文章编号：1003-0077(2011)00-0050-08

英汉《小王子》抽象语义图结构的对比分析

李 斌[1]，闻 媛[1]，卜丽君[1]，曲维光[2]，薛念文[3]

(1. 南京师范大学 文学院，江苏 南京 210097；
2. 南京师范大学 计算机科学与技术学院，江苏 南京 210023；
3. 布兰迪斯大学 计算机系，美国 沃尔瑟姆 02453)

摘 要：AMR(抽象语义表示)是国际上一种新的句子语义表示方法，有着接近于中间语言的表示能力，其研发者已经建立了英文《小王子》等 AMR 语料库。AMR 与以往的句法语义表示方法的最大不同在于两个方面，首先采用图结构来表示句子的语义；其次允许添加原句之外的概念节点来表示隐含的语义。该文针对汉语特点，在制定中文 AMR 标注规范的基础上，标注完成了中文版《小王子》的 AMR 语料库，标注一致性的 Smatch 值为 0.83。统计结果显示，英汉双语含图结构句子具有很高的相关性，且含有图的句子比例高达 40%左右，额外添加的概念节点则存在较大差异。最后讨论了 AMR 在汉语句子语义表示以及跨语言对比方面的优势。

关键词：抽象语义表示；语义图；英汉对比；自然语言处理

中图分类号：TP391 **文献标识码**：A

A Comparative Analysis of the AMR Graphs Between English and Chinese Corpus of *the Little Prince*

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

(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. Computer Science Department, Brandeis University, Waltham, MA 02453, USA)

Abstract: AMR is a new representation of the abstract meaning of a sentence, which is close to the Interlingua. The English AMR corpus including *the Little Prince* has been released. The major differences between AMR and the previous syntactic and semantic representation lie in two aspects. First, AMR uses a graph. Second, it allows adding concept nodes which are omitted in a sentence. In this paper, we design the Chinese AMR annotation specification and construct the Chinese *Little Prince* AMR corpus, achieving an inter-agreement Smatch value is 0.83. The bilingual comparison shows that the graph structures in English and Chinese sentences are highly correlated. With a proportion of 40% sentences having graph structure. But the added concept nodes are different. We also discuss AMR's ability to represent the semantic meaning of Chinese sentences as well as the advantages of AMR in cross language comparison.

Key words: abstract semantic representation; semantic graph; English-Chinese comparison; natural language processing

1 引言

抽象语义表示(Abstract Meaning Representation)简称为 AMR，是一种新型的句子语义表示方式，由美国宾夕法尼亚大学的语言数据联盟(LDC)、南加州大学、科罗拉多大学等科研机构的多位学者共同提出[1]。与传统的基于树的句法语义表示方法

收稿日期：2016-09-15 定稿日期：2016-10-20

基金项目：江苏高校哲学社会科学研究项目(2016SJB740004)；国家科技支撑计划课题(2014BAK04B02)；国家自然科学基金(61272221)

不同，AMR 使用单根有向无环图①来表示一个句子的语义。这种表示方法相比树结构拥有较大的优势：首先，单根结构保持了句子的树形主干；其次，有向无环图使用图结构可以较好地描写一个名词由多个谓词支配所形成的论元共享（argument sharing）等现象；第三，AMR 还允许补充出句中隐含或省略的成分，以还原出较为完整的句子语义。这三大优点，使得 AMR 一经公布，就引起了国际上的重视，涌现了从跨语言翻译价值角度进行的讨论[2]、自动分析技术[3]、转化应用[4]等多方面的研究论文。AMR 配套发布了包括《小王子》在内的两万多句英文语料库，2016 年的 SemEval 语义评测也举办了英文 AMR 的自动分析竞赛项目②。英文《小王子》语料中带有图结构的句子比例高达 42%③，带有补充概念节点的句子比例也在 10%以上，说明了 AMR 使用图结构和补充概念节点的有效性和合理性，也使得学术界对于句子的结构有了新的认识。

