

文章编号: 1003-0077(2022)01-0021-10

## 基于概念关系对齐的中文抽象语义表示解析评测方法

肖力铭<sup>1</sup>, 李 斌<sup>1</sup>, 许智星<sup>1</sup>, 霍凯蕊<sup>1</sup>, 冯敏萱<sup>1</sup>, 周俊生<sup>2</sup>, 曲维光<sup>2</sup>

(1. 南京师范大学 文学院, 江苏 南京 210097; 2. 南京师范大学 计算机与电子信息学院, 江苏 南京 210023)

**摘 要:** 抽象语义表示 (Abstract Meaning Representation, AMR) 是一种句子语义表示方法, 能够将句子的语义表示为一个单根有向无环图。随着中文 AMR 语料库规模的扩大, 解析系统的研究也相继展开, 将句子自动解析为中文 AMR。然而, 现有的 AMR 解析评测方法并不能处理中文 AMR 的重要组成部分——概念对齐和关系对齐信息, 尤其是关系对齐中对应到有向弧上的虚词信息。因此, 为了弥补中文 AMR 解析评测在这两个方面的空缺, 该文在 Smatch 指标的基础上加入了描写概念对齐和关系对齐的三元组, 得到用以评测中文 AMR 的整体性指标 Align-Smatch。选取 100 句人工标注语料与标准语料进行评测对照实验, 结果显示, Align-Smatch 有效兼容了对齐信息, 对有向弧的评测比 Smatch 更合理。该文还提出了概念对齐指标、关系对齐指标、隐含概念指标共三个分项指标, 以进一步评测中文 AMR 解析器在对齐子任务中的分项性能。

**关键词:** 抽象语义表示; 评测方法; 概念对齐; 关系对齐; 语义分析

**中图分类号:** TP391

**文献标识码:** A

### A Novel Evaluation Method for Chinese Abstract Meaning Representation Parsing Based on Alignment of Concept and Relation

XIAO Liming<sup>1</sup>, LI Bin<sup>1</sup>, XU Zhixing<sup>1</sup>, HUO Kairui<sup>1</sup>,  
FENG Minxuan<sup>1</sup>, ZHOU Junsheng<sup>2</sup>, QU Weiguang<sup>2</sup>

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;  
2. School of Computer, Electronics and Information, Nanjing Normal University, Nanjing, Jiangsu 210023, China)

**Abstract:** Abstract Meaning Representation is a sentence-level meaning representation, which abstracts a sentence's meaning into a rooted acyclic directed graph. With the continuous expansion of Chinese AMR corpus, more and more scholars have developed parsing systems to automatically parse sentences into Chinese AMR. To make up for the vacancy of Chinese AMR parsing evaluation methods, we have improved the Smatch algorithm of generating triples to make it compatible with concept alignment and relation alignment. We finally complete a new integrity metric Align-Smatch for parsing evaluation. Compared on 100 manually annotated AMR and gold AMR, it is revealed that Align-Smatch works well in alignments and more robust in evaluating arcs. We also put forward some fine-grained metric for evaluating concept alignment, relation alignment and implicit concepts, in order to further measure parsers' performance in subtasks.

**Keywords:** abstract meaning representation; evaluation method; concept alignment; relation alignment; semantic parsing

## 0 引言

随着词法、句法分析技术的日益成熟, 自然语言处理已经整体推进到了语义分析层面。句子语义作为重点和难点, 更占据了语义分析的核心地位<sup>[1]</sup>。

针对整句语义形式化表示的缺失, 以及句子语义标注存在的领域相关性的问题, Banarescu 等人<sup>[2]</sup>于 2013 年提出了一种与领域无关的整句语义表示方法——抽象语义表示 (Abstract Meaning Representation, AMR)。AMR 能将一个句子的语义抽象为一个单根有向无环图, 其不仅可以描写一

收稿日期: 2021-03-30 定稿日期: 2021-05-23

基金项目: 国家社会科学基金(18BYY127); 国家自然科学基金(61772278); 江苏省社会科学基金(20JYB004)

个名词由多个谓词支配所形成的论元共享现象,还允许补充隐含语义以完整地表示句义。强大的语义表示能力使得 AMR 一经问世便引起了广泛的关注,涌现出大量关于 AMR 自动解析技术、AMR 转换应用等方面的文章。图 1 以“I have been having some trouble with money.”为例给出了 AMR 的两种展现形式:文本缩进法(上)和图示法(下)。

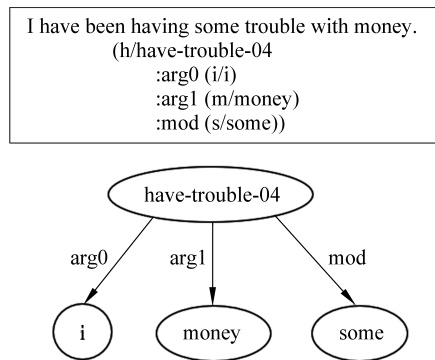


图 1 AMR 示例

不过,AMR 在体系上仍存在两大问题。首先,AMR 只关注实词,忽视虚词,对整句的语义表示仍不够完整<sup>[2-3]</sup>,如图 1 句子中的“with”未得到表示。其次,AMR 缺乏实词和概念的对齐(alignment),导致在相关模型的训练中需要使用对齐器(aligner)进行单独的对齐步骤。但对齐器自身存在误差,比如目前精度最高的 JAMR-Aligner<sup>[4]</sup>在遇到同一个词根的不同形态同时出现时,容易做出错误的判断<sup>[5]</sup>。如图 1 句子中同时出现了 have 和 having,对齐器可能错误地将 have-trouble-04 与 have 对齐。对齐器的误差将降低模型约 5%~10%的分析精度<sup>[6-7]</sup>。

