

自然语言句子抽象语义表示 AMR 研究综述

曲维光^{1,2} 周俊生¹ 吴晓东¹ 戴茹冰² 顾 敏¹ 顾彦慧¹

(1. 南京师范大学计算机科学与技术学院, 南京, 210023; 2. 南京师范大学文学院, 南京, 210097)

摘 要: 句子的语义处理是自然语言处理的重要难题与挑战。抽象语义表示 (Abstract meaning representation, AMR) 是近几年国际上新兴的句子级语义表示方法, 突破了传统的句法树结构的限制, 将一个句子语义抽象为一个单根有向无环图, 很好地解决了论元共享问题, 成为语言资源建设和句子语义解析的研究热点。本文从 AMR 概念与规范、解析算法和应用等方面对 AMR 相关研究进行系统的梳理, 特别对 AMR 的各种解析算法进行了比较深入的分析 and 比较, 指出了现有算法存在的问题和不足, 同时介绍了中文 AMR 的开发进展, 最后展望了 AMR 未来的研究方向。

关键词: 抽象语义表示; 句法语义; 语义解析; 自然语言处理

中图分类号: TP391 **文献标志码:** A

Survey on Abstract Meaning Representation

Qu Weiguang^{1,2}, Zhou Junsheng¹, Wu Xiaodong¹, Dai Rubing², Gu Min¹, Gu Yanhui¹

(1. School of Computer Science and Technology, Nanjing Normal University, Nanjing, 210023, China; 2. School of Chinese language and literature, Nanjing Normal University, Nanjing, 210097, China)

Abstract: Semantic processing is a key challenge in natural language processing. Abstract meaning representation (AMR) is a novel framework of representing the meaning of a sentence. Instead of a tree, it abstracts the meaning of a sentence into a rooted acyclic directed graph, which solves the argument sharing problem. Thus, the corpus construction and automated parsing of AMR become a heated research field. This paper introduces the AMR's basic concept, annotation guidelines, parsing algorithm and applications. Then we discuss the problems and shortcomings of the parsing algorithms by comparison experiments. We also introduce the development of Chinese AMR researches. At last, we discuss the potentials of AMR, which are fruitful for Chinese semantic processing.

Key words: abstract meaning representation; syntactic semantics; semantic parsing; natural language processing

引 言

语义是语言形式所要表达的内在含义, 如何实现对自然语言句子的完整语义理解, 是人工智能和自然语言处理研究领域的一个重要研究目标。从某种意义上讲, 自然语言处理研究的最终目标就是在语义理解的基础上实现各种类型的应用, 但是由于语义的模糊性、多义性, 实现自然语言的语义理解是

基金项目: 国家自然科学基金(61272221, 61472191)资助项目; 江苏省高校哲学社会科学研究项目(2016SJB74004)资助项目。
收稿日期: 2016-11-27; 修订日期: 2017-01-20

一个巨大的挑战^[1]。近年来语义标注相关的研究有所进展,能够分别进行命名实体识别、词义消歧、语义角色关系和共指消解等各种单独语义的标注与分析任务,每一种语义任务都有各自的评测方法,而缺少一个针对整句进行逻辑语义表示的规范和语料库。为此,2013 年 Banarescu 等提出了一种语义表示语言^[2],即抽象语义表示(Abstract meaning representation, AMR),并开发了一个较大规模的标注语料库,它由自然语言句子和与其对应的用 AMR 形式表示的句子的逻辑语义图构成。AMR 建立了统一的标注规范,其简单可读的特点方便人们理解,AMR 表示的逻辑语义也方便计算机的处理。自此,对 AMR 解析与应用等相关问题的研究受到了国内外学者的高度关注,引发了一股 AMR 研究热潮。可以预见,今后数年里 AMR 一定会受到更多研究者的关注。

本文将从 AMR 的定义与规范、AMR 解析算法和 AMR 应用等方面对抽象语义表示相关研究进展进行系统的梳理,特别对各种 AMR 解析算法进行比较深入的分析和比较,并指出一些算法存在的问题和不足,同时对中文 AMR 的开发进行简介。最后展望了 AMR 未来的研究方向,以期对其在国内的研究起到一定的推动作用。

1 AMR 的定义与规范

1.1 AMR 的定义

传统的句子语义解析任务一般是针对一个特定的领域(例如地理数据库查询、航班信息查询等)设计一个相对简单的形式化的意义表示语言,然后再采用该意义表示语言对句子进行相应的逻辑意义标注^[3,4]。AMR 则是一种全新的、领域无关的句子语义表示方法,将一个句子的语义抽象为一个单根有向无环图。图 1 给出了一个英文句子的 AMR 图表示的示例。其中,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,且忽略虚词和形态变化体现较虚的语义

(如单复数、时态等)。这种表示方法相比树结构拥有较大的优势:(1)单根结构保持了句子的树形主干;(2)图结构的使用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现象;(3)AMR 允许补充句中隐含或省略的成分,以还原出较为完整的句子语义,能够更加全面地描写语义,并有利于语义的自动生成。

1.2 AMR 的标注规范

AMR 图中的每个节点表示一个语义概念,语义概念可以是单词(如“he”,“accent”等)、也可以是 PropBank framesets^[5](如“try-01”,“affect-02”等)或者是特殊的关键词。其中特殊关键词包括实体类型(如“country”,“date-entity”等)、量词(如

