文章编号：1003-0077（2017）00-0000-00

融合概念对齐信息的中文 AMR 语料库的构建*

李斌[1]，闻媛[1]，宋丽[1]，卜丽君[1]，曲维光[2,3]，薛念文[4]

（1.南京师范大学 文学院，江苏省 南京市 210097；

2.南京师范大学 计算机科学与技术学院，江苏省 南京市 210023；

3.闽江学院 福建省信息处理与智能控制重点实验室，福建省 福州市 350121；

4.布兰迪斯大学 计算机学院，美国 沃尔瑟姆市 02453）

摘要：作为一种新的句子语义表示方法，抽象语义表示（AMR）将一个句子抽象为单根有向无环图，已经建立了较大规模的英文语料库。然而，句子中的词语和 AMR 图的概念对齐信息缺失，使得自动分析效果和语料标注质量受到影响，同时中文尚无较大规模的 AMR 语料库。本文介绍了中文 AMR 语料库的构建工作，针对汉语特点调整了 AMR 的标注体系，增加对复句关系的标注，提出了融合概念对齐的一体化标注方案，解决了中英文输入法频繁切换的问题，增加了错别字纠正和未标注词高亮功能，提高了标注效率。然后，从 CTB 中选取了 6923 句进行人工标注，形成中文 AMR 语料库，统计得到图和环的比例分别为 48%和 1%，以及利用对齐信息才能获取的非投影句的比例 32%，为中文 AMR 的理论和自动分析研究奠定基础。

关键词：抽象语义表示；语义图；句子语义；语言知识库；

中图分类号：TP391　　**文献标识码：**A

Construction of Chinese Abstract Meaning Representation Corpus with Concept-to-word Alignment

LI Bin[1], WEN Yuan[1], SONG Li[1], BU Lijun[1], QU Weiguang[2,3], XUE Nianwen[4]

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
2. School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu 210023, China;
3. Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, Fujian, 350121, China;
4. Michtom School of Computer Science, Brandeis University, Waltham, MA 02453, USA)

Abstract: As a new sentence-level meaning representation, Abstract Meaning Representation (AMR) uses a rooted acyclic directed graph to represent the meaning of a sentence. A large AMR bank has been constructed for English, but the concepts of an AMR graph are not aligned to the words in a sentence, which artificially increases the difficulty in manual annotation as well as automatic parsing. This paper describes the construction of a Chinese AMR corpus, based on guidelines adapted from English but accounting for Chinese-specific properties. We also designed an efficient annotation framework that incorporates concept-to-word alignment, taking advantage of the morphology-poor nature of Chinese. We have annotated the AMRs of 6923 sentences selected from the Chinese TreeBank, among which 48% of the sentences are graphs, 1% of the sentences are cycles, and 32% have non-projective subtrees. We plan to publicly release this data for use in linguistic and NLP research.

Key words: Abstract Semantic Representation; Semantic Graph; sentence meaning; language knowledgebase;

1 引言

句子语义的自动分析，是继词法分析、句法分析之后，自然语言处理学界寻求突破的重点课题。但是关于句子语义的理论众说纷纭[1][2][3]，又缺乏大规模的语义标注实践，使得语义自动分析停滞不前。究其原因，语义不同于句法，语义结构很难像句法结构那样简明地表示出来。在这种背景下，以美国南加州大学 Kevin Knight 为代表的欧美多位学者近年共同提出了 AMR（Abstract Meaning Representation，抽象语义表示）的句子语义表示方法[4]，标注了英语《小王子》等文学、新闻、生物领域的语料。AMR 较好地解决了句子语义表示的三个难题：（1）以句法语义为重点，兼顾词汇语义，允许增删原句词语，表示能力强[5]；（2）表示方法简明直观，人工标注一致性已

* **收稿日期：**　　**定稿日期：**

基金项目：江苏高校哲学社会科学研究项目（2016SJB740004）；国家自然科学基金（61472191，61272221）；福建省信息处理与智能控制重点实验室开放基金项目（MJUKF201705）。