2016 年, Li 等人<sup>[8]</sup>将 AMR 引入到汉语中,进行了三大调整和改良。首先,在体系上增加了对于虚词的表示,如增加表示“体”“个体量词”等语义关系的标签。其次,在表示方法上提出了概念、关系与句子中的词语对齐,特别是把虚词标注在节点或有向弧上<sup>[3,9]</sup>。最后,根据汉语的语言特点规定了具体的标注细则,如规定了重叠式、离合词、连动结构等特殊结构的处理方法。最终,设计了一套中文 AMR 的标注规范,构建了一个规模约为两万句的中文 AMR 语料库<sup>[8-9]</sup>。

随着语料库规模的扩大,越来越多的学者参与到中文 AMR 的自动解析工作中<sup>[10-14]</sup>。他们研制的解析系统能在给定句子的情况下,预测并输出该句相应的 AMR 结构。在 2020 年国际自然语言学习会议 CoNLL 发布的跨语言语义表示方法解析评测

任务 MRP2020 中,中文 AMR 的自动分析达到了 0.81 的  $F_1$  值<sup>[15-16]</sup>,是目前最好的结果。

尽管如此,这些成果仍不能体现中文 AMR 解析的真实水平。包括 CoNLL 在内,所有实验和评测所使用的评测指标都是基于英文 AMR 设计的,并不适合中文 AMR。Smatch<sup>[17]</sup>、SemBleu<sup>[18]</sup>等整体性指标无法兼容中文 AMR 所做出的调整,评测语料只能删除对齐信息。中文 AMR 的概念和关系对齐信息,特别是标注在有向弧上的虚词并没有得到解析和评估。CoNLL 使用的评测工具 MTool<sup>[15]</sup>能够评测概念对齐信息,但也无法衡量关系对齐信息。其次,一些 AMR 分项评测指标<sup>[19-20]</sup>主要围绕概念识别和关系识别子任务设计,也不涉及对齐信息。

为了弥补中文 AMR 解析评测在对齐信息上的空缺,为未来中文 AMR 解析工作的发展提供新的标准和方向,本文基于 Smatch 指标进行改进,提出了 Align-Smatch 指标。同时,为了给评测提供细粒度的可读性报告,本文围绕对齐信息,提出了概念对齐指标、关系对齐指标、隐含概念指标共三个分项评测指标,以更好地服务于中文 AMR 解析评测。

## 1 概念对齐和关系对齐

AMR 的本质是把句子中的词以及词之间的联系抽象为“概念”和“关系”,反映在 AMR 图中则分别是“节点”和“有向弧”。具体来说,词被抽象为概念节点,词之间的关系被抽象为带有语义角色标签的有向弧。

这种抽象的语义表示方式,一方面使得 AMR 拥有较强的语义表示能力,能够增加、删除、修改概念节点,对隐含概念进行补充标注。如图 2 中,AMR 在标注专有名词“中国”时,将其中的隐含概念“country”抽象为节点补出。

但另一方面,也使得 AMR 难以反映概念和词的映射关系,因为它不进行概念对齐。AMR 是基于英文设计的,考虑到英文词语有形态变化,而概念没有,所以在 AMR 标注过程中,难以使用词语编号来代表概念,只能取词语的首字母作为概念节点的编号,或按照节点的出现顺序分配编号<sup>[9]</sup>,如表 1 左侧所示。这导致计算机不能直接对概念溯源,也无法从 AMR 中还原句子的语序,加大了 AMR 解析等应用技术的难度。

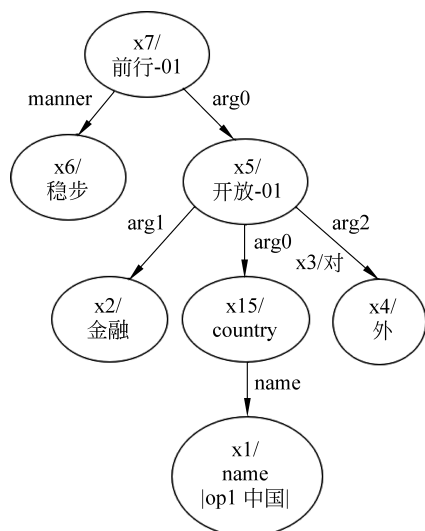


图2 “中国金融对外开放稳步前行”中文 AMR 图示

表1 概念对齐标注

中国	金融	对	外	开放	稳步	前行
x1	x2	x3	x4	x5	x6	x7
概念与词不对齐的标注			概念与词对齐的标注			
(x0 / 前行-01 :manner() (x1 / 稳步) :arg0() (x2 / 开放-01 :arg1() (x3 / 金融) :arg2() (x4 / 外) :arg0() (c / country :name() (x5 / name :op1 中国))))			(x7 / 前行-01 :manner() (x6 / 稳步) :arg0() (x5 / 开放-01 :arg1() (x2 / 金融) :arg2() (x4 / 外) :arg0() (x15 / country :name() (x1 / name :op1 中国))))			

为了解决这一问题,李斌等人提出了融合概念对齐的中文 AMR 标注方案,实现了中文 AMR 的概念对齐<sup>[9]</sup>。通过对分词后的原始句子按照线性有序的原则,给每个词分配编号,每个概念节点也被赋予了相应的编号。编号采用“x+数字”的形式表示,如果数字不大于句子的总词数,则代表与该节点对齐的词的编号。如表1所示,加粗部分即为对齐后的编号表示,保证了大部分概念节点与原始句子中的词对齐了。没有对齐的概念则是被补出的隐含概念,如“x15”节点的概念 country,它被赋予了大于句子总词数的编号。