另一方面，汉语的句法语义自动分析研究，也开始从句法树走向了语义图。Ding 等[5]加工了汉语语义依存图库，其中带有图结构的句子仅 10%左右。虽然句子的标注体系不同，也没有增添概念节点的机制，但相比英文《小王子》带有图结构的句子比例 42%，仍有较大差异。这促使我们试图分析英文的图结构到底由哪些因素造成，汉语中图结构的情况又如何。为了使中英文数据能够在可比较的语料库上进行分析，我们根据英文 AMR 的标注规范[6]，设计了中文 AMR 标注规范，标注了和英文《小王子》句对齐的中文《小王子》1 562 句。由两位语言学研究生分别独立标注，标注一致性的 Smatch 值为 0.83。统计结果显示，中文《小王子》含有图的句子比例也高达 36%左右，且与英语具有很高的一致性。而英汉双语的补充的概念节点的数量却存在较大差异，体现出语言结构的差异。

2　AMR 简介及相关研究

2.1　AMR 简介

AMR（Abstract Meaning Representation，抽象语义表示）是句子语义的一种表示方法，将一个句子的语义抽象为一个单根有向无环图。在这个语义图上，句子中的实词抽象为概念节点，实词之间的关系抽象为带有语义关系标签的有向弧，同时忽略虚词和形态变化体现的较虚的语义（如 the、单复数、时、体等等）。图 1 分别给出了“The boy wants to go to school”及中文翻译“男孩想去学校”的 AMR 表示。

图 1　“The boy wants to go to school”的 AMR 英汉表示方法

图 1 中，每个概念节点都有一个字母开头的编号。“想（want）”作为句子唯一的根节点，编号分别是 x 和 w，“男孩（boy）”作为“想（want）”的 arg0（施事），“去（go）”作为“想（want）”的 arg1（受事）。这里与传统的句法分析或语义角色标注有一些差异，英文做了词形还原，省略了冠词 the、形态标记（动词的数、介词 to），而汉语则没有词形方面的变化。与传统表示方法的主要不同在于对论元共享现象的处理，例如“想（want）”和“去（go）”的 arg0 都是“男孩（boy）”。传统的句法分析方法受限于树结构，往往舍弃“男孩-去”这个关系；而语义角色标注会保留两个关系，形成图结构。AMR 为了保留论元共享的信息，又避免图结构的凌乱显示，允许重复使用词语的编号 b 和 x1，使得 AMR 在保持树状层次结构的同时，保有图结构的信息。

为了明确谓词及其论元之间的语义关系，AMR 要求标注谓词的具体义项。因为一个谓词会有多个义项，而不同义项下的论元框架会存在差异。在图 1 中，动词“想（want）”被标注了“－01”的信息，表示此处的“想（want）”使用的是其第一个义项的论元框架。

AMR 暂时忽略语言中语义较虚的成分，如英文中“名词的数、动词的数、有定/无定、时、体”等由形态变化体现的语义。而它最令人称道之处，在于它允许根据整体语义增删概念节点，能够弥补传统句法表示的严重缺陷。例如，The injured was taken home.（受伤的被送回家了）。在短语结构文法和依存文法的框架下，The injured（受伤的）只能作为一

① Banarescu 等[6]指出，在技术操作上仍有约 0.3%的句子的 AMR 结构存在环。

② http://alt.qcri.org/semeval2016/task8/。

③ 虽然 AMR 采用单根有向图表示句子语义，但很多句子没有形成图结构，仍为树结构。英文《小王子》语料中，剩余 58%的句子为单根树结构。

个整体来处理，其语义难以得到揭示。

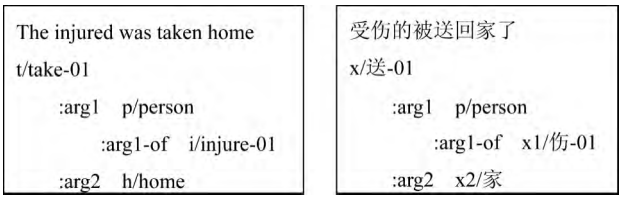


图2 “The injured was taken home”的AMR英汉表示方法

图2给出了AMR的处理方式。AMR允许补充句子中省略的成分，将“person(人)”补充出来，作为“take(送)”的arg1(受事)，也作为“injure(伤)”的arg1(受事)，更完整地表示了句子的语义。