到达 0.83，并建立了四万多句的大规模英语语料库；（3）统计机器学习方法在英语 AMR 自动分析上已经取得了 0.62 左右的 F 值[6]。

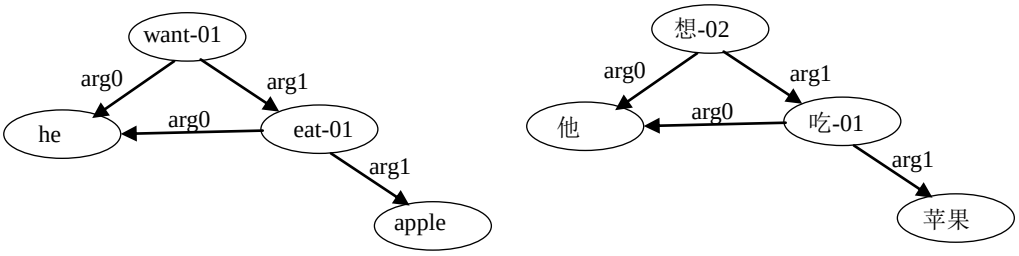
不过，除了分析精度不高，尚在初创期的 AMR 也存在三个问题：（1）忽略虚词、形态变化的各种语法意义和复句关系；（2）AMR 图上的概念没有与原句的词语对齐，而自动对齐的 F 值仅有 90%左右[7]，这就意味着对齐问题解决后自动分析还有一定提升空间；（3）其他语言的语料还很少，图结构的跨语言有效性有待进一步论证。因此，补充一定的语法意义，增加复句关系和对齐信息，提高 AMR 的自动分析效果，增加更多语种的 AMR 语料，深入探讨 AMR 的语言学理论价值[8]，就成为目前该领域最为迫切的研究内容。而目前，中文 AMR 的语料库规模很小，只有 1000 多句缺少概念对齐的《小王子》语料[9]，亟需调整标注方法，扩大语料规模。

本文根据 AMR 的基本理论和原则，给出了融合 AMR 语义图与原句对齐的一体化人工标注方案和软件平台，结合汉语的特点制定汉语 AMR 的标注体系和标注规范，建立了近 7000 句的中文 AMR（下文简称 CAMR）语料库。然后，针对汉语中的语义图、非投影树、环等结构进行统计分析，阐述对齐版 CAMR 语料库的价值。

2 AMR 简介与语料标注体系

AMR 是一种抽象的句子语义的表示方法，不同于传统的树形结构（tree），它将一个句子的语义抽象为一个单根有向无环图（a single rooted, acyclic, directed graph）。所谓“抽象”是指，把句子中的实词抽象为概念节点，把实词之间的关系抽象为带有语义关系标签的有向弧，忽略虚词和由形态变化体现的较虚的语义（如冠词、单复数、时态等等），同时还允许补充句子中省略或缺失的概念。AMR 虽采用图结构，但其单根的要求使得句子依然以依存树结构为主体，层次鲜明。下面以实例来说明 AMR 作为句义表示方法的两个优点。

（1）**采用图结构处理论元共享问题。**AMR 与传统句义表示方法的主要差异在于对论元共享现象的处理。例如，在英语句子“He wants to eat the apple”及汉语翻译“他想吃苹果”中，传统的句法分析方法，如短语结构文法和依存文法，都限于树形结构，会舍弃“他-吃”这个施事关系；而 AMR 则将两个关系都保留，形成图结构，解决了“他”同时作为“想”和“吃”的 arg0（施事）问题。



A: He wants to eat the apple
w/want-01
:arg0 h/he
:arg1 e/eat-01
:arg0 h
:arg1 a/apple

B: 他想吃苹果
x/想-02
:arg0 x1/他
:arg1 x2/吃-01
:arg0 x1
:arg1 x3/苹果

图1：论元共享形成的图结构示例

上图给出了 AMR 的两种展现形式，图示法（上）和文本缩进法（下）。句 A 和句 B 中，每个概念节点都有一个字母编号，关系 arg0、arg1 等取自 PropBank 的论元关系[10]。“want（想）”作为句子唯一的根结点，“he”和“他”的编号分别为 h 和 x1，是“want（想）”的 arg0（施事），也是“eat（吃）”的 arg0（施事）。为了明确谓词及其论元之间的语义关系，AMR 标注了谓词在 PropBank 词典中的具体义项，如谓词“want-01”，表示此处的“want”使用的是其第一个义项的论元框架。

（2）**允许重新分析和补充概念，能更完整地表示一个句子的语义。**AMR 更为灵活之处在

于允许根据整体语义来增删概念节点，这样能够弥补传统的句法表示的缺陷。句 C 和句 D 给出了“The dancer has gone”及其汉语翻译“跳舞的走了”的 AMR 表示。AMR 可以根据上下文将“dancer”重新分析为跳舞的 arg0（施事）person（人）。中文则可以添加出概念 person，作为“跳舞”的 arg0。AMR 的这个特点，解决了传统的句法表示方法无法应对的省略和词内分析困境，具有较高的语言学价值和应用价值。

图2：概念补充的句子示例

此外，AMR 还允许删除一些在意义上冗余的实词，使得句子的基本意义更加简明。比如，“他回答道”可以省略“道”。

AMR 在句子语义表示方面的优点，使其一经问世就获得了学界的高度关注，从 2014 年到 2016 年，自然语言处理领域的 ACL、EMNLP、NAACL 等国际顶级会议出现了多篇关于 AMR 自动分析和应用的论文，2016 年的国际语义评测 SemEval 也进行了英语 AMR 自动分析的专项评测[6]。相比之下，国内 AMR 的相关研究较少。比较相近的工作有哈工大的语义依存语料图库[11]和武汉大学的概念依存图库[12]及自动分析技术，两者都采用依存图结构来表示句子的语义，但没有 AMR 的概念补充、删除和替换的抽象机制，未限制为单根，存在多根的情况，其标注体系与规范也尚未公开。北京大学探索汉语“意合语法”的“词库-构式”知识库采用的是生成词汇学和构式语法的理论框架[13]，与 AMR 差别较大。中文《小王子》AMR 语料库只有 1000 多个句子，没有进行概念对齐[9]，对标注体系的介绍也不够完整。因此，本文基于 6000 多句的标注实践，着力介绍中文 AMR 的概念对齐标注方法以及标注体系的特色，最后用统计数据说明该语料库的理论与应用价值。