除了概念对齐,AMR 还直接忽略了如介词、冠词等虚词,认为这些词对句子的语义没有贡献。但在解析中,一些研究尝试将虚词利用起来,提高模型关系预测的精度。如对齐器 ISI<sup>[21]</sup>将介词与语义角色关系进行对齐,Brandt 等<sup>[22]</sup>在解析模型中加入介词语义角色标注信息。虽然这些尝试尚未取得显著成果,但虚词在 AMR 解析中的作用仍有研究空间。

其次,虚词对连接上下文有很大的作用,有利于更为精准地将 AMR 还原为句子、语篇。

因此,中文 AMR 选择保留虚词进行标注,将表示句子的体意义和语气意义的虚词处理为概念节点;将表示实词间关系意义的虚词视为语义关系的映射,与语义角色标签一同标注在有向弧上<sup>[3]</sup>。有向弧上的虚词同样带有编号,这完成了语义关系与词的对齐,实现了关系对齐。如表2所示,加粗部分即是把原始句子中表示实词“开放”和“外”之间关系的虚词“对”放进了 AMR 中,使用编号 x3 将表示对象的语义关系 arg2 与虚词“对”进行对齐。

表2 关系对齐标注

关系与词不对齐的标注	关系与词对齐的标注
(x7 / 前行-01 :manner() (x6 / 稳步) :arg0() (x5 / 开放-01 :arg1() (x2 / 金融) :arg2() (x4 / 外) :arg0() (x15 / country :name() (x1 / name :op1 中国))))	(x7 / 前行-01 :manner() (x6 / 稳步) :arg0() (x5 / 开放-01 :arg1() (x2 / 金融) : <b>arg2(x3 / 对)</b> (x4 / 外) :arg0() (x15 / country :name() (x1 / name :op1 中国))))

概念对齐和关系对齐的增加不仅改良了 AMR 表示体系,还为评价解析系统的优劣提供了新的维度,即对齐能力。该维度旨在衡量系统捕捉句子和 AMR 映射关系的程度,有利于对解析结果进行错误分析,而不仅仅是判断节点和有向弧是否正确。具体来说,概念对齐信息反映了概念和词的关系,而关系对齐信息反映了语义关系和词的关系,都是衡量对齐能力的重要因素。此外,概念对齐信息将隐含概念也区分了出来,而隐含概念大多来自复句关系和专有名词,对隐含概念进行评测有利于衡量 AMR 语义表示的完整度。这些都是缺乏对齐信息的 AMR 难以做到的。

## 2 基于概念关系对齐的中文抽象语义表示解析评测方法

目前还没有解析器能够处理带有概念、关系对齐的中文 AMR。为了推动中文 AMR 解析的进一步发展,首先应提供能够兼容概念、关系对齐的评测方法。为此,在整体性指标方面,我们在现有的评测方法 Smatch 的基础上,加入概念对齐和关系对齐的表示,提出了 Align-Smatch 指标,来总体评估中文 AMR 解析器的精度。在分项指标方面,设计了

概念对齐指标、关系对齐指标及隐含概念指标来评测解析器在对齐能力上的表现。

## 2.1 Align Smatch

整体性指标直接返回一个 0—1 之间的数值来衡量两个 AMR 图的匹配程度。其中, Smatch 是目前使用最广泛的 AMR 解析评测指标。对于两个需要进行匹配的 AMR 图, Smatch 首先会重新命名 AMR 图的节点, 并将每个 AMR 图转化为一个三元组的集合, 然后利用爬山算法 (Hill-climbing method) 进行贪心式搜索以获取两个集合最大的三元组匹配个数, 最终返回准确率 (Precision)、召回率 (Recall) 和  $F_1$  值。

Smatch 生成的每个三元组集合一般包含三个三元组类别: 对于一个节点  $N_1$ , 有三元组  $\text{instance}(N_1, C)$  表示该节点的概念 (Concept), 有三元组  $P(N_1, V)$  表示该节点具有的属性 (Property) 和属性值 (Value)。特别的, 当  $P = \text{“TOP”}$  时, 节点  $N_1$  是顶点。节点  $N_1$  和节点  $N_2$  之间的有向弧用三元组  $R(N_1, N_2)$  表示,  $R$  代表语义角色 (Role), 节点  $N_1$  是源节点, 节点  $N_2$  是目标节点。

以图 2“中国金融对外开放稳步前行”的中文 AMR 为例, 表 3 列举了由 Smatch 生成的三元组集合, 共 15 个三元组。与中文 AMR 相比, 节点被 Smatch 重新命名, 缺失了用来表示概念对齐的节点编号、被标在有向弧上的虚词“对”, 而且关系对齐也没有得到体现。可见, Smatch 指标并不适合评测中文 AMR。

表 3 Smatch 三元组

类别	三元组	数量
节点	instance (a0, 前行-01), instance (a1, 稳步), instance (a2, 开放-01), instance (a3, 金融), instance (a4, country), instance (a5, 外), instance (a6, name)	7
有向弧	manner (a0, a1), arg0 (a0, a2), arg0 (a2, a4), arg1 (a2, a3), arg2 (a2, a5), name (a4, a6)	6
节点属性	TOP (a0, “top”), op1 (a6, 中国)	2

除此之外, Smatch 本身还存在两个不足:

(1) Smatch 在比较有向弧的三元组时, 只考虑语义角色标签是否相同, 而没有考察节点的概念是否一致。这容易出现两个语义完全不同的 AMR 却有较高分数的情况<sup>[18]</sup>。