“monetary-quantity”,“distance-quantity”等)以及逻辑连词(“and”等)。有向边的标注表示的是两个概念之间的关系。值得注意的是,有多个语义角色的词在有向图中只用一个节点来表示。所以当实体在句子中的作用很重要时,允许给有向图中的节点设置多个父节点(如图 1 中的节点“he”有两个父节点“try-01”和“affect-02”)。总体上,AMR 涵盖约一百种概念关系,表 1 中给出了部分关系示例。AMR 还包括几乎所有关系的逆关系(如“:arg0”对应的“:arg0-of”、“:location”对应的“:location-of”等)。此外每一个关系都有相关的具体化形式。例如,当想要具体化关系“:location”时,它对应的具体化形式就是转化为概念为“be-located-at-91”的节点。所以 AMR 是一种可以扩展和修正的语义表示,详细的 AMR 规

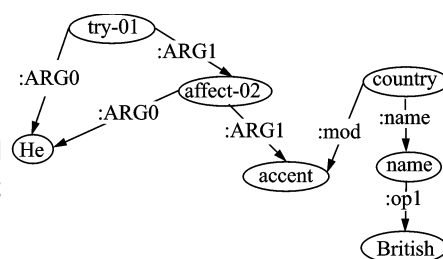


图 1 英文句子“He tries to affect a British accent.”的 AMR 图表示

Fig.1 AMR graph of sentence “He tries to affect a British accent.”

表 1 AMR 概念关系

Tab.1 AMR concept relations

| | |
|--------------------|---|
| PropBank framesets | :arg0, :arg1, :arg2... |
| 一般语义关系 | :age, :destination, :location, :name... |
| 日期实体关系 | :day, :month, :time... |
| 列表关系 | :op1, :op2, :op3... |
| 数量关系 | :quant, :unit, :scale... |

范参见 <http://amr.isi.edu>。AMR 概念和概念关系的定义可以表示出任意形式的句子。AMR 把所有的词用一种既合理又统一的方式考虑进来,能辅助完成很多基于语义的任务,在解决实际问题时有很大的潜在应用价值。

1.3 AMR 的评测方法

在设计好 AMR 表示规范后,AMR 解析的任务就是对给定的输入句子,预测和输出一个相应的 AMR 图结构。为评价一个 AMR 解析输出图结构的准确性,一种最简单的方法是整句准确性计算,即输出 AMR 图与人工标注 AMR 图(称为参考 AMR 图)完全一致时准确率为 1,否则为 0。显然这种评价方法粒度过粗,更合适的方法是评价一个输出 AMR 图结构的部分准确率,即准确率在 0-1 之间。然而,计算一个输出 AMR 图与一个参考 AMR 图之间部分匹配的程度是非常困难的问题,因为这两个图的节点集之间可能并不具有直接的匹配和映射关系,确定这两个图之间的最优顶点对齐关系实质上是一个 NP 完全问题。

针对 AMR 解析评测广泛采用的是一种称为 Smatch 的度量方法^[6]。它在对两个 AMR 图进行匹配度计算时,首先将每个 AMR 图转化成一个逻辑三元组(triple)的集合,其中每个三元组表示图中的一个顶点或一条边;然后,Smatch 方法计算两个三元组集合之间的匹配或重叠程度,度量指标也分为准确率(precision)、召回率(recall)和 F_1 值。更准确地说,Smatch 方法是通过搜索两个图之间的变量(节点)集的最优匹配而获取的最大的 F_1 值。为给出多项式时间的 Smatch 值计算,一种方法是将计算问题转化为一个整数线性规划 ILP 问题,然后利用现有的 ILP 求解算法求解一个近似解;或者采用启发式爬山算法进行贪心式搜索以获取近似最优解。考虑到评测效率,目前的 Smatch 值计算主要是采用爬山算法进行近似求解^[6]。

2 AMR 解析算法

近年来,对 AMR 解析问题的研究受到国内外研究者的广泛关注,先后提出多种不同的解析模型与算法。依据解析过程与解析策略的不同,将现有的 AMR 解析算法划分成基于图的解析方法、基于转换的解析、基于组合范畴语法的解析方法和基于机器翻译的解析方法等 4 种不同类型,并对这 4 种类型的解析算法分别进行分析与比较。

2.1 基于图的解析方法

Flanigan 等在 2014 年提出了第 1 个 AMR 解析器 JAMR^[7]。JAMR 是一个基于统计的模型,由词串直接得到 AMR 有向图。JAMR 使用管道式的方法分阶段完成 AMR 解析的两个子任务:概念识别和关系识别。

(1)概念识别阶段的任务是将输入句子中的词片段映射到概念图片段。其中,概念图片段大部分情况只含有单个的标注概念节点,但它有时也可能是一个含有多个概念节点和标注边的子图结构。JAMR 将概念识别任务看成是一个序列化标注和切分任务,使用半马尔科夫模型^[8]进行概念识别,并设计了一个动态规划算法进行解码。概念识别阶段的输出结果是概念图片段的序列。

(2)关系识别是在第 1 阶段输出的概念图片段序列中,通过添加带标注的有向边而构造一个 AMR 图。JAMR 将关系识别看作是一个基于约束的组合优化问题,在给定约束集的条件下,搜索具有最大值的子图。子图的分值则采用按边分解的模式和线性模型进行计算,即有

$$\text{score}(E_G; \theta) = \sum_{e \in E_G} \theta^T f(e) \quad (1)$$

式中: E_G 为子图的边集, θ 为模型的参数, $f(e)$ 为用于计算边 e 的特征函数。具体地, JAMR 分两个步骤实现关系识别:首先,设计了一个基于最小生成树 Kruskal 算法而改造的算法 MSCG 生成一个包含所有概念图片段的最大连通子图;然后采用拉格朗日松弛算法^[9]调整该生成图以满足确定性约束,使得从图中每个顶点发出的有向边的关系标注类型(主要针对边的标注 ARG0-ARG5)不重复。