3 融合概念对齐的中文 AMR 的标注体系

中文 AMR 沿用了英文 AMR 的体系和规范，本着与英文 AMR 保持兼容、兼顾汉语特点、提高标注质量的原则，我们进行了三个方面的改进：（1）增加概念对齐信息，并将其融合到语料标注过程中；（2）增加复句关系的标注；（3）对中文特殊现象的标注予以具体规定。

3.1 概念对齐

中文 AMR 语料库的构建之初，使用的是美国南加州大学的 AMR 标注工具[1]。在标注中文语料时发现三个较大的缺点：（1）对于一个句子，哪些词语已经标注过，哪些没有标注过，没有任何提示。当一个句子过长时，标注人员特别容易重复标注或者漏标一些词语。（2）由于需要输入汉字，来回切换中英文输入法较为耗时。（3）每个概念都分配了一个标签，但除了区分不同的概念外，并没有什么更好的用途，如“p/person”中“/”左边的“p”。而这三个问题，都源于 AMR 不做概念对齐。如果有了概念和词语的对齐信息，则可以直接利用词语编号代替词语的手工输入，减少中英文切换，同时记录下来哪些词语已经被标注过。

如果用词语在句子中的编号作为概念标签，就可以做到概念对齐。不过，AMR 没有做概念对齐，肯定有特殊的考虑。比如，英文词语有形态变化，而概念没有，所以很难直接用词语编号来代表概念，另外，AMR 有时将 teacher（教师）这样的词内部分析为“person :arg0-of teach”，也使得利用编号进行概念对齐存在较大困难。然而，汉语几乎没有形态变化，在大多情况下，概念和词的形式是一样的。只是汉语中的一些特殊结构和用法，如重叠式（认认真真）、离合词（帮了一个忙）、错别字（窝-我）等，给对齐方案的设计带来了困难。因此，我们提出了以词语编号为基础的概念对齐标注方法，解决了四个问题：（1）做到了概念和词语的对齐；（2）应对重叠式、离合词等形式变化和错别字问题；（3）极大减少输入法来回切换，基本用英文输入，减少了汉字输入时间。（4）使用词语高亮警示，防止标注时漏标词语。

[1] https://www.isi.edu/cgi-bin/div3/mt/amr-editor/login-gen-v1.7.cgi。

具体方案为双层编号法，即根据每个概念对应的词语的编号和词内的字的编号来做对齐。在一般情况下，只需输入“x+词语编号”即可表示一个概念。表 1 给出了具体的标注样例，每条输入语句为“支配概念的编号 :语义关系 被支配概念的编号(校正概念)”。

表1：CAMR对英文标注法的对齐改进及输入过程

谁[1] 帮[2] 了[3] 窝（我）[4] 这么[5] 大[6] 的[7] 忙[8]？[9]		
AMR 无对齐的文本表示	CAMR 带对齐的文本表示	CAMR 的人工输入
x1/帮忙-01 :arg0 a/amr-unknown :arg1 x1/我 :degree x2/大 :degree x3/这么 :mode i/interrogative	x2_x8/帮忙-01 :aspect x3/了 :arg0 x1/ amr-unknown :arg1 x4/我 :degree x6/大 :degree x5/这么 :mode x9/interrogative	root :top x2_x8 x2_x8 :aspect x3 x2_x8 :arg0 x1(amr-unknown) x2_x8 :arg1 x4(我) x2_x8 :degree x6 x6: degree x5 x2_x8 :mode x9(interrogative)

从概念对齐的标注方法来看，在 CAMR 图上的每个概念都尽可能地对应到原始句子的词语乃至标点上。离合词“帮忙”用 x2_x8 表示；疑问词“谁”对应词语编号 x1，并标出 amr-unknown（表示未知的概念）；错别字“窝”的编号为 x4，将其校正为“我”；疑问语气 interrogative，则对应问号“？”的编号 x9。

对于词内分析、分词错误等特殊情况，则使用词内的汉字编号方法解决对齐问题。例如“土地/拥有者”这个例子，“拥有者”作为一个词不容易标出和“土地”的关系，需要进行内部拆分，标注如下。