(2) Smatch 会为每个 AMR 图的根节点加入 TOP 属性三元组, 但比较时并不考虑两个根节点的概念是否相同, 这使得两个根节点概念不一样的 AMR 也会有匹配上的 TOP 属性三元组。

以图 3 和图 4 为例, 这两个语义完全不同的句子在 Smatch 中能达到约 40% 的  $F_1$  值, 共有两条边  $\text{arg0}$ ,  $\text{arg1}$  与 TOP 属性得到匹配, 这显然是不合理的。

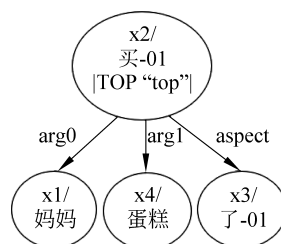


图 3 “妈妈买了蛋糕”中文 AMR 图示

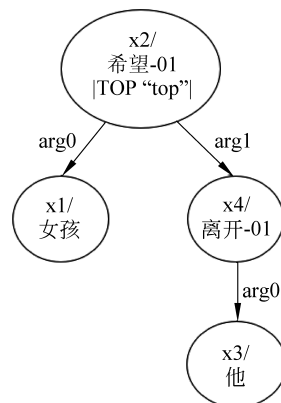


图 4 “女孩希望他离开”中文 AMR 图示

为了解决上述问题, 我们首先对 Smatch 进行了两处修正。一是在匹配有向弧三元组时, 在语义角色相同的前提下, 限制源节点和目标节点必须相同, 否则不予匹配。二是将表示根节点的三元组由属性移到有向弧的类别中, 表示为 TOP (a0, a0), 即视为一条指向自身的有向弧, 如图 5 所示。这使得在匹配根节点时, 会根据第一处修正, 考察根节点的概念是否一致, 从而避免了上述不同的根节点却能相互匹配的问题。图 3 和图 4 的 Smatch 值在第一处修正后降低为 0.13, 在第二处修正后变为 0, 更加符合人类的直觉。下文使用“FIX”代表修正版本。



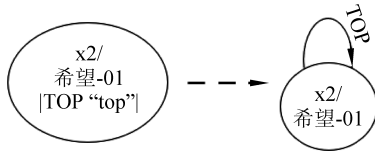


图 5 Align-Smatch 根节点图示

之后,我们在 Smatch 中融入了一个描写概念对齐的三元组和一个描写关系对齐的四元组。假如句子含有  $L$  个词,若节点  $N_1$  的索引  $I \leq L$ ,则有一个  $P = \text{"anchor"}, V = I$  的节点属性三元组  $\text{anchor}(N_1, I)$  表示节点  $N_1$  的概念对齐信息。对于关系对齐,Smatch 已有包含语义角色标签的三元组,只需再表示对应语义关系的词即可。我们将有向弧上的词  $WA(\text{Word in Arc})$  和它的索引值  $I$  视为一个整体,用一个四元组  $((WA, I), N_1, N_2)$  来表示,  $N_1, N_2$  分别表示有向弧的源节点和目标节点。以图 2 中  $x_5$  节点到  $x_4$  节点的有向弧为例,它带有语义角色标签  $\text{arg2}$  和一个介词“对”,因此要表示成  $\text{arg2}(x_5, x_4)$  和  $(\text{对}, 3)(x_5, x_4)$ <sup>①</sup>。这两个多元组表示  $\text{arg2}$  和“对”都在这条有向弧上,表示出了关系对齐信息。

表 4 展示了 Align-Smatch 形成的多元组。与表 3 相比,表 4 用有向弧类别表示根节点,而不是用节点属性类别表示;在节点属性类别中增加了表示概念对齐的三元组;在有向弧类别中增加了表示关系对齐的四元组。Align-Smatch 多元组更加完整地表示了中文 AMR。

表 4 Align-Smatch 多元组

类别	多元组	数量
节点	$\text{instance}(a_0, \text{前行-01}),$ $\text{instance}(a_1, \text{稳步}),$ $\text{instance}(a_2, \text{开放-01}),$ $\text{instance}(a_3, \text{金融}),$ $\text{instance}(a_4, \text{country}),$ $\text{instance}(a_5, \text{外}),$ $\text{instance}(a_6, \text{name})$	7
有向弧	$\text{top}(a_0, a_0),$ $\text{manner}(a_0, a_1),$ $\text{arg0}(a_0, a_2),$ $\text{arg0}(a_2, a_4),$ $\text{arg1}(a_2, a_3),$ $\text{arg2}(a_2, a_5),$ $(\text{对}, 3)(a_2, a_5),$ $\text{name}(a_4, a_6)$	8

续表

类别	多元组	数量
节点属性	$\text{anchor}(a_0, 7),$ $\text{anchor}(a_1, 6),$ $\text{anchor}(a_2, 5),$ $\text{anchor}(a_3, 2),$ $\text{anchor}(a_5, 4),$ $\text{anchor}(a_6, 1),$ $\text{op1}(a_6, \text{中国})$	7

为了生成表 4 中表示对齐信息的多元组,我们改进了 Smatch 生成三元组的算法,使其能够识别对齐信息。新算法采用 Shift-Reduce 方法依次处理中文 AMR 的字符  $w_n$ , 使用一个栈  $\text{Node\_Stack}(\text{NS})$  来储存节点,一个栈  $\text{Arc\_Stack}(\text{AS})$  来储存有向弧上的词,一个栈  $\text{Relation\_Stack}(\text{RS})$  来储存语义角色,一个缓存区(Buffer)来中转语义角色。最终,通过中文 AMR 的四个符号“)”“(“”: “/”生成相对应的状态码(State),从而对字符串  $s$ 、栈和缓存区进行相应的操作(Action),将输入的中文 AMR 转变为多元组集合  $T_w$ 。表 5 为多元组生成算法的伪码。