为了训练解析器,需要将训练语料中的句子和标注的 AMR 图进行对齐,即需要知道句子中的词片段对应的 AMR 图的概念片段。为此,JAMR 建立一个自动对齐器来实现对齐。自动对齐器定义了一系列启发式规则,对每个句子通过执行一个贪心算法过程去匹配定义的规则,以实现句子与 AMR 图的对齐,进而构造一个包含所有词片段到概念图片段的映射关系的对齐表。另外,Pourdarnghani 等也尝试采用统计机器翻译中常用的 IBM 词对齐模型来实现句子与 AMR 图之间的对齐,也取得了基本相当的对齐效果^[10]。Werling 等认为在 JAMR 中的关系识别子任务并不是解析过程中最大的挑战^[11],其已经取得了较好的性能,解析过程中主要难点则在于概念识别。因此,他们提出一种关于改进概念识别的新算法 NER++。该算法将由输入句子中的词片段映射到概念图片段的任务看成是依次对句子中的词(或短语)施加不同类型的动作而生成相应概念图片段的过程,并具体设计了 9 个不同的动作:IDENTITY, VERB, VALUE, LEMMA, DICT, NAME, PERSON, DATE 和 NONE。然后,再引入一个最大熵分类器,将各个词片段到概念图片段的映射通过动作分类而实现。相对于 JAMR,该方法获取了一定的解析性能提升;然而该方法的整体框架还是和 JAMR 一样,采用了管道式方法。

为缓解管道式 AMR 解析模型的误差传播,Zhou 等提出了一种增量式联合模型同时执行 AMR 解析的两个子任务^[12],即将给定的句子 x ,直接映射到 AMR 图 y ,该联合解码问题的目标函数可定义为

$$\hat{y} = \underset{y' \in \text{gen}(x)}{\operatorname{argmax}} (\mathbf{w}^T \cdot \varphi(x, y') + \mathbf{w}'^T \cdot f(y')) \quad (2)$$

式中:第 1 项用于计算推导出的概念片段序列的分值,第 2 项则是一个推导出的 AMR 子图的分值。显然,执行式(2)所表示的联合解码问题是一个难解的问题。为此,他们首先提出了一种新算法 CWBS 来实现增量式的关系识别,其基本思想是在追加当前的概念片段到左边的部分图时,仅考察当前概念片段与左边的各个连通分量之间的关系。具体地,设计了一种从右到左的嵌套柱状搜索算法,搜索当前片段与左边各连通分量之间的最优边集,并设计了 SHIFT, LEFTARC, RIGHTARC 和 LEFT & RIGHT-ARCS4 种类型的动作。CWBS 算法不仅能保证最终输出 AMR 图的连通性,还可以有效避免搜索过程中的噪声;然后,再采用一种类似于 multiple-beam 的 segment-based 解码器来实现概念识别,并将关系识别的 CWBS 算法集成到一个联合分析框架中,从而实现两个子任务的增量式联合分析。对模型的参数训练,采用一种改进的参数更新方法 Max-volation perceptron^[13],它在最大错误发生的位置对参数进行更新,这种参数更新算法的收敛速度更快并且精度也有所提高。实验结果显示,该联合模型获得了明显的性能改进。

2.2 基于转换(Transition-Based)的解析方法

以往的研究表明,基于转换的分析方法在依存分析、短语结构分析等句法分析任务中获得了成功的应用^[14,15]。Wang 等提出了一种通过对依存树进行转换的方法 CAMR 来实现 AMR 解析^[16],该算法主要包括两个步骤:(1)使用现有的依存解析器将句子生成相应的依存句法树;(2)采用一种新的转换算法将依存树转换为 AMR 图。解析器建立了一个转换动作的集合^[16],主要通过这个集合来完成依存树向 AMR 图的转换。动作集合的定义对最终的解析结果非常重要,集合中主要包括 8 个动作:SWAP-lr, REATTACHk-lr, REENTRANCEk-lr, REPLACE HEAD, MERGE, NEXT EDGE-lr, NEXT NODE-lc 和 DELETE NODE。例如,图 2 中的 SWAP 展示了节点间依存关系转换的过程。首先将节点“and”删除,然后将节点“and”插入到节点“Korea”之前,并修改节点相对应的边,这样通过 SWAP 这个动作来修改依存树与 AMR 图不同的依存关系。对句子的依存树进行 AMR 解析时,通过执行一个贪心解码算法选择一个最优的动作序列来完成依存树到 AMR 图的转换。基于依存树转换的 AMR 解析避免了其他解析器中 AMR 对齐中产生的错误,该算法的主要错误则出现在依存树转换到 AMR 图的过程中。他们又进一步对该算法进行了改进^[17],主要是定义了一种用于推导隐含概念的新动作 INFER,并在模型中加入了语义角色标注、词义消歧和共指消解等扩展特征,从而获得了更好的解析性能。

Goodman 等认为 CAMR 中采用的贪心解码算法不可避免地存在错误累积和传播问题,因此他们在 CAMR 算法的基础上探索扩展模仿学习^[18]到结构化预测过程以实现 AMR 解析^[19]。具体地,他们

主要提出噪音消减和针对性搜索两种新的手段有效地减小搜索的空间,提高模仿学习的效率和性能,从而改进了 AMR 解析的准确性。

2.3 基于 CCG 的解析方法

Artzi 等提出了一个基于组合范畴语法 (Combinatory categorial grammar, CCG) 的 AMR 解析算法^[20]。CCG 是一种可在句法和语义之间提供透明接口的范畴形式体系^[21], 它能很好地表示出句子的语言特性,已广泛用于传统的语义解析任务^[22,23]。该算法的基本思想是采用 CCG 的句法分析过程生成组合性的 AMR 结构,再采用因子图对非组合性语言现象(如指代关系等)建模求解。算法的处理过程分成两个阶段:

(1)通过 CCG 构造基于 lambda 演算表示形式的组合性 AMR 结构,即通过对句子执行 CCG 分析得到 CCG 分析树,得到未指定的逻辑形式,其中包含有未确定的变量。在 CCG 分析的推导过程中,仍采用传统的 CKY 分析算法去挑选最优的未指定的逻辑形式。对推导的概率值计算则采用对数线性模型。若定义 $D(z)$ 为包含最后逻辑形式 z 的推导子集, $\theta \in \mathbf{R}^l$ 是一个 l 维的参数向量,则可定义 z 的概率值为

$$p(z | x; \theta) = \sum_{d \in D(z)} p(d | x; \theta) \quad (3)$$

而对于一个推导 d ,其概率计算公式为

$$p(d | x; \theta) = \frac{e^{\theta \cdot \varphi(x, d)}}{\sum_{d' \in \text{GEN}(x)} e^{\theta \cdot \varphi(x, d')}} \quad (4)$$

(2)采用因子图及对数线性模型将逻辑形式中的未确定的变量从候选集中搜索最优的变量赋值,从而得到完整的逻辑形式,最后再通过自定义的一套简单可逆的转换过程,将逻辑形式转换成标准 AMR 形式。这个基于 CCG 的两阶段 AMR 解析模型的学习。除了需要进行模型参数估计之外,面临的一大挑战是如何从训练数据中有效地归纳 CCG 词典的词汇项。显然,CCG 词汇生成的质量对 CCG 分析的准确性有着重要的影响。而 CCG 词典词汇项的组合搜索空间巨大,是一个异常困难的问题^[24]。因而如何有效解决 CCG 词汇生成问题成为制约基于 CCG 的 AMR 解析模型提高性能的主要瓶颈。

2.4 基于机器翻译框架的解析方法

Pust 等将 AMR 解析看作是句法的机器翻译 (Syntax-based machine translation, SBMT) 问题^[25],也就是充分利用现有的串到树 SBMT 框架,将句子转换为相应的 AMR 图。为此,首先需要将 AMR 图转换为一种适用于 SBMT 机制的树结构。为实现从 AMR 图到树结构的转换,他们提出了一种多步骤的转换过程,包括修改 AMR 图为句法风格树结构、树重构、树重标注和重排序等多个步骤。由于图与树在结构上具有本质上的差异性,因此在第 1 个转换步骤中将会不可避免地丢弃一些信息。例如,对于图中具有多个父节点的节点,在转换后的树结构中仅能保留单个父节点。显然,这种从 AMR 图到树结构的转换是一种有损转换。进一步,他们在 SBMT 框架下引入了一种建立在 AMR 结构上的语言模型。特别地,他们将多种外部的语义资源加入到模型中,包含层次的语义范畴、词汇形态规范化和基于富含语义信息规则的对齐等,从而使系统获得了较好的解析性能。

2.5 各种解析算法的性能比较

目前,各种 AMR 解析器的实验主要是在几个不同版本的公开数据集上进行,包括 LDC2013E117, LDC2014T12 和 LDC2014T12-proxy,其中数据集 LDC2014T12-proxy 是仅包含了数据集 LDC2014T12 中的新闻语料部分。实验结果的评价都采用了目前公认的 Smatch 值,分为准确率 (P)、召回率 (R) 和 F_1 值。为表示方便,采用一些缩写名称表示前述的各种 AMR 解析器。其中, JAMR 表示文献[7]的解析系统,NER++表示文献[11]的解析系统,JOINT 表示文献[12]的解析系统,CAMR 表示文献[17]

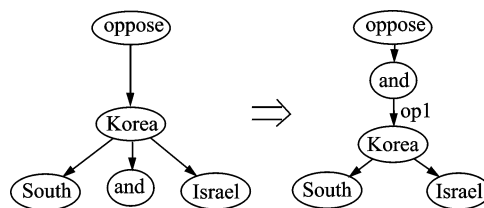


图2 SWAP 动作示例

Fig.2 Example of action SWAP

的解析系统,IMITATION 表示文献[19]的解析,CCG 表示文献[20]的解析系统,SBMT 表示文献[25]的解析系统。在以上的各个 AMR 解析器中,系统 CAMR 和 SBMT 在训练模型时除了使用训练数据外,还额外地使用了较多的外部语义资源。表 2 中给出了各个 AMR 解析系统在 3 个不同的数据集上的实验结果。

表 2 各种 AMR 解析系统的实验性能比较

Tab.2 Comparison of AMR parsers

| Dataset | System | P | R | F_1 |
|------------------|-------------|------|------|-------|
| LDC2013E117 | JAMR(fixed) | 0.67 | 0.58 | 0.62 |
| | NER++ | 0.66 | 0.59 | 0.62 |
| | CAMR | 0.71 | 0.69 | 0.70 |
| | JOINT | 0.73 | 0.69 | 0.71 |
| LDC2014T12-proxy | JAMR(fixed) | 0.68 | 0.59 | 0.63 |
| | NER++ | 0.67 | 0.58 | 0.62 |
| | CCG | 0.67 | 0.66 | 0.66 |
| | CAMR | 0.72 | 0.67 | 0.70 |
| | IMITATION | 0.73 | 0.68 | 0.70 |
| | JOINT | 0.74 | 0.68 | 0.71 |
| LDC2014T12 | JAMR(fixed) | 0.64 | 0.53 | 0.58 |
| | CAMR | 0.70 | 0.62 | 0.66 |
| | JOINT | 0.70 | 0.62 | 0.66 |
| | SBMT | | | 0.67 |