表2：词内编号拆分标注示例

土地[1] 拥有者[2]
x2_3/者 :arg0-of x2_1_2/拥有-01 :arg1 x1/土地

x2_3 表示的是第 2 个词的第 3 个字“者”，x2_1_2 则表示第 2 个词的前两个字“拥有”。这种方式比较直观，便于标注人员记忆，也便于后期编程统计处理。即使出现类似 x2_1_2_3_4_5_6 等较长的情况，录入稍显麻烦，也比用中文输入法打出每个汉字要快，而且这种情况出现极少，不会对标注造成影响。

相比于英文 AMR 的标注方式，CAMR 的双层编号法基本不需打中文，大量减少了中英文输入法切换，减轻了标注人员的操作量。稍微耗时的则是离合词这样少见的需要输入多个词语编号的情况。对于普通的句子来说，一个词的编号很简单，输入速度快。表 3 给出了一个大致的估算，如果不计编号合并和拆分的情况，每条关系的输入平均可以节约 4 次敲击，节约 23%的录入时间。即使计入编号合并、拆分、纠正错别字等复杂情况，综合考量，相比英文 AMR，人工录入的时间也较少。

表3：一般情况下英文AMR和CAMR标注方法输入一条关系的效率估算

标注平台	核心节点平均敲击数（一般相同）	关系平均敲击数（相同）	依存节点平均敲击次数（一般不同）	合计
英文 AMR	输入编号 3+空格 1	输入关系 5+空格 1	整词拼音 5+中英文切换 2	17
CAMR 对齐一体化	输入编号 3+空格 1	输入关系 5+空格 1	输入编号 3	13

根据双层编号法，我们开发了 CAMR 专用的人工标注平台 CAMR AnnoKit，既做到了概念对齐，又可以校正词形、修改错别字，同时极大减少输入法来回切换，基本用英文输入，减少了汉字输入时间。还借助对齐信息，使用词语高亮警示，防止漏标词语，标注速度和质量有了明显的提高。

3.2 标注体系

CAMR 大体上沿用了英文 AMR 采取的 OntoNotes 的标注体系，使用了 5 个核心的语义关系标签、44 个非核心语义关系标签以及 109 个专名概念。

（1）语义关系。核心语义关系是指谓词自身的事件框架的若干语义角色，包括 5 种：arg0（原型施事）、arg1（原型受事）、arg2（间接宾语、工具等）、arg3（起点、属性等）、arg4（终点）。而英文非核心语义关系则为 40 种，我们新增了 4 种。CAMR 共使用 44 种标签来表示一般的非核心语义角色关系，这些标签参考了 AMR 的标注体系，并根据中文的特点进行了修改和补充，考虑到与 AMR 的兼容性，CAMR 的非核心语义角色关系标签仍然使用英文单词。分别如下：

表4：中文AMR非核心关系

序号	非核心关系	中文说明	序号	非核心关系	中文说明
1	accompanier	伴随	**23**	medium	媒介
2	*aspect	体	**24**	mod	修饰
3	beneficiary	受益者	**25**	mode	语气
4	cause	起因	**26**	name	名称
5	compared-to	参照物	**27**	ord	序数
6	consist-of	构成（材料）	**28**	part-of	部分
7	condition	条件	**29**	path	路径
8	cost	花费	**30**	polarity	极性
9	*cunit	中文特殊量词	**31**	polite	礼貌
10	degree	程度	**32**	poss	领属
11	destination	目的地	**33**	purpose	目的
12	direction	方向	**34**	quant	数字
13	domain	陈述	**35**	range	跨度
14	duration	时长	**36**	source	源
15	example	例子	**37**	subevent	子事件
16	extent	范围程度	**38**	subset	子集
17	*perspective	方面	**39**	superset	父集
18	frequency	频率	**40**	*tense	时
19	instrument	工具	**41**	time	时间
20	li	数字编号	**42**	topic	话题
21	location	处所	**43**	unit	度量单位
22	manner	方式	**44**	value	值

注：加*的表示中文 AMR 新增的 4 种关系。

下面重点说明一下我们增加的关系标签，tense（时）、aspect（体）、cunit（中文特殊量词）、perspective（方面）。在英文 AMR 的体系中，语义较虚的时、体等是不予标注的。按这个标准，中文的特殊量词（如个、只等）也被排除在外。但我们觉得句子中的时、体和中文量词的信息相当重要，能够使句义更为完整，值得标注出来，且不会耗费多少时间与精力。perspective 则是我们不得不增加的关系类型，如“他在经济上独立了”，“经济”和“独立”的关系，难以用 AMR 的关系来标注，所以把“经济”作为“独立”的 perspective。