表 5 Align-Smatch 多元组生成算法伪码

算法 1 中文 AMR→多元组

```

Input: Chinese AMR  $w = w_0, \dots, w_n$ 
Output: a set of tuples  $T_w = \{t_1, \dots, t_m\}$ 
1:  $T_w \leftarrow \{\}$ 
2: for  $i \leftarrow w_0$  to  $w_n$  do
3:   if  $i = )$  or  $($  or  $:$  or  $/$  then
4:      $\text{state} \leftarrow 0$  to 3
5:      $t_m \leftarrow \text{Action}(\text{state}, s)$ 
6:      $T_w.\text{Push}(t_m)$ 
7:   else
8:      $s \leftarrow i$ 
9:   end for
10: return  $T_w$ 

```

算法 1 与 Smatch 生成三元组的算法的不同主要在于对符号“(”的操作,以及新增的 AS 栈和缓存区。在英文 AMR 中,“(”后是节点,但在中文 AMR 中,“(”后还可能是有向弧上的词。以表 2 右侧的中文 AMR“:  $\text{arg2}(x_3/\text{对})(x_4/\text{外})$ ”为例,共有两个“(”,第一个“(”的后面是有向弧上的词“对”,第二个“(”的后面是概念节点“外”。为了对二者进行区分,在算法 1 中,第一个“(”之前的语义角色不再直接进入 RS 栈,而是先进入缓存区,

① 为了与三元组表示一致,具体例子中将有向弧上的词和该词的索引值置于括号外。

直到操作符变为“)”才从缓存区弹到 RS 栈。这样,“(”和“)”之间的操作符“/”就能根据缓存区的状态,对前后字符串的归属做出判断:如果缓存区

不为空,则字符串为有向弧上的词,否则为节点。表 6 是以表 2 右侧的中文 AMR 为例的算法 1 部分运行流程。

表 6 以“中国金融对外开放稳步前行”中文 AMR 为例的算法 1 部分流程

NS	AS	RS	缓存区	字符串	操作符	多元组 <sup>①</sup>
.....	.....	.....	.....	.....	.....	.....
[x7, x5]	[]	[]	[]	arg1	(	—
[x7, x5]	[]	[]	[arg1]	‘,’	)	—
[x7, x5]	[]	[arg1]	[]	‘,’	(	—
[x7, x5]	[]	[arg1]	[]	x2	/	arg1 (x5, x2) anchor (x2, 2)
[x7, x5, x2]	[]	[]	[]	金融	)	instance (x2, 金融)
[x7, x5]	[]	[]	[]	‘,’	:	—
[x7, x5]	[]	[]	[]	arg2	(	—
[x7, x5]	[]	[]	[arg2]	x3	/	—
[x7, x5]	[x3]	[]	[arg2]	对	)	—
[x7, x5]	[x3, 对]	[arg2]	[]	‘,’	(	—
[x7, x5]	[x3, 对]	[arg2]	[]	x4	/	arg2 (x5, x4) (对, 3)(x5, x4) anchor (x4, 4)
.....	.....	.....	.....	.....	.....	.....

通过算法 1 将中文 AMR 语料(Gold AMR) $G_g$ 和解析器输出结果(Parsed AMR) $G_p$ 转变为两个多元组集合  $T_g$  和  $T_p$  后,我们沿用 Smatch 的爬山算法来获得两个集合最佳的多元组匹配数。

爬山算法首先对  $T_g$  和  $T_p$  多元组中的节点进行初始化匹配,得到首次匹配的多元组数和节点映射集。例如, $T_g$  有三元组 instance (a0, 前行—01), $T_p$  有三元组 instance (b0, 前行—01),如果它们概念相同,则形成映射对(a0, b0)。接着,在初始的映射集中通过“交换(Swap)”和“移动(Move)”两个操作寻找更好的匹配结果。

“交换”发生在两个映射对之间,如有(a0, b0), (a2, b2)两个映射对,交换彼此的映射对象后,形成(a0, b2), (a2, b0)两个新的映射对。

“移动”发生在一个映射对和未被匹配的节点之间,如有(a0, b0)映射对和未匹配的节点 b5,则会尝试构建映射对(a0, b5)。

最终,通过多次初始化以避免陷入局部最优的

困境<sup>[15]</sup>,搜索到两个多元组集合最优的匹配数。默认初始化次数为 5 次,具体如表 7 所示。

表 7 Align-Smatch 爬山算法伪码

#### 算法 2 爬山算法

```

Input: Chinese AMR Tuples  $T_g$  and  $T_p$ 
Output: Num(Match Tuples)
1: Restarts $\leftarrow$ 5
2:  $N \leftarrow \{\}$ 
3: for i $\leftarrow$ 1 to Restarts do
4:   Mapping, Match Num $\leftarrow$ Initialize( $T_g, T_p$ )
5:   Swap_Gain $\leftarrow$ Swap(Mapping)
6:   Move_Gain $\leftarrow$ Move(Mapping)
7:   Match Num += Swap_Gain + Move_Gain
8:   N.Push(Match Num)
9: end for
10: return Max(N)

```

根据匹配结果,将返回准确率  $P$ 、召回率  $R$ 、 $F_1$  值来评估解析器的精度。

由此,概念关系对齐信息融入到了 Smatch 指

① “—”表示没有生成多元组。将各多元组中的节点按照出现顺序从 0 开始重新编号,可得到表 4 的多元组。

标中,并能对中文 AMR 进行评测,我们将该指标称为 Align-Smatch。

$$\begin{aligned} P &= \frac{\text{num}(\text{Match Tuples})}{\text{num}(T_p)} \\ R &= \frac{\text{num}(\text{Match Tuples})}{\text{num}(T_g)} \\ F_1 &= \frac{P \times R}{P + R} \times 2 \end{aligned} \quad (1)$$