AMR的补充概念节点和删除语义较虚的词语的方式，对汉语来说也很重要。一方面，汉语的“的”字结构(如“受伤的”)，在传统的句法分析中也被当作一个整体来对待，难以体现出其真正的语义。而在AMR的补充概念的方式下，person(人)的补充使得意义得到了较为完整的表达，“受”的被动义也由“person : arg1-of 伤”描写出来，体现出AMR对于中文语义表示的价值。

另一方面，AMR也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明确。比如，“他回答说”可以省略为“他回答”。此外，AMR还规定了一部分近义词可以使用最常见或歧义较少的单词进行替换，如在句中表示“好像”的意思的“like”替换为“resemble-01”。

AMR的抽象语义表示方法给句子语义以更加清晰的表达，受到学界的密切关注，但也褒贬不一[2]。赞扬者认为这种表示方法整体上简洁有效，弥补了句法树在表示语义上的缺陷，接近真正意义上的中间语言(interlingua)；批评者则认为忽略形态变化所表达的意义是难以接受的。不过这一缺点对于汉语来说并不那么重要，因为汉语本来就没有形态变化，甚至被一些语言学家称作“语义型语言”[7]。从上面的例子我们也可以看出，由于没有形态变化，汉语表示为AMR以后，损失的信息远比英文少。换言之，相比英文、德文等具有形态变化的印欧语言，AMR更适于表示汉语的句法语义。除去形态变化，汉语在句法分析时遇到的常见难题，如造成论元共享的连动句、兼语句等可以通过图结构得到很好的解决；“的”字结构等省略句子成分的结构也可以通过补充概念来解决。

AMR对于句子语义较为简洁而完整的表示、可计算评测的特点，使其至少具有三点潜在价值：(1)提升智能问答、文本摘要、事件分析等应用技术；(2)作为机器翻译的中间语言，提升机器翻译效果；(3)为句子级别之上的篇章语义表示奠定研究基础。因此，提高AMR的自动分析效果，增加更多语种的AMR语料就成为目前该领域最为迫切的研究内容。而中文AMR语料的构建一方面可以满足中文句法语义分析的应用需求，另一方面对于汉语的句法语义研究也有重要的语言学价值。

2.2 句子的图结构研究

传统的句法分析以树作为句子的基本结构[8-9]。而随着框架语义学(Frame Semantics)的兴起[10-11]，语义角色的标注(Semantic Role Labelling)工作也逐步展开[12]。当一个句子中多个谓词共享同一个名词性成分时，多个谓词及其语义角色就会形成图结构。根据2009年依存和语义角色标注评测CoNLL2009 Shared Task语料，英语和汉语由于语义角色的论元共享现象，出现了较多的图结构[13]。2014和2015年的SemEval国际评测则直接引入了语义依存图(Semantic Dependency Graph)，在DM、PAS、PCEDT三个英文语料上由重入的回边(reentrance)造成的图结构的句子分别占到了27.35%、29.40%和9.27%[14]。英文《小王子》AMR语料库上具有图结构的句子比例更高达42%。

英语句子的表示方法不仅使用了图结构，而且图结构的比例也确实较高。但是，汉语句子图结构的情况依然不够清楚。CoNLL2009语义角色标注数据[13]只标注了谓词及其论元的语义关系，所以并不能忠实地反映出汉语完整句子的图结构情况。借鉴Oepen等[14]的体系，Ding等[5]加工了中文语义依存图库，其中带有图结构的句子仅10%左右，与英文语义依存图和英文AMR的差异较大。Xue等[2]从机器翻译的中间语言角度，对英语、汉语和捷克语各100句的三语平行语料库进行了AMR的对比分析。其中，汉语的语料也出现了图结构。不过，100句的语料在规模上比较小，没有专门从图结构的角度进行分析。

因此，基于英汉平行语料构建更大规模的AMR语义图库，可以更好地比较两种语言中图结构的对应情况、图结构存在的比例、图结构的共性和差异等，以进一步观察AMR的跨语言表示能力和AMR对于汉语的句法语义表示能力。