（2）复句关系的表示方法。AMR 对于复句关系重视不足，仅有 cause、condition、concession 等几种语义关系和 and、or 两个概念。关系和概念纠缠在一起，不利于统计分析和计算应用。我们统一设置了 10 个概念而非关系来表示小句之间的复句关系，每个小句由编号基于 1 的 argX 来建立关系。因为关系往往只关联两个成分，而像并列、时序、递进复句的小句成分往往多于 2 个。例如，我们增加的时序（temporal）概念，在处理复句“她吃过晚饭，去跳舞，又逛了夜市”时，将 temporal 作为根结点，arg1 为“她吃晚饭”，arg2 为“她去跳舞”，arg3 为“她逛夜市”。

表5：CAMR的10种复句概念

序号	复句概念	中文说明	序号	复句概念	中文说明
1	and	并列	6	or	选择
2	causation	因果	7	concession	让步
3	condition	条件	8	orx	排他选择
4	contrast	转折	9	progression	递进
5	temporal	时序	10	expansion	解说

（3）专名体系。在专名方面，AMR 给出了一个含有 109 个专名的体系，用于专有名词的标注和缺失概念的标注。如“北京”标注为“city :name 北京”，“跳舞的”补充“person”，标注为“person :arg0-of 跳舞”。表 6 中，每行为一大类，一个大类的代表词放在第一位，其他则为下位小类的标签。如事物的小类不明确，则用大类；如果大类也不明确，则使用顶级标签 thing（事物）。

表6：专有名词（Name Entity）类别表

专名类别及部分下位分类	数量
thing（一般事物及类型不明确的事物）	1
person（人），family（家庭），animal（动物），language（语言），nationality（国籍）……	8
organization（组织），company（公司），government-organization（政府组织），military（军队）……	11
location（位置），city（市），city-district（市区），county（县），state（州），province（省）……	21
facility（设施），airport（机场），station（车站），port（港口），tunnel（隧道）……	21
event（事件），incident（偶发事件），natural-disaster（自然灾害），earthquake（地震）……	8
product（产品），vehicle（车辆），ship（船舶），aircraft（飞机），aircraft-type（飞机类型）……	12
publication（出版物），book（书籍），newspaper（报纸），magazine（杂志），journal（期刊）	5
natural-object（天然物体）	1
molecular-physical-entity（分子物理实体），small-molecule（小分子），protein（蛋白）……	15
law（法律），treaty（条约），award（奖励），food-dish（食品），disease（疾病）……	6
合　　计	109

（4）其他概念和关系。此外，还延续了 AMR 关于时间、日期、地址、度量衡等具体概念和单位的标注方法，皆参考了 AMR 标注规范 1.2.2 版[14]。

3.3 汉语特殊现象的规定

汉语的特殊语言现象的标注方法，依然以英文 AMR 的标注规范为基础，将汉语特殊结构的语义用概念或语义关系表示出来。限于篇幅，下面仅略为介绍主要的方面，详细的标注规范，请参考已公开的版本[2]。

（1）首先是增加对汉语特殊量词、时和体的标注，还增加了 perspective（方面）这一关系，详见上文。

（2）对重叠式进行还原，如“看看”还原为“看”，“开开心心”还原为“开心”，“打扫打扫”还原为“打扫”等。如果重叠式有特殊而明确的含义，也予以标注，如“年年”按照 AMR 处理频率表达“每年”的方式进行标注。

（3）对汉语离合式一般采取“合”的方式，如“游了一下午泳”处理为“游泳”持续的时间为“一下午”。

（4）对于连动、兼语结构，一般按照论元共享进行处理，如上文“他想吃苹果”，“想”支配“吃”，且“想”和“吃”共享 arg0“他”。

（5）对动补结构，根据其语义进行标注，如“走不了”处理为“不能走”、“唱哭”处理为“哭”是“唱”的结果。

4 语料分析

4.1 人工标注及基本统计

基于 CAMR 的对齐一体化标注工具，我们标注了宾州中文树库 CTB8.0 语料（以下简称 CTB）的网络媒体语料 6923 句（原语料共 7022 句，其中 99 句存在断句错误或句子意义错乱，未予标注）。在其中随机抽样的 500 句语料上，双人标注一致性达到 0.83 的 Smatch 值[15]，与英文 AMR 基本相当。标注时，谓词所采用的语义角色框架则使用中文谓词库（CPB）的谓词框架词典[16]。该词典是从 CPB 标注语料中抽取出来的，含有每个谓词在不同义项下的语义角色框架，共收录了 24510 个中文谓词（包括动词、形容词等）的 26650 个义项的不同语义角色框架。这部词典较好地覆盖了 CTB 和《小王子》的语料。少量没有覆盖到的谓词，其语义角色则根据标注规范从

[2] http://www.cs.brandeis.edu/~clp/camr/res/CAMR_GL_v1.2.pdf。

AMR 规定的语义关系中选取。下面给出基本的统计数据和对于图、环、非投影树及较为重要的现象的统计分析。

在 6923 个句子中，平均句长为 22.36 个词，平均概念数 19.24 个，回边 6778 条，3360 个句子是图，比例为 48.53%。平均每句添加 2.92 个概念，其中 0.84 个为专名添加，2.08 个概念是额外添加的。复句概念属于添加出来的概念节点，我们增设的概念 temporal（时序）和 progression（递进）允许多个 argx 的出现，在语料标注中确实体现出了优点。语料中共有 206 个 temporal 复句，最多有 6 个分句；progression 复句 251 个，最多 3 个分句。说明本文的标注方法对复句关系的描写能力要比用关系标签更好。