我们在 CAMR 1.0 语料库<sup>①</sup>中随机抽取了 100 句带有关对齐的中文 AMR 作为标准语料(G)。由于目前没有能够处理关系对齐信息的中文 AMR 解析器,我们使用人工标注数据作为实验的对照语料,选择了两名标注人员对这 100 个句子进行重新标注,得到对照语料 A 和 B。表 8 是三个语料的基本统计数据。

表 8 实验语料基本统计数据

指标	语料		
	G	A	B
平均句长	17.2		
平均节点数	15.4	14.9	15.7
平均节点数 (概念对齐)	12.6	12.6	12.8
平均弧数	16.1	15.7	16.4
平均弧数 (关系对齐)	2.5	2.4	2.1

将对照语料和标准语料两两组合,得到三组实验对象: A-B、A-G、B-G,分别使用 Smatch、加入概念对齐的 Concept-Smatch、加入概念关系对齐的 Align-Smatch 共三个指标进行评测,其中,Concept-Smatch 和 Align-Smatch 分别用修正版(FIX)再进行一次评测。表 9 为实验结果。

表 9 整体性指标评测实验结果

指标	组别		
	A-B	A-G	B-G
Smatch	0.73	0.78	0.82
Concept-Smatch	0.78	0.81	0.86
Concept-Smatch (FIX)	0.76	0.78	0.85
Align-Smatch	0.73	0.79	0.81
Align-Smatch (FIX)	0.70	0.77	0.81

从表 9 可以看出: 同一个指标下, B-G > A-

G > A-B,说明语料 B 的标注更为标准。A-B 的 Smatch 分数小于 AMR 要求的一致性准则 0.83<sup>[2]</sup>,说明 A 和 B 标注一致性较差。

同一组实验对象下,融入概念对齐信息的 Concept-Smatch 修正前后的分数均大于 Smatch,提升了约 3 到 5 个百分点。这主要归因于中文 AMR 标注平台对于概念对齐标注的支持。标注人员只需输入词语编号就能实现概念对齐,还用词语高亮警示防止漏标词语<sup>[9]</sup>,提高了概念对齐信息的标注准确率,进而提高了评测分数。

Align-Smatch 的分数修正前后则普遍低于 Smatch 和 Concept-Smatch,说明 A、B 在标注有向弧上的词时存在较多错误。错误主要集中在以下几点:

(1) 虚词框架结构。如“所发出的信息”位于有向弧上的词是一个框架结构“所……的”,但漏标了“所”。

(2) 有向弧的源节点或目标节点。如“只可惜这些人虽有一颗心”的连词“虽”应标在表示转折关系的节点 contrast 和概念节点“有”之间的有向弧上,但错误地标注在“可惜”和“有”之间的有向弧上。

(3) 副词。中文 AMR 考虑到副词类别繁多、内部虚化程度不同的特点,将大多数副词与其他虚词区分开来,标注为节点概念。对于一些语义较虚的副词,标注可能会出现错误。如“昨晚未雨绸缪了一下”的“一下”是频率副词,用在动词后表略微之意,语义较虚。中文 AMR 将其处理为概念节点,与另一个概念节点“未雨绸缪”之间是表示频率的 frequency 关系。B 标注则处理为与 frequency 关系对齐的词,因此无法得到匹配。

B 标注在关系对齐信息上还存在较大的漏标问题,它标有关系对齐信息的平均弧数仅为 2.1 条,相比标准语料少了 0.4 条,因此 B-G 的 Align-Smatch 分数相较 Concept-Smatch 的下降幅度要大于 A-G。

修正版的分数均不高于普通版本,这是因为修正版对于有向弧的匹配要更为严格,但也更为合理。以“人们越来越开始怀疑他们所发出的信息,是因为自己屡屡被骗。”一句为例,图 6 为人工标注 A 的中文 AMR 图示,图 7 为标准语料 G 的中文 AMR 图示。加粗部分为无法匹配的部分。