3　中文《小王子》AMR 的标注

《小王子》英文 AMR 库①提供了 1 562 句的标注数据，并附带了句对齐的中文《小王子》生语料。在此基础上标注中文《小王子》的 AMR，便可得到英汉句对齐的双语 AMR 语料库。我们首先根据中文宾州树库(CTB)[15]的分词规范，对中文《小王子》语料进行了自动分词和人工校对；其次，参照英文 AMR 标注规范[6]，制定了中文 AMR 标注规范；然后，标注了中文《小王子》的 AMR 语料库[16]②。

制定中文 AMR 的标注规范，是一件难度较大的工作。现有的 AMR 规范毕竟是根据英语的语言现象制定出来的，对于汉语中特有的量词(本、台)、重叠式(认认真真)、离合词(帮忙—帮了一个忙)、动补结构(跑得快、吃不了)等现象，还缺少具体的规定和处理方法。我们参考 AMR 的基本原则，经过大量的试标与讨论，制定出较为详细的标注规范。限于篇幅，现简述如下。

(1) 语义关系参照 AMR 的标准，分为核心语义关系与非核心语义关系。核心语义关系与英文 AMR 相同，沿用 Propbank[12] 和 Chinese Propbank[15] 的标注体系，共有五个：ARG0(原形施事)、ARG1(原形受事)、ARG2(间接宾语、工具等)、ARG3(出发点、受益者等)、ARG4(终点)。非核心语义关系，包括 accompanier(伴随)、age(年龄)、beneficiary(受益者)等共计 43 个。

此外，还有一些比较特殊的关系标签，如 and(和)、or(或)等概念的分项关系 op1、op2 等，用于 multi-sentence(句群)的分项关系 snt1、snt2 等。

(2) 按照 AMR 省略较虚的语义成分的原则，汉语特有的量词“本、张、台”等应该被省略，重叠式“认认真真”应该被还原为“认真”。

(3) 汉语离合式采取“合”的方式，如“帮了一个忙”的谓词合并为“帮忙”。

(4) 对于汉语中较为复杂的动补结构，根据句子中的具体语义进行标注。动补结构通常分为多种类型，如表示程度的“跑得快”、表示可能的“吃不了”、表示体的“做完作业”、表示结果的“看清楚、跑丢”等，均在规范中予以规定。

(5) 对于汉语“的”字结构为代表的需要补充概念节点的情况，也分门别类地予以规定。

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[17]。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 24 510 个中文谓词(包括动词、形容词等)的 26 650 个义项的不同语义角色框架。这部词典较好地覆盖了《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 AMR 规定的语义关系中选取。

中文《小王子》的 AMR 数据，由两位语言学研究生分别独立标注(语料 A、B)，标注一致性的 Smatch 值[18] 为 0.83，与英文小王子的标注一致性达到了同等质量[1]。

4　英汉对比统计和分析

下面对本文使用的两个标注语料进行含有图结构的基本情况进行统计对比，并进行相应的统计检验，观察《小王子》英语语料和《小王子》汉语语料的差异性和相关性。具体来说，统计英汉对齐的句子中是否含有图结构以及含有图结构的个数，检验汉语和英语中图结构存在情况的差异性和一致性，并对产生差异的原因进行分析。

4.1　基本统计数据

对于英汉《小王子》全部 1 562 句语料，汉语的两份人工标注结果(A、B)和英语人工标注的结果呈现出一定的共性和差异。表 1 给出了三份语料的图结构的统计数据。英语语料中，总共出现了 1 293 条回边，造成了 663 个图结构的句子。而汉语的语料 A 和 B 仅分别出现了 1 037 和 1 040 条回边，分别造成 548 和 576 个句子出现图结构。

表 1　《小王子》英汉 AMR 语料库的图结构统计

统计项	含图结构句数/回边条数	总句数	含图结构比例/%
《小王子》英语语料	663/1 293	1 562	42.45
《小王子》汉语语料 A	548/1 037	1 562	35.08
《小王子》汉语语料 B	576/1 040	1 562	36.88

表 1 的数据体现出：(1)英汉双语出现图结构的句子都较多。两种语言都有约 40%的句子出现了图结构。(2)英语的图结构比汉语略多一些。英语含有图结构的句子比例为 42.45%，汉语的比例则略低，分