通过比较表 2 中各个 AMR 解析器的实验数据可以看出:(1)目前英文 AMR 的自动解析效果普遍不高,最好成绩的 F_1 值仅为 0.71。(2)AMR 的自动解析需要一定的外部资源。在 LDC2013E117 和 LDC2014T12-proxy 两个数据集上,系统 JOINT 均获得最高的 F_1 值;在数据集 LDC2014T12 上,系统 SBMT 产生了最高的 F_1 值,这也说明在 AMR 解析模型中合理地使用一些外部语义资源能有助于解析性能的提高。因此,AMR 的自动解析技术还有很大的提升空间,在算法和资源建设方面也需要不断研究。此外,影响 AMR 解析效果还有一个重要因素。在自动解析之前,AMR 需要进行概念和词语的自动对齐,而目前对齐的 F_1 值仅有 0.9 左右^[7,10],如果能提升自动对齐的精度,则可以较大幅度提升解析效果。

3 AMR 的应用

随着 AMR 标注规范的提出和 AMR 解析算法研究的快速发展,有研究者开始对如何利用 AMR 结构帮助解决各种自然语言处理领域的应用问题进行探索性研究。目前已有文献显示,在文本摘要、实体链接、生物信息抽取和标题生成等典型应用问题上^[26-32],利用 AMR 结构确实能够有效或显著地提高效果和性能。以下将以文本摘要和实体链接这两种典型应用为例,介绍 AMR 的应用方法与效果。

3.1 利用 AMR 实现文本抽象性摘要

不同于传统的压缩式或抽取式摘要生成技术^[33-35],抽象性摘要是指一种对给定的内容要求从头生成相应的文本摘要的技术。Liu 等利用 AMR 实现了一个抽象性摘要生成系统^[26],该系统包含 3 个基本步骤:(1)使用现有的 AMR 解析器将源文本中的每个句子分别解析成对应的 AMR 图;(2)将第 1 步中得到的多个 AMR 图经过合并和转化操作生成单个摘要 AMR 图;(3)从单个的摘要 AMR 图生成摘要文本。图 3 给出了一个简单的摘要生成过程实例。其中,第(2)步是一个关键和困难的步骤,第(2)步的

实现又划分成两个阶段:第一阶段对第(1)步得到的多个句子 AMR 图通过执行概念融合、句子连接以及图扩展等步骤,生成单个连通的 AMR 图(称之为源图);第二阶段再从第一阶段得到的源图中挑选部分节点和边构造一个子图(称之为摘要图),而从源图中选择生成摘要子图的任务可以看成是一个结构化预测问题,其目标是在不改变内容主要语义信息的条件下输出最简洁和紧凑的表示结构。具体的,通过引入合适的特征和约束表示,采用整数线性规划求解该结构化预测任务,生成一个摘要子图。采用 ROUGE-1 得分^[36]对该抽象性摘要系统的摘要结果进行评测,最后在测试集上利用标准 AMR 标注和 JAMR 解析结果生成的 Oracle 摘要的 F 值分别达到了 65.8% 和 57.8%,这个初步的探索性研究的实验结果显示,利用 AMR 结构生成抽象性文本摘要的方法能够产生良好的结果。

3.2 基于 AMR 的无监督实体链接实现

实体链接是指将在源文本中提及的实体与知识库中的相关实体关联到一起^[37-39]。Pan 等利用 AMR 实现了一个无监督的实体链接系统^[27],该系统利用 AMR 结构中富含的深度语义信息提供描述各个实体指称出现的上下文特征的丰富表示,从而有利于协作者实体指称的最优集合的选择,进而有助于实现实体指称的消歧;在此基础上再采用无监督的图推导方法进行上下文比较,从而实现实体链接的目标。基于 AMR 的实体消歧方法可分解为 3 个基本步骤:(1)将实体指称出现的源文本解析成 AMR,然后对给定的实体指称 m 采用基于规则的方法构造一个以 m 为中心的星形图状知识网络;(2)通过规则方法从知识库构造各个实体的知识网络,作为实体的上下文表示形式;(3)对于每个待链接的实体指称,利用实体显著度计算生成一个候选实体的列表,再采用基于上下文协作者的信息,对当前实体指称与各个候选实体之间根据之前构造的知识网络图进行 Jaccard 相似度计算,以重排序的方式实现实体链接。采用这种无监督的实体链接算法在增加人工实体链接标注的 AMR 数据集 LDC2013E117 上进行了实验,如果基于人工标注的 AMR 结构,在测试集上获得了 91.13% 的准确率,略高于目前有监督的实体链接方法达到的最高准确率 91.01%,该有监督的实体链接方法是通过规则方法组合了两个不同的系统^[40,41],结合了两者的优势,并在大规模的标注数据集 TAC-KBP2009-2014 上训练重排序模型获取的;如果基于 AMR 解析器的输出,该算法也获得了 88.52% 的准确率。实验结果表明,相对于实体链接的其他方法,AMR 结构确实能够更好地捕捉和表示实体指称的上下文信息。

4 中文 AMR 的开发

与传统的基于树的句法语义表示方法不同,AMR 使用单根有向无环图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。因此,增加更多语种的 AMR 语料和多语言自动解析技术就成为目前该领域最为迫切的研究内容。捷克语、汉语也开始标注了 AMR 小规模语料库^[42],微软依托自建的句子逻辑语义语料库,使用与 AMR 映射的规则方法,开发了英语、法语、德语、西班牙语和日语的 AMR 解析工具^[43]。不过,受制于中文 AMR 语料的缺乏,针对中文的 AMR 自动解析方面的研究尚未