① <https://catalog.ldc.upenn.edu/LDC2019T07>

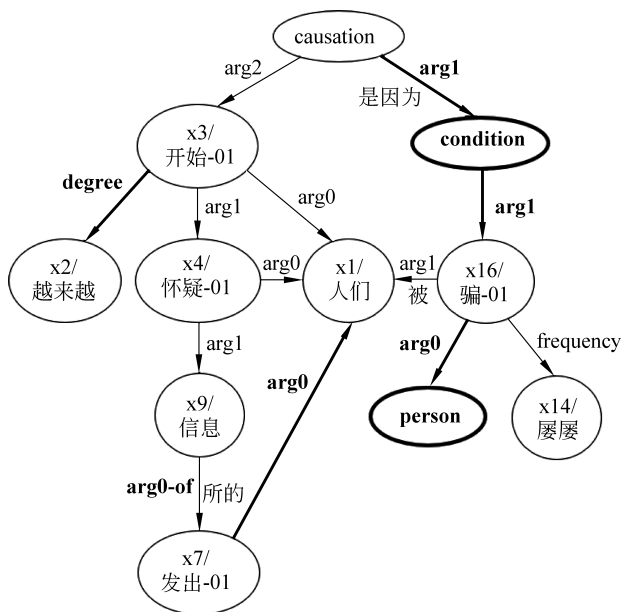


图6 中文 AMR 人工标注图示

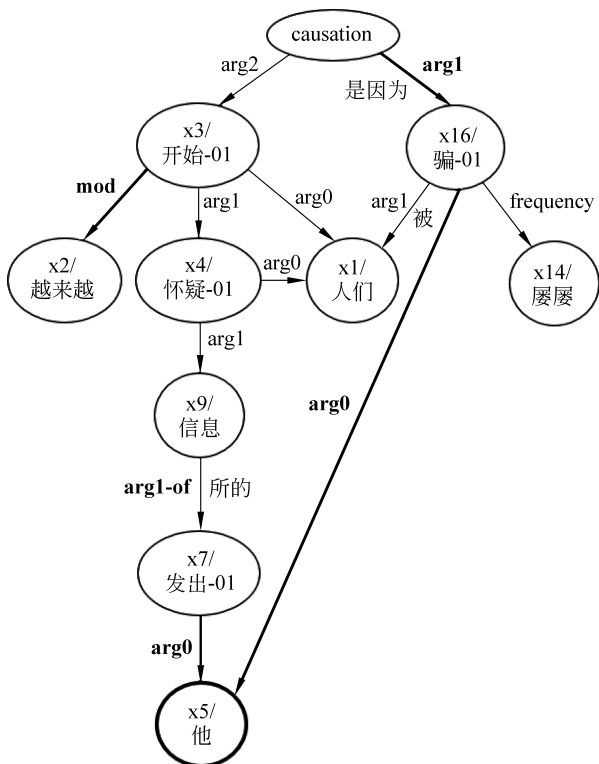


图7 中文 AMR 标准图示

两图的 Align-Smatch 得分为 0.76, 修正后为 0.73, 正确匹配的三元组数减少了一个。这是因为修正前匹配有向弧的算法主要考虑语义角色是否相同, 所以会将图 6 的  $\text{arg0}(x16, \text{person})$  有向弧与图 7 的  $\text{arg0}(x16, x5)$  有向弧相匹配, 这显然是错误的, 因为它们的目标节点并不相同。修正后的算法避免了这种情况的发生。

有趣的是, B-G 的评测分数并没有在修正前后有所波动, 而 A-G 则下调约两个百分点。这可能是因为 B 标注的有向弧更为贴合标准语料, 而语料 A 标注的有向弧可能出现源节点或目标节点的错误。

## 2.2 三个分项指标

仅仅使用整体性指标还不足以评测解析系统的性能, 因为整体性指标颗粒较粗, 可读性差, 无法体现解析系统在概念识别、关系识别等单维度子任务的表现情况, 也无法反映解析系统的优势和劣势, 不利于系统的进一步改进。因此, AMR 解析评测还需要设立分项指标, 多维度地评测解析系统, 满足具体任务的考察需求。

Damonte 等人<sup>[19]</sup>提出了九个维度的分项指标: 不考虑语义角色的 Smatch 指标、不考虑语义框架的 Smatch 指标、名词短语指标、边的重入性指标、概念指标、专有名词指标、含 wiki 属性的专有名词指标、极性属性指标、语义角色指标。我们在此基础上结合中文 AMR 的特点, 增加了概念对齐指标、关系对齐指标, 以及隐含概念指标。

概念对齐指标 (Concept Alignment, CA) 专注于衡量  $G_g$  和  $G_p$  两个中文 AMR 概念对齐信息的匹配程度, 可以考察解析器在概念对齐子任务中的表现情况。从中文 AMR 中提取节点  $N$ , 获取概念映射到词的索引值  $I$ , 以及概念  $C$ , 转换为三元组  $C(N, I)$  用以计算概念对齐指标。概念相同时, 若两个三元组的索引值也相同, 则三元组正确个数加一。最终返回正确率  $P_{CA}$ 、召回率  $R_{CA}$  和  $F_{ICA}$  值。

$$P_{CA} = \frac{\text{num}(\text{True CA Triples}(G_p))}{\text{num}(\text{All CA Triples}(G_p))} \quad (2)$$

$$R_{CA} = \frac{\text{num}(\text{True CA Triples}(G_p))}{\text{num}(\text{ALL CA Triples}(G_g))}$$

$$F_{ICA} = \frac{P_{CA} \times R_{CA}}{P_{CA} + R_{CA}} \times 2$$

概念对齐指标还可以用于英文 AMR 解析评测。近年来, 随着深度学习的兴起, 不少研究<sup>[6-7, 20, 23-24]</sup>借助基于注意力机制的编码器-解码器模型, 让解析器自动获取对齐信息, 从而舍弃了容易带来错误传播的外部对齐器。虽然不使用对齐器后, 解析结果得到了提升, 但这些解析器的对齐精度也不得而知。因此, 需要一个概念对齐指标来评估解析器在对齐子任务中的表现情况。

① 为方便区分为对齐信息, 图 6、图 7 不为隐含概念节点添加编号。



关系对齐指标(Relation Alignment, RA)考察解析器将词对应到有向弧的能力。将有向弧上的词 WA、词的索引值  $I$ , 和有向弧的源节点  $N_1$ 、目标节点  $N_2$  转换为四元组  $((WA, I), N_1, N_2)$ , 计算两个关系对齐四元组的匹配情况。有向弧上的词和索引相同时, 若源节点和目标节点也相同, 则四元组正确个数加一。最终得到正确率  $P_{RA}$ 、召回率  $R_{RA}$  和  $F1_{RA}$  值。

$$\begin{aligned} P_{RA} &= \frac{\text{num}(\text{True RA Quadruples}(G_p))}{\text{num}(\text{ALL RA Quadruples}(G_p))} \\ R_{RA} &= \frac{\text{num}(\text{True RA Quadruples}(G_p))}{\text{num}(\text{ALL RA Quadruples}(G_g))} \\ F1_{RA} &= \frac{P_{RA} \times R_{RA}}{P_{RA} + R_{RA}} \times 2 \end{aligned} \quad (3)$$