① 语料下载地址 http://amr.isi.edu/。

② 语料下载地址 http://www.cs.brandeis.edu/~clp/camr/camr.html。

别为35.08%和36.88%[①]。从形成图结构的回边的数量看,英语也是略高于汉语。(3)英汉双语出现图结构的一致性较高。英语句子出现图结构,则有着对译关系的汉语句子也倾向于出现图结构。Pearson检验显示,汉语A和B两个语料与英文语料是否含有图结构的相关系数分别为0.555和0.565;而单个句字中含有图结构的数量的相关系数为0.695和0.705。这些结果均在0.01的水平上显著。

这三点统计结果,已经可以回答本文的基本问题,即汉语和英语的图结构比例到底相差多少。在双语平行语料上,能够清楚地看出具有图结构的句子数量较大,AMR的图结构的表示方法具有合理性。但是,我们依然想弄清楚,形成图结构的回边的比例,以及英语图结构的句子多于汉语的原因。

4.2 图结构对比分析

汉语和英语产生图结构的主要原因都是语义角色的共享,即同一个语义角色被不同的论元结构所分享。和图1相似,《小王子》语料中出现了大量的论元共享的句子,见图3。

So then I chose another profession, and learned to pilot airplanes.	后来，我只好选择了另外一个职业，我学会了开飞机。
(c2 / cause-01 :ARG1 (a / and :op1 (c / choose-01 **:ARG0 (i / i)** :ARG1 (p / profession :mod (a2 / another))) :op2 (l / learn-01 **:ARG0 i** :ARG1 (p2 / pilot-01 **:ARG0 i** :ARG1(a3/ airplane)))))	(x2 / temporal :arg2 (x4 / and :op1 (x5 / 选择-01 **:arg0 (x6 / 我)** :arg1 (x7 / 职业 :mod (x8 / 另) :quant (x9 / 1))) :op2 (x10 / 学会-01 **:arg0 (x11 / x6)** :arg1 (x12 / 开 **:arg0 (x13 / x6)** :arg1 (x14 / 飞机)))))

图3 英汉双语的论元共享实例

图3中,无论是英语的"So then I chose another profession, and learned to pilot airplanes."还是汉语的"后来,我只好选择了另外一个职业,我学会了开飞机。"都涉及三个主要动词"选择(choose)"、"学会(learn)"和"开(pilot)"。而这三个动词都共同分享了同一个施事(arg0)——我(I),在语义图中就会有三条弧指向"我(I)"这个词,形成图结构。论元共享无论在汉语还是在英语中都十分普遍,这也是汉语和英语中图结构比例都较高的主要原因。而传统的短语结构语法和依存语法都不允许出现图结构,到了框架语义学、依存图和AMR的研究中,才使用图结构。而AMR体系下,含有图结构的句子比例更高。所以,在论元共享中哪些语义角色特别容易引起图结构就成为需要统计的对象。

表2针对汉语和英语做了相应的统计,发现:(1)汉语和英语图结构中arg0、arg1和arg2共享引发图结构的情况都比较普遍,英语语料中占到77.11%,汉语语料(B)中占到85.19%;(2)两种语言中,arg0、arg1和arg2的数量依次递减。arg0(原型施事)所占的比例约在一半以上,arg1(原型受事)和arg2(原型与事)的比例较低;(3)英汉差异较大之处在于,汉语中的arg0比例明显较多(71.06%),明显多于英语(46.64%),也就是说,汉语中原型施事(arg0)的共享是比较普遍的,虽然英语中这种情况也比较多,但是在分布上比汉语更加均匀一些。

从语法的角度来解释这种差异性并不难。我们知道arg0无论在汉语还是英语中做主语的情况是比较多的,而汉语中主语省略情况也相对较多。当多个动词连续出现时,汉语更倾向于将一个主语放在最前面,后面的主语承前省略。而英语则更加注重句子结构的完整性。所以汉语中这种由于arg0的共享产生图结构的比例就更高,也是可以理解的。