4.2 图结构

文献[9]已经统计了中文《小王子》AMR 语料中图结构的比例约为 36%。我们统计了英文 AMR 语料 11875 句中图结构的比例，平均为 49%。本文新标注的 CTB 语料则为 48.53%，图结构句子的比例与英文 AMR 语料相当，基本可以说明图结构在中英文语料中的普遍性。

表7: 中英文AMR语料含图结构的句子统计

AMR 语料	总句数	含图句数	含图比例
英文 bolt	1062	722	0.68
英文 dfa	1703	898	0.53
英文 mt09sdl	204	137	0.67
英文 proxy	6603	2954	0.45
英文 xinhua	741	423	0.57
英文小王子	1562	663	0.43
英文合计	11875	5797	0.49
中文小王子	1562	576	0.36
中文 CTB	6923	3360	0.48
中文合计	8485	3936	0.46

我们观察了图结构在句长上的分布，发现随着句子长度的增加，图结构出现的比例越来越高。图 3 给出了句长和图结构的分布曲线，5 个词以上的句子才开始出现图结构，20 个词以上的句子中，图结构出现的比例则超过了 50%。在超过 65 个词之后，几乎全是图结构，偶有几个句子不是图，造成了下降的尖峰。这也说明，和一般的句法分析问题相似，长句分析仍然是自动分析算法的难点。

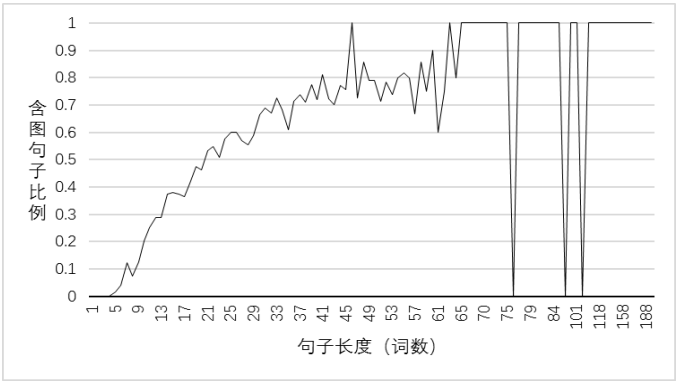


图3：中文CTB语料句子长度与图结构的关系

从回边和概念添加的情况看，自动分析难度也较大。在图结构的句子中，平均每句含 2.02 条回边。从全部句子来看，回边的数量最小值为 0，最大值为 14，方差为 1.41。而对于添加的概念来说，最小值为 0，最大值为 26，方差为 2.42。这对于机器自动判别哪里增加边和概念来说，确实具有挑战。从目前英文的自动分析结果来看，添加概念和判定概念之间的关系效果都不理想[6]。不过，语料的统计也可以提供一些有用的信息。根据文献[9]统计中英文《小王子》语料的结果，仅 arg0、arg1 和 arg2 三种关系造成的论元共享就分别占到回边总数的 85.79%和 75.11%。而在 CTB 语料中，这三种关系造成的论元共享比例为 86.43%。说明在自动分析给初始特征时，除了树形结构，也应给出语义角色 SRL 的自动标注结果。对我们标注的 CAMR 语料来说，复句关

系的判定也最好能够先行给出。

4.3 环

虽然 AMR 采用有向无环图来表示句子的语义结构，但是在具体标注中，英文语料约 0.3%的句子出现了环（cycle）[14]，即一个概念结点经过其他结点重新指向了自己。例如表 8 中的句子，“英雄”是“敢于”的施事 arg0，“救”是“敢于”的 arg0，而“英雄”又是“救”的施事 arg0，形成了一个环。为了避免打印环时形成死循环，我们采用的方式是，对于出现 2 次以上的概念编号，仅输出其第一次出现时的子树，后续出现则只打印其编号。

表8：带环的句子示例

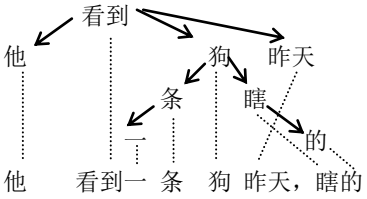
感谢 1 敢于 2 救 3 人 4 的 5 英雄 6
x1/感谢-02
:arg1 **x6/英雄**
:arg0-of x2/敢于-01
:arg0 x3/救-01
:arg0 **x6**
:arg1 x4/人