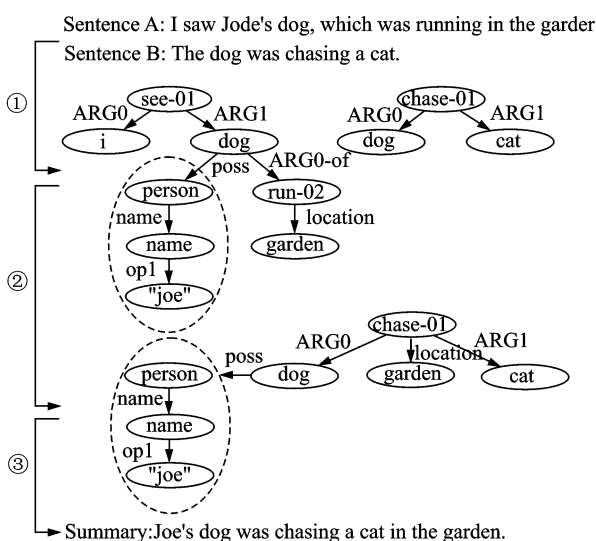


图3 一个简单的英文抽象性摘要生成示例

Fig.3 An example of a simple abstract generation

展开。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。南京师范大学 Li 等近来已经开始启动中文 AMR 的规范设计和语料库标注工作^[44]。Li 等(2016)基于 AMR 的框架结构,将 AMR 语义表示体系引入到汉语中,同时也充分考虑汉语与英语的表达差异性,重点解决了 AMR 概念和词语对齐的问题,初步建立了一套汉语抽象语义的表示方法和标注规范法^[44]。他们首先根据中文宾州树库(Chinese Penn treebank, CTB)的分词规范,对 1562 句中文《小王子》语料进行了自动分词和人工校对。然后,根据中文 Propbank 的动词框架词典^[45],人工标注了每个句子的中文 AMR,从而获得了一个与英文《小王子》句对齐的双语 AMR 语料库,参见 <http://www.cs.brandeis.edu/~clp/camr/camr.html>。在此基础上,李斌等(2017)进一步比较了两种语言中图结构的对应情况、图结构存在的比例与图结构的共性和差异等,认为 AMR 的跨语言表示能力和 AMR 对于汉语的句法语义表示能力都较为理想^[46]。目前,他们正在进行对 CTB 的 AMR 标注工作,以建立较大规模中文 AMR 标注语料库,将为汉语 AMR 的自动解析与应用提供良好的研究基础。

5 结束语

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,使其至少具有两点潜在价值:(1)提升文本摘要、事件分析和智能问答等应用技术;(2)为句子级别之上的篇章语义表示奠定研究基础。因而自 AMR 提出之后,就引起了国际上的高度重视,涌现了从自动解析技术到转化应用等多方面的研究论文。纵观 AMR 的研究现状,可以发现,一方面由于 AMR 图结构能够富含句子中完整的语义信息,因而能有利地促进各种自然语言应用问题的解决;另一方面,AMR 解析器的性能是 AMR 结构能否有效应用的关键。然而现有的 AMR 解析算法研究结果显示,当前 AMR 解析算法的准确性还比较有限,最高 F_1 值仅在 70% 左右,相对于句法分析算法当前所能达到的准确性,还存在不小的差距。因此认为今后在以下方面值得进一步展开更多、更深入的 AMR 相关研究。(1)概念对齐与生成是 AMR 解析的一个重要子任务,其性能的高低将是提高 AMR 解析准确性的重要瓶颈,而目前这方面的研究相对较少。由于概念对齐与生成是 AMR 图构造的一个前提任务,如何充分利用各种外部语义资源、设计有效的算法对输入句子中的各个词或短语生成有效的概念候选,特别是一些隐含或缺省概念的添加,是一个困难和重要的研究问题。(2)作为一种句子语义表示形式,AMR 结构与句子的句法结构(特别是依存句法结构)之间存在很强的关联性。如果能将句子的 AMR 解析与句法分析两者构建成一个联合分析过程,使得句法分析与语义分析两个任务之间能够实现信息交互,相互利用以促进彼此的结构消歧,将会大大有益于句法分析与 AMR 语义解析两者的性能提升。(3)近年来神经网络和深度学习已经被广泛应用于自然语言处理的多个研究问题,利用神经网络学习生成词、短语和句子等分布式表示,在句子切分、依存句法分析和成分句法分析等任务上的应用均获得了明显的收益^[47-50]。然而,将神经网络模型应用于 AMR 解析任务还是一个有待研究的问题。不同于传统的句法分析任务,AMR 结构描述的是概念节点之间的语义关系,因此,如何学习生成有效的概念分布式表示是一个首先需要解决的问题。(4)进一步更广泛地开展 AMR 在多个自然语言处理应用领域中的应用研究非常有意义。AMR 作为一种句子级抽象语义表示形式,其主要设计目标在于描述和揭示句子中蕴含的完整的、深度的语义信息,以利于和推动完美解决各种自然语言处理应用问题。目前的一些 AMR 应用研究已经初步显示了 AMR 语义结构的应用价值,然而还有必要针对更多的自然语言处理的应用问题展开探索性的 AMR 应用研究,例如信息抽取、问答系统、文本匹配和机器阅读理解等。

参考文献:

- [1] 孙茂松,刘挺,姬东鸿. 语言计算的重要国际前沿[J]. 中文信息学报,2014, 28(1):1-8.
Sun Maosong, Liu Ting, Ji Donghong. An important international forefront of language computing [J]. Journal of Chinese