除此以外,表3也给出了其他类型的语义角色共享导致的图结构情况的比例,主要是非核心语义关系下的图结构。在非核心语义关系内,poss(领属关系)和domain(系动词关系)也造成了较多的图结构。需要注意的是,由于种类较多,这里并没有穷尽所有的语义关系。

① 人工对比后发现,语料B的图结构数量略多,且A、B之间的差异对后文的统计影响较小,所以后文统计仅使用汉语语料B的数据。

表 2　汉语和英语中 arg0、arg1、arg2 的共享导致的图结构

语料类别	图结构总个数	ARG0	ARG1	ARG2	总计
汉语语料 B	1040	739	128	19	886
	100.00%	71.06%	12.31%	1.83%	85.19%
英语语料	1 293	603	298	96	997
	100.00%	46.64%	23.05%	7.42%	77.11%

表 3　汉语和英语中其他语义关系共享导致的图结构

语料类别	图结构总个数	ARG3	poss	domain	location	part-of	beneficiary	cause
汉语语料 B	1 040	6	88	20	2	12	9	3
	100.00%	0.58%	8.46%	1.92%	0.19%	1.15%	0.87%	0.29%
英语语料	1 293	1	136	46	3	53	7	0
	100.00%	0.08%	10.52%	3.56%	0.23%	4.10%	0.54%	0.00%

汉语图较英语图较少的原因，主要在于汉语翻译得较为简洁。我们分析了英语存在图结构而汉语是树结构的句子，多是源于英语句子较长、语义关系较为复杂所致。例如，“I answered you with the first thing that came into my head”这句话，“I”既作为“answer”的 arg0，即“the person who answers”，同时，“head”与“I”又是“part-of”的关系（身体的一部分），这样一来，就形成一条回边，造成图结构。但是汉语的句子则十分简洁，译为“我是随便回答你的”，没有出现图结构。

整体上来说，英汉《小王子》的句子中出现图结构的比例较为接近，且呈现出较高的相关性。造成图结构的原因在于论元共享，尤以 arg0、arg1、arg2 和 poss 为主。

4.3　添加的概念节点分析

AMR 允许添加概念节点，是其与传统的句法语义分析体系最大的不同。通过 thing（物）、person（人）、company（公司）等概念节点的添加，可以使得句子语义的表示更为自然和完整。AMR 拥有一个完整的命名实体概念集合，可以用来表示添加的概念节点，thing 和 person 只是最为常用的两个概念。不过，AMR 并不标注每个概念与原来句子中的词语的对应关系，这些命名实体也以英文单词表示，和句子中的词语没有形式上的差别，所以统计英文《小王子》中概念节点的添加情况较为困难①。我们只以最为常见的 thing 和 person 两个概念，来观察概念添加在两种语言中的作用。表 4～表 7 根据概念添加的类型，分别给出了英汉《小王子》中添加这两种概念的统计数据。Thing 在英汉语料中分别出现了 86 次和 38 次，person 在英汉语料中分别出现了 97 次和 8 次。概念添加在英语语料中所出现的句子比例超过了 10%，而汉语句子的比例较低。

表 4　英语语料中添加 thing 的统计

thing	名词内部分析	what、how 等引导的从句	some of it 等	其他	总计
出现次数	55	22	3	6	86
出现次数比例/%	63.95	25.58	3.49	6.98	100
出现句子数	52	22	3	6	83
出现句子数比例/%	62.65	26.51	3.61	7.23	100

① 有专门针对 AMR 概念和原句词语对齐的研究，如 Pourdamghani 等[19]，对齐的 Smatch 值为 90%左右。

表5 汉语语料B中添加thing的统计

thing	的字结构	数量结构	所字结构	“碰到什么吃什么”	补回其他省略	总计
出现次数	14	17	3	1	3	38
出现次数比例/%	36.84	44.74	7.89	2.63	7.89	100.00
出现句子数	13	16	3	1	3	36
出现句子数比例/%	36.11	44.44	8.33	2.78	8.33	100.00

表6 英语语料中添加person的统计

person	指人名词内部分析	have-role	include	who引导的从句	其他	总计
出现次数	65	18	5	4	5	97
出现次数比例/%	67.01	18.56	5.15	4.12	5.15	100
出现句子数	64	18	5	4	3	94
出现句子数比例/%	68.09	19.15	5.32	4.26	3.19	100