当然，这也可以说是 AMR 的体系所导致的，动词短语“敢于救”做定语时，“敢于”在上层，则要使用反关系 argX-of 形成“英雄-敢于”的关系，加上“英雄-救”的关系，便会形成环。如果不使用反关系，则会导致一个句子出现多个根（“感谢”和“敢于”），丧失了树形主干，破坏了整体结构，所以被 AMR 舍弃。这种技术上的处理，可以进一步引发语言学上的探讨。比如在依存文法中，一般将“敢于”处理为“救”的状语，也不标注其论元结构。AMR 的标注方法则兼顾各个谓词的论元结构，不掩藏问题，能够更好地表示句子的语义结构，却引发了环。我们在设计标注软件时，对于带环的句子在输出时进行了约束处理，使其避免出现死循环的情况。

在 CAMR 语料中，有 76 个带环的句子，约占句子总数的 1%，这个比例超过了英文 AMR 语料的 0.3%。这些带环的句子，基本上都是由动词短语做定语引发。而 CAMR 所标注的丰富的论元结构，能够有助于发现和讨论“环”的问题。

4.4 非投影树结构

非投影树，是依存语法中存在的特殊现象，指依存树上的结点垂直投影到句子上出现交叉的现象。这种现象对于语言学来说，具有较高的研究价值，对于自动分析的算法来说存在挑战。如，图 4 中的句子“他看到一条狗昨天，瞎的”，在依存树向原句做投影时，就会出现“昨天”和“瞎”的交叉。非投影的句子在英文等其他语料库的建设中几乎是普遍存在的，而在汉语研究和语料库建设中尚未引起重视。



注：图中的箭头表示依存关系，虚线表示投影关系。

图4：依存树的非投影句示例

非投影结构之前极少被讨论，原因是依存语法在早期严格排斥非投影树，只允许投影树的存在[17][18]。而在欧洲语序自由语言的的语料标注过程中，学者们感到投影树的限制太强，不利于表示句子的结构，于是提出突破投影树，采用无限制的依存树（unrestricted dependency trees）进行句子结构的描写[19]。之后，多种语言的描写都采用了这种无限制依存树。文献[20]对 12 种语言的依存树库进行了非投影树的统计，西班牙语最低 1.72%，荷兰语最高 36.44%，其他语言如日语 5.29%、阿拉伯语 11.16%、捷克语 23.15%、德语 27.75%等，均有着较多的非投影树。虽然国际上也有中文依存分析的评测 CoNLL06、07 和其他语料，但这些依存树库在人工标注时有意

识地遵循了投影树原则，或者是从短语结构树库转换而来，所以无法统计出中文语料的非投影树。文献[21]对“中文非投射语义依存现象”进行了讨论，但实际研究的是超越树结构的图结构，并不是严格的非投影树。

在缺乏对齐信息的英文 AMR 语料上很难进行非投影树的数量统计。而本文的融合对齐信息的 CAMR 语料，可以做出具体的统计分析。AMR 虽然采取了有向无环图结构，但主体架构依然是单根树，加之我们的对齐信息，使得我们能够从语料中获取含有非投影子树的句子。CAMR 语料中有 2238 个句子含非投影子树，占句子总数的 32.3%。经过初步分析，非投影句主要有两种情况，一种是话题化，如“软环境，我看行”。如果恢复为投影树则为“我看软环境行”。而过去的中文依存树库在标注过程中把话题作为主语，没有很好地表示出句子的语义结构。第二种是源于 AMR 的体系将情态词作为谓词的父节点。例如，“他可以住这儿”，按 AMR 的规范，“可以”作为“住”的父节点，就会导致非投影子树。当然，还存在离合词造成的非投影等其他几种情况。

表9：非投影句的CAMR标注示例

软 1 环境 2 ，3 我 4 看 5 行 6	他 1 可以 2 住 3 这儿 4
x5/看-11 :arg0 x4/我 :arg1 x6/行-04 **:arg0 x2/环境** **:arg0-of x1/软-01**	**x2/可以** :arg0 x3/住-01 :arg0 x1/他 :arg1 x4/这儿

英文 AMR 缺乏对齐信息，难以提取非投影句，中文 CAMR 融合对齐信息的一体化标注体系能有效地获取非投影句，并易于定位非投影结构。关于非投影结构的细致分析和理论探讨，我们将另文展开。

5 结论与未来工作

针对 AMR 语料缺乏对齐信息的不足及中文 AMR 标注过程中存在的问题，本文首先提出了融合概念对齐信息的一体化标注方法和标注平台，解决了中英文输入法频繁切换的问题、增加了错别字纠正和未标注词高亮功能，提高了标注效率。其次，在标注体系上针对中文的语言现象进行了概念和关系的调整。然后，在标注完成的 6923 句的 CTB 语料上，对图、环和论元共享问题进行了统计分析，为自动分析提供了相应的建议。最后，借助中文 AMR 的对齐信息，我们得以统计出含中文非投影子树的句子比例，这对于语言学研究和自动分析算法的设计都具有较大价值。