- Information Processing, 2014; 28(1): 1-8.
- [2] Banarescu L, Bonial C, Cai S, et al. Abstract meaning representation for sembanking[C] // Proceedings of the 7th Linguistic Annotation Workshop. Sophia, Bulgaria; The Association for Computational Linguistics, 2013; 178-186.
 - [3] Zettlemoyer L S, Collins M. Learning to map sentences to logical form; Structured classification with probabilistic categorial grammars[C] // Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence (UAI), Edinburgh, Scotland; AUAI Press, 2005; 658-666.
 - [4] Wong Y, Mooney R J. Learning for semantic parsing with statistical machine translation[C] // Proceedings of the North American Chapter of the ACL (NAACL 2006). New York, USA; The Association for Computational Linguistics, 2006; 439-446.
 - [5] Kingsbury P, Palmer M. From treebank to propbank[C] // Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas, Canary Islands, Spain; European Language Resources Association, 2002; 1989-1993.
 - [6] Cai S, Knight K. Smatch: An evaluation metric for semantic feature structures[C] // Proceedings of the Association for Computational Linguistics (ACL 2013). Sofia, Bulgaria; The Association for Computational Linguistics, 2013; 748-752.
 - [7] Flanigan J, Thomson S, Carbonell J, et al. A discriminative graph-based parser for the abstract meaning representation[C] // Proceedings of the Association for Computational Linguistics (ACL 2014). Baltimore, MD, USA; The Association for Computer Linguistic, 2014; 1426-1436.
 - [8] Janssen J, Limnios N. Semi-Markov models and applications[M]. [S.l.]: Springer, 1999.
 - [9] Rush A M, Collins M. A tutorial on dual decomposition and Lagrangian relaxation for inference in natural language processing[J]. Journal of Artificial Intelligence Research, 2012, 45(1): 305-362.
 - [10] Pourdamghani N, Yang G, Hermjakob U, et al. Aligning English strings with abstract meaning representation graphs[C] // Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014). Doha, Qatar; The Association for Computational Linguistics, 2014; 425-429.
 - [11] Werling K, Angeli G, Manning C. Robust subgraph generation improves abstract meaning representation parsing[C] // Proceedings of the Association for Computational Linguistics (ACL 2015). Beijing, China; The Association for Computational Linguistics, 2015; 982 - 991.
 - [12] Zhou J, Xu F, Uszkoreit H, et al. AMR parsing with an incremental joint model[C] // Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, Texas, USA; The Association for Computational Linguistics, 2016; 178-186.
 - [13] Huang L, Fayong S, Guo Y. Structured perceptron with inexact search[C] // Proceedings of the Human Language Technologies; Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2012). Montreal, Canada; The Association for Computational Linguistics, 2012; 142-151.
 - [14] Nivre J. Algorithms for deterministic incremental dependency parsing[J]. Computational Linguistics, 2008, 34(4): 513-553.
 - [15] Zhang Y, Clark S. Syntactic processing using the generalized perceptron and beam search[J]. Computational Linguistics, 2011, 37(1): 105-151.
 - [16] Wang C, Xue N W, Pradhan S. A transition-based algorithm for AMR parsing[C] // Proceedings of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies (NAACL 2015). Denver, Colorado, USA; The Association for Computational Linguistics, 2015; 366-375.
 - [17] Wang C, Xue N, Pradhan S. Boosting transition-based AMR parsing with refined actions and auxiliary analyzers[C] // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL 2015). Beijing, China; The Association for Computational Linguistics, 2015; 857-862.
 - [18] Silver D, Bagnell J, Stentz A. High performance outdoor navigation from overhead data using imitation learning[C] // Robotics; Science and Systems IV. Zurich, Switzerland; The MIT Press, 2008; 262-269.
 - [19] Goodman J, Vlachos A. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing[C] // Proceedings of Association for Computational Linguistics (ACL 2016). Berlin, Germany; The Association for Computational Linguistics, 2016; 1-11.
 - [20] Artzi Y, Lee K, Zettlemoyer L. Broad-coverage CCG semantic parsing with AMR[C] // Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal; The Association for Computational Linguistics, 2015; 1699-1710.
 - [21] Steedman M. The syntactic process[M]. USA; The MIT Press, 2000.