表7 汉语语料B中添加person的统计

person	的字结构	其他省略	总计
出现次数	7	1	8
出现次数比例/%	87.50	12.50	100.00
出现句子数	5	1	6
出现句子数比例/%	83.33	16.67	100.00

从这些数据可以看出：(1)概念添加对于英语来说作用更大。英语中由词缀或形态变化构成的名词，往往被AMR进行内部分析。例如，带有-ing的drawing(图画)分析为thing : arg1-of draw(thing是画的受事)，带有-er的admirer(仰慕者)分析为person : arg0-of admire(person是仰慕的施事)等。这种描写方式的优劣也许存在争议，但是对于what/how等引导的从句来说，显得不可或缺。例如，what you like，处理为thing : arg1-of like和like : arg0 you. 对于“some of it”之类的短语，补充为“some+thing of it”也显得更为完整。(2)汉语中出现的数量略少，但对于刻画“的”字结构、“所”字结构、数量结构非常有效。如前文所述，AMR的概念添加方式对于“的”字结构有着良好的表示能力，能够补充出转指的成分thing、person等。“所”字结构如“所思”、“所想”、“所言”等，一般省略了动词的宾语，借助thing等概念可以很好地补充出来。数量结构，如承接上文省略的“我也买了一个”和连动结构中的“吃一个少一个”，都省略了名词性成分，也需要根据上下文来补充概念节点。(3)如果去掉词语内部结构的分析造成的概念添加，则英汉双语在补充原句中省略的词语方面数量较为接近。英语补充thing的总数86减去词语内部分析的55，则剩余31个较为纯粹的添加操作，与汉语添加thing的36个非常接近。英语补充person的总数97个，减去名词内部分析的65个，剩余32个较为纯粹的概念添加操作，与汉语的8个差距缩小了很多。一方面，AMR对词语内部的分析，刻画出英汉在构词和形态变化上的差异；另一方面，AMR通过概念添加的方式对两种语言句子中省略成分的补充较为有效。这也加深了我们对于两种语言的理解。一般来说，英语比较强调句子结构的完整性，而汉语句子中成分省略现象较多。但是通过AMR的标注数据来看，英汉都存在成分省略的现象。最为可贵的是，AMR的这种标注方法使得英汉句子在语义层面上得到了较为接近的表示，显示了其充当跨语言翻译的中间语言的潜力。

5 结论及未来工作

本文通过标注汉语《小王子》AMR语料库，与英文《小王子》AMR语料库进行对比分析，得出的主要结论是：(1)汉语和英语中都含有较高比例的图结构，分别为36%和42%左右，说明图结构在汉语和英语中都是普遍存在的；(2)添加概念节点的方式能够更好地描写句子中省略的词语的语义。特别对于汉语的“的”字结构、“所”字结构和数量结构，具有良好的补充能力。

这两点结论体现出AMR确实具有良好的句子语义表征能力。一方面，具备图结构的句子比例较高说明图结构的引入确有必要；另一方面，AMR能

够补充出句子中省略的成分，以完整地表征句子的语义，便于进行跨语言的比较。

当然，本文的工作还是初步的，需要在以下几个方面深入研究。首先，统计分析英汉《小王子》语料库中每一个句子在 AMR 表示上的异同，以进一步探究 AMR 的跨语言表示能力和英汉两种语言本身在词汇和句法上的差异；其次，标注更大规模的汉语 AMR 语料库，以研究汉语的句法语义问题，同时为汉语 AMR 自动分析技术提供训练和测试数据。然后，与英语、捷克语等其他语言的 AMR 语料库进行跨语言对比研究；最后，AMR 是句子级别的语义表示方法，汉语中的成分省略特别是主语省略情况会导致我们处理时丢失一些语义上应该存在的图结构，还需要考虑篇章级别 AMR 的标注方法。

参考文献

[1] Banarescu L, Bonial C, Cai S, et al. Abstract Meaning Representation for Sembanking[C]//Proceedings of the 7th Linguistic Annotation Workshop, Sophia, Bulgaria, 2013.

[2] Xue N, Bojar O, Hajič J, et al. Not an Interlingua, but Close: Comparison of English AMRs to Chinese and Czech[C]//Proceedings of the 9th International Conference on Language Resources and Evalua-tion (LREC'14), Reykjavik, Iceland, May 26-31, 2014: 1765-1772.

[3] Flanigan J, Thomson S, Carbonell J, et al. A Discriminative Graph-Based Parser for the Abstract Meaning Representation[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 1426-1436.

[4] Liu F, Flanigan J, Thomson S, et al. Toward Abstractive Summarization Using Semantic Representations Human Language Technologies[C]//Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL, Denver, Colorado, May 31-June 5, 2015: 1077-1086.

[5] Ding Y, Shao Y, Che W, et al. Dependency Graph Based Chinese Semantic Parsing[C]//Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer International Publishing, 2014: 58-69.

[6] Banarescu L, Bonial C, Cai S, et al. Abstract Meaning Repre-sentation (AMR) 1.2.2 Specification[DB/OL]. [2015]. https://github.com/amrisi/amr-guidelines/blob/master/amr.md.

[7] 徐通锵. 语言论——语义型语言的结构原理和研究方法[M]. 长春：东北师范大学出版社. 1997.

[8] Chomsky N. Syntactic Structures[M]. The Hague/Paris: Mouton, 1957.

[9] Tesnière L. Eléments de syntaxe structurale[M]. Paris: Librairie C. Klincksieck, 1959.

[10] Fillmore C J. Frame Semantics[J]. Encyclopedia of Language & Linguistics, 2006: 613-620.

[11] Baker Collin F, Charles J Fillmore, John B Lowe. The Berkeley FrameNet Project[C]//Proceedings of COLING/ACL-98, Montreal, 1998: 86-90.

[12] Palmer M. Daniel G, Paul K. The Proposition Bank: An Annotated Corpus of Semantic Roles[J]. Computational Linguistics, 2005, 31(1): 71-106.

[13] Hajič, Jan, Ciaramita M, et al. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages[C]//Proceedings of the 13th Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2009: 1-18.

[14] Oepen S, Kuhlmann M, Miyao Y, et al. SemEval 2014 Task 8: Broad-Coverage Semantic Dependency Parsing[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014: 63-72.

[15] Xue N, Xia F, Chiou F, et al. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus[J]. Natural Language Engineering, 2005, 11(2): 207-238.

[16] Bin Li, YuanWen, Lijun Bu, et al. Annotating the Little Prince with Chinese AMRs[C]//Proceedings of the 10th Linguistic Annotation Workshop. Berlin, Aug, 2016.

[17] Nianwen Xue, Martha Palmer. Adding Semantic Roles to the Chinese Treebank[J]. Natural Language Engineering, 2009, 15(1): 143-172.

[18] 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.

[19] 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.

（下转第 74 页）

孙世昶(1979—),博士,讲师,主要研究领域为机器学习与文本挖掘。
E-mail:ssc@dlnu.edu.cn

林鸿飞(1962—),通信作者,博士,教授,主要研究领域为文本挖掘和信息检索。
E-mail:lhf@dlut.edu.cn

孟佳娜(1972—),博士,教授,主要研究领域为文本挖掘。
E-mail:mjn@dlnu.edu.cn

(上接第 57 页)

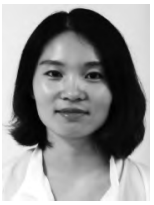
李斌(1981—),博士,副教授,主要研究领域为计算语言学。
E-mail:libin.njnu@gmail.com

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

卜丽君(1990—),硕士研究生,主要研究领域为计算语言学。
E-mail:blj_njnu@163.com

(上接第 65 页)

于东(1982—),博士,副教授,主要研究领域为自然语言处理。
E-mail:yudong_bluc@126.com



赵艳(1994—),硕士研究生,主要研究领域为语言信息处理。
E-mail:zhaoyan 0819@126.com

韦林煊(1995—),本科生,主要研究领域为语言信息处理。
E-mail:515984350@qq.com