在今后的工作中，我们将继续扩大中文 AMR 的语料标注规模，进行语言学理论研究和自动分析算法研究。首先，将 CTB 语料标注完整，并扩展至其他领域语料，发布给学界使用。其次，借助 CTB 中已有的树形结构和对齐信息，比较 AMR 和传统句法树的异同与价值。最后，在较大规模的中文 AMR 语料基础上进行自动分析技术的研究，建立中文 AMR 自动分析系统。

参考文献

[1] Katz J J, Fodor J A. The Structure of a Semantic Theory [J]. Language, Vol. 39(2), 1963: 170–210.
[2] Montague. Universal Grammar[J]. Theoria, 1970, Vol.36: 373–398.
[3] Jackendoff R. Towards an Explanatory Semantic Representation [J]. Linguistic Inquiry, Vol. 7(1), 1976: 89–150.
[4] Banarescu L, Bonial C, Cai S, et al. Abstract Meaning Representation for Sembanking [C]//Proceedings of the 7th Linguistic Annotation Workshop, Sophia, Bulgaria, 2013: 178–86.
[5] Bos J. Expressive Power of Abstract Meaning Representations [J]. Computational Linguistics, 2016, Vol.42(3): 527–535.
[6] May J. SemEval-2016 Task 8: Meaning Representation Parsing [C]//Proceedings of SemEval-2016, San Diego, California, 2016: 1063–1073.
[7] Pourdamghani N, Gao Y, Hermjakob U, et al. Aligning English Strings with Abstract Meaning Representation Graphs [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 425–429.
[8] Xue N, Bojar O, Hajič J, Palmer M, et al. Not an Interlingua, but Close: Comparison of English AMRs to Chinese

and Czech [C]//Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, May 26-31, 2014: 1765–1772.
[9] 李斌，闻媛，卜丽君，曲维光，薛念文. 英汉《小王子》AMR 语义图结构的对比分析[J]. 中文信息学报，2017 年第 1 期：50–57.
[10] Palmer M, Daniel G, Paul K. The Proposition Bank: An Annotated Corpus of Semantic Roles [J]. Computational Linguistics, 2005, Vol.31(1): 71–106.
[11] Wang Y, Guo J, Che W, et al. Transition-Based Chinese Semantic Dependency Graph Parsing [C]//Proceedings of China National Conference on Chinese Computational Linguistics. Yantai, China. 2016: 12–24.
[12] Chen B, Ji D. Chinese Semantic Parsing Based on Dependency Graph and Feature Structure [C]//International Conference on Electronic14 and Mechanical Engineering and Information Technology. 2011, Vol.4: 1731–1734.
[13] 袁毓林，詹卫东，施春宏. 汉语"词库—构式"互动的语法描写体系及其教学应用[J]. 语言教学与研究，2014 年第 2 期：17–25.
[14] Banarescu L, Bonial C, Cai S, et al. Abstract Meaning Representation (AMR) 1.2.2 Specification[DB/OL]. [2015]. https://github. com/amrisi/amr-guidelines/blob/master/amr.md.
[15] Cai S, Knight K. Smatch: an Evaluation Metric for Semantic Feature Structures [C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, August 4-9, 2013: 748–752.
[16] Xue N, Xia F, Chiou F, Palmer M. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus [J]. Natural Language Engineering, 2005, Vol.11(2): 207–238.
[17] Hays D. Dependency Theory: A Formalism and Some Observations [J]. Languages, Vol.40(4), 1964: 511–525.
[18] Percival W K. Refelections on the History of Dependency Notions in Linguistics [J]. Historiographia Linguistica 1990, Vol.17: 29–47.
[19] Holan Tomăš, Vladislav Kuboň, Karel Oliva, and Martin Plătek.1998. Two Useful Measures of Word Order Complexity [C]//Alain Polguère and Sylvain Kahane, eds. Proceedings of Dependency-Based Grammars Workshop, COLING/ACL: 21–28.
[20] Havelka Jiří 2007. Beyond Projectivity: Multilingual Evaluation of Constraints and Measures on Non-Projective Structures [C]//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic: 608–615.
[21] 郑丽娟，邵艳秋，杨尔弘. 中文非投射语义依存现象分析研究[J]. 中文信息学报，2014 年第 6 期：41–47.

李斌（1981—），博士，副教授，主要研究领域为计算语言学。
E-mail: libin.njnu@gmail.com

闻媛（1992—），硕士生，主要研究领域为计算语言学。
E-mail: wenyuan.njnu@gmail.com

宋丽（1993—），硕士生，主要研究领域为计算语言学。
E-mail: songli1105@sina.com

作者联系方式：李斌　地址：江苏省南京市鼓楼区宁海路 122 号南京师范大学随园校区文学院
邮编：210097　电话：13813878144 电子邮箱：libin.njnu@gmail.com