- [22] Zettlemoyer L S, Collins M. Online learning of relaxed CCG grammars for parsing to logical form[C] // Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2007). Prague, Czech Republic; The Association for Computational Linguistics, 2007; 678-687.
- [23] Kwiatkowski T, Zettlemoyer L, Goldwater S, et al. Inducing probabilistic CCG grammars from logical form with higher-order unification[C] // Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2010). MIT Stata Center, Massachusetts, USA; The Association for Computational Linguistics, 2010; 1223-1233.
- [24] Kwiatkowski T, Zettlemoyer L, Goldwater S, et al. Lexical generalization in CCG grammar induction for semantic parsing [C] // Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2011). Edinburgh, UK; The Association for Computational Linguistics, 2011; 1512-1523.
- [25] Pust M, Hermjakob U, Knight K, et al. Using syntax-based machine translation to parse English into abstract meaning representation[C] // Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal; The Association for Computational Linguistics, 2015; 482-489.
- [26] Liu F, Flanigan J, Thomson S, et al. Toward abstractive summarization using semantic representations[C] // The 2015 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies (NAACL 2015). Denver, Colorado, USA; The Association for Computational Linguistics, 2015; 1077-1086.
- [27] Pan X, Cassidy T, Hermjakob U, et al. Unsupervised entity linking with abstract meaning representation[C] // The 2015 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies (NAACL 2015). Denver, Colorado, USA; The Association for Computational Linguistics, 2015; 1130-1139.
- [28] Takase S, Suzuki J, Okazaki N, et al. Neural headline generation on abstract meaning representation[C]// Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, Texas, USA; The Association for Computer Linguistic, 2016; 1054-1059.
- [29] Sachan M, Xing E. Machine comprehension using rich semantic representations[C] // Proceedings of Association for Computational Linguistics (ACL 2016). Berlin, Germany; The Association for Computational Linguistics, 2016; 486-492.
- [30] Garg S, Galstyan A, Hermjakob U, et al. Extracting biomolecular interactions using semantic parsing of biomedical text [C]// Proceedings of Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016). Phoenix, Arizona, USA; AAAI Press, 2016; 2718-2726.
- [31] Pourdamghani N, Knight K, Hermjakob U. Generating English from abstract meaning representations[C] // Proceedings of the Ninth International Natural Language Generation Conference. Edinburgh, UK; The Association for Computational Linguistics, 2016; 21-25.
- [32] Mitra A, Baral C. Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning[C] // Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016). Phoenix, Arizona, USA; AAAI Press, 2016; 2779-2785.
- [33] Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization[C] // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Lisbon, Portugal; The Association for Computational Linguistics, 2015; 379-389.
- [34] Nallapati R, Zhou B, Santos C, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[C] // Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany; The Association for Computational Linguistics, 2016; 280-290.
- [35] Cheng J, Lapata M. Neural summarization by extracting sentences and words[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany; The Association for Computational Linguistics, 2016; 484-494.
- [36] Lin C. ROUGE: A package for automatic evaluation of summaries[C] // Proceedings of ACL Workshop on Text Summarization Branches Out. Barcelona, Spain; The Association for Computational Linguistics, 2004; 25-26.
- [37] Hachey B, Radford W, Nothman J, et al. Evaluating entity linking with Wikipedia[J]. Artificial Intelligence, 2013, 194(3): 130-150.
- [38] Sil A, Yates A. Re-ranking for joint named-entity recognition and linking[C] // Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013). San Francisco, CA, USA; ACM, 2013; 2369-2374.
- [39] Wick M, Singh S, Pandya H, et al. A joint model for discovering and linking entities[C] // Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (AKBC@CIKM 2013). San Francisco, California, USA; ACM, 2013; 67-72.
- [40] Chen Z, Ji H. Collaborative ranking: A case study on entity linking[C] // Proceedings of Empirical Methods in Natural Lan-

- guage Processing (EMNLP 2011). Edinburgh, UK: The Association for Computational Linguistics, 2011: 771-781.
- [41] Cheng X, Roth D. Relational inference for wikification[C] // Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2013). Seattle, Washington, USA: The Association for Computational Linguistics, 2013: 1787-1796.
- [42] Xue N, Bojar O, Hajic J, 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. Reykjavik, Iceland: European Language Resources Association ELRA, 2014: 1765-1772.
- [43] Vanderwende L, Menezes A, Quirk C. An AMR parser for English, French, German, Spanish and Japanese and a new AMR-annotated corpus[C] // The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2015). Denver, Colorado, USA: The Association for Computational Linguistics, 2015: 26-30.
- [44] Li B, Wen Y, Qu W, et al. Annotating the little prince with Chinese AMRs[C] // Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016. Berlin, Germany: The Association for Computational Linguistics, 2016: 7-15.
- [45] Xue N, Palmer M. Adding semantic roles to the Chinese Treebank[J]. *Natural Language Engineering*, 2009, 15(1): 143-172.
- [46] 李斌, 闻媛, 卜丽君, 等. 英汉《小王子》AMR 语义图结构的对比分析[J]. *中文信息学报*, 2017(1): 33-41.
Li Bin, Wen Yuan, Pu Lijun, et al. A contrastive analysis of AMR semantic graph for the Little Prince in English and Chinese[J]. *Journal of Chinese Information Processing*, 2017(1): 33-41.
- [47] Cai D, Zhao H. Neural word segmentation learning for Chinese[C] // Proceedings of Association for Computational Linguistics (ACL 2016). Berlin, Germany: The Association for Computational Linguistics, 2016: 409-420.
- [48] Zhou H, Zhang Y, Huang S, et al. A neural probabilistic structured-prediction model for transition-based dependency parsing [C] // Proceedings of Association for Computational Linguistics (ACL 2015). Beijing, China: The Association for Computational Linguistics, 2015: 1213-1222.
- [49] Socher R, Bauer J, Manning C D, et al. Parsing with compositional vector grammar[C] // Proceedings of Association for Computational Linguistics (ACL 2013). Sofia, Bulgaria: The Association for Computational Linguistics, 2013: 455-465.
- [50] Watanabe T, Sumita E. Transition-based neural constituent parsing[C] // Proceedings of Association for Computational Linguistics (ACL 2015). Beijing, China: The Association for Computational Linguistics, 2015: 1169-1179.

作者简介:



曲维光 (1964-), 男, 教授, 博士生导师, 研究方向: 自然语言处理, E-mail: wgqu@nynu.edu.cn.



吴晓东 (1989-), 男, 硕士研究生, 研究方向: 自然语言处理, E-mail: 1041871851@qq.com.



顾敏 (1993-), 女, 硕士研究生, 研究方向: 自然语言处理, E-mail: 178408229@qq.com.



周俊生 (1972-), 男, 教授, 硕士生导师, 研究方向: 自然语言处理, E-mail: zhoujs@nynu.edu.cn.



戴茹冰 (1989-), 女, 博士研究生, 研究方向: 计算语言学, 词汇语义学, E-mail: ice_dr@163.com.



顾彦慧 (1978-), 男, 副教授, 硕士生导师, 研究方向: 自然语言处理, E-mail: gu@nynu.edu.cn.