中文 AMR 里, 还存在没有概念对齐的节点。这些节点是不出现在句子中的隐含概念, 包括复句关系概念, 如“causation”概念表示因果关系; 专有名词概念, 如“上海”中隐含的“city”概念。隐含概念与句子语义的完整度以及专名识别紧密相关, 有必要对解析器生成隐含概念的能力进行考察。因此, 我们提出隐含概念指标(Implicit Concept, IC), 通过统计两个 AMR 图没有概念对齐的节点个数, 计算正确率  $P_{IC}$ 、召回率  $R_{IC}$  和  $F1_{IC}$  值, 得到该指标。

$$\begin{aligned} P_{IC} &= \frac{\text{num}(\text{True Unaligned Nodes}(G_p))}{\text{num}(\text{ALL Unaligned Nodes}(G_p))} \\ R_{IC} &= \frac{\text{num}(\text{True Unaligned Nodes}(G_p))}{\text{num}(\text{ALL Unaligned Nodes}(G_g))} \\ F1_{IC} &= \frac{P_{IC} \times R_{IC}}{P_{IC} + R_{IC}} \times 2 \end{aligned} \quad (4)$$

表 10 是 A—G、B—G 在这三个分项指标下的表现情况。二者在概念对齐指标上均有较高的分数, 与整体性指标反映的趋势一致。从该分数也可以看出, 即使有标注平台的帮助, 标注概念对齐信息仍会出现错误。特别是遇到一个概念对齐多个词、一个概念对齐语素的情况, 标注者容易错误地组合语义或拆分语义, 导致错误的概念对齐信息产生。句中的同形词和长距离依存的词也容易被错标概念对齐信息。

表 10 分项指标评测结果

指标	组别	
	A—G	B—G
$F1_{CA}$	0.94	0.95
$F1_{RA}$	0.71	0.65
$F1_{IC}$	0.75	0.85

A、B 在关系对齐指标的分数则远低于概念对齐指标, 说明关系对齐的标注难度更大。A 的关系对齐指标分数比 B 高约 6 个百分点, 说明 A 更善于标注关系对齐信息, 这与整体性指标的对照实验结果相互印证。

B 在隐含概念指标的分数上比 A 高约 10 个百分点, 显然 B 更善于补出句子的隐含概念。A 则存在漏标, 根据表 8 可算出, G 的平均隐含概念节点数约为 3 个, A 仅有约 2.3 个。具体错误类别中, A 有约 70% 的复句关系概念标注错误, B 有约 67% 的专有名词概念标注错误。

分项指标的加入将在未来让我们可以更全面、更多样地评测一个中文 AMR 解析器, 帮助发现每一个解析器的优势和劣势, 从而推动中文 AMR 解析的发展。

### 3 结语

由于目前的 AMR 解析评测指标无法兼容中文 AMR 的概念对齐信息和关系对齐信息, 使得评测结果无法真实反映中文 AMR 解析的水平, 阻碍了中文 AMR 解析的进一步发展。为此, 本文在 Smatch 指标的基础上, 加入了描写概念、关系对齐的三元组, 提出了面向中文 AMR 解析的整体性指标 Align-Smatch, 并设置了两份人工标注语料 A、B 与标准语料的对照实验。实验结果显示, 两份人工标注语料一致性较低, 语料 B 的标注更为标准, 且修正后的评测指标更为合理。

本文还提出了概念对齐指标、关系对齐指标、隐含概念指标共三个分项评估方法, 用以呈现中文 AMR 解析器在概念对齐、关系对齐、隐含概念生成等子任务中的表现情况。三个分项指标反映出 A 更善于标注关系对齐信息、语料 B 更善于标注隐含概念, 突显了各自的优势和不足。

本文发现的人工标注错误对中文 AMR 解析系统的研发也有一定的借鉴意义, 如在生成有向弧上的词时考虑是否为虚词框架、在概念对齐子任务中关注概念对齐的不同类别, 以及同形词和长距离依存等问题。

下一步工作中, 我们将把这四个评估方法, 用于目前正在开发中的包含对齐信息的中文 AMR 解析器, 以全面评测中文 AMR 的自动分析效果, 实验情况也将推进评测方法的调整。通用语义表示方法(Uniform Meaning Representation, UMR)<sup>[25]</sup>的研究也正在如火如荼地展开, 未来我们将尝试让

Align-Smatch 兼容其他语言的 AMR, 促进跨语言 AMR 解析评测工具的建设。

## 参考文献

- [1] 孙茂松, 刘挺, 姬东鸿, 等. 语言计算的重要国际前沿[J]. 中文信息学报, 2014, 28(1): 1-8.
- [2] Banarescu L, Bonial C, Cai S, et al. Abstract meaning representation for sembanking[C]//Proceedings of the Linguistic Annotation Workshop and Interoperability with Discourse, 2013: 178-186.
- [3] 戴玉玲, 戴茹冰, 冯敏萱, 等. 基于关系对齐的汉语虚词抽象语义表示与分析[J]. 中文信息学报, 2020, 34(04): 24-32.
- [4] 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.
- [5] Liu Y, Che W, Zheng B, et al. An AMR aligner tuned by transition-based parser[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 2422-2430.
- [6] Lyu C, Titov I. AMR parsing as graph prediction with latent alignment[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 397-407.
- [7] Zhang S, Ma X, Duh K, et al. AMR parsