注语音对话转写语料，针对实际对话特点和新出现的问题完善DAMR标注体系，使之能够适用各个领域的对话语料，以验证DAMR的效果；第二，使用DAMR标注体系标注语料，构建一个大规模的对话AMR语料库，并进行统计分析。第三，我们希望通过对对话的标注语料库进行学习，提高对话自动分析的效果。

参考文献

Allison Adams. 2000. Dependency Parsing and Dialogue Systems. UPPSALA University.

Amanda Stent. 2000. Rhetorical Structure in Dialog. *Proceedings of the First International Conference on Natural Language Generation*, 247–252.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking *Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186.

Carlos D.Martinez. 2002. A Labelling Proposal to Annotate Dialogues. *Proceedings of the Third International Conference on Language Resources and Evaluation(LREC)*, 1577–1582.

Christian Raymond. 2007. The LUNA Corpus: an Annotation Scheme for a Multi-domain Multi Lingual Dialogue Corpus. *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, 185–186.

Claire N. Bonial, Lucia Donatelli, Jessica Ervin, Clare R. Voss. 2019. Abstract Meaning Representation for Human-Robot Dialogue. *Proceedings of the Society for Computation in Linguistics*.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for Dialogue. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 684–695.

Harry Bunt. 2010. The DIT++ Taxonomy for Functional Dialogue Markup. *Proceeding of 8th Int. Conf. on Automous Agents and Multiagent Systems(AAMAS,2009)*, 13–24.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, David Traum. 2010. Towards an ISO Standard for Dialogue Act Annotation. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.. 2548–2555.

HCRC group. 1996. *HCRC Dialogue Structure Coding Manual*. University of Edinburgh.

James Allen, Mark Core. 1997. *Draft of DAMSL: Dialog Act Markup in Several Layers*. Lancaster University.

James Allen, Peter Heeman. 1994. *TRAINS Spoken Dialog Corpus*. University of Rochester.

Jurafsky D, Shriberg E, Biasca D. 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual*.

Keyan Zhou, Aijun Li, Zhigang Yin, Chengqing Zong. 2010. CASIA-CASSIL: a Chinese Telephone Conversation Corpus in Real Scenarios with Multi-leveled Annotation *Proceedings of the International Conference on Language Resources and Evaluation(LREC 2010)*. 2407–2413.

Lynn Carlson, Daniel Marcu, Mary Ellen Okurovsky. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

M. POESIO, F. Bruneseaux, L. Romary. 1999. The MATE Meta-scheme for Coreference in Dialogues in Multiple Languages. *ACL Workshop Towards Standards and Toolos for Discourse Tagging*, 65–74.

Nianwen Xue, Qishen Su, Sooyoung Jeong. 2016. Annotating the Discourse and Dialogue Structure of SMS Message Conversations. *Proceedings of LAW X –The 10th Linguistic Annotation Workshop*. 180–-187.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Martha Palmer. 2018. AMR Beyond the Sentence: the Multi-sentence AMR corpus. *Proceedings of the 27th International Conference on Computational Linguistics*. 3693–-3702.

PDTB-Group. 2009. *The Penn Discourse Treebank 2.0 Annotation Manual*. University of Pennsylvania.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad. 2016. Annotation of Discourse Relations for Conversational Spoken Dialogs. *Proceedings of International Conference on Language Resources and Evaluation(LREC2010)*. 2084–2090.

Stephan Oepen, Omri Abend, Jan Hajic, Daniel Hershcovich, Marco Kuhlmann, Tim O'Gorman, Nianwen Xue, Jayeol Chun, Milan Straka, Zdenka Uresova. 2016. MRP 2019: Cross-Framework Meaning Representation Parsing. *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*. 1–27.

方称宇, 曹竟, 刘晓月. 2013. 基于语料库的最新ISO会话行为标注体系的研究：从SWBD-DAMSL到SWBD-ISO. 当代语言学. 15(4): 439–458.

李斌, 闻媛, 宋丽，卜丽君，曲维光，薛念文. 2017. 融合概念对齐信息的中文AMR语料库的构建. 中文信息学报. 31(6):93–102.

王珊, 刘锐. 2016. 谈话节目语料库的构建与会话结构分析. 中文信息学报. 30(6): 140–146.

郑桂东. 2018. 多轮对话语料构建中的离群对话分析. 哈尔滨工业大学.

周强. 2017. 汉语日常会话的对话行为分析标注研究. 中文信息学报. 31(06):75–82.

周小强, 王晓龙, 陈清财. 2017. 交互式问答的关系结构体系及标注. 中文信息学报. 32(5): 1–10.

宗成庆，吴华，黄泰翼，徐波. 1999. 限定领域汉语口语对话语料分析. 全国第五届计算机语言联合学术会议. 115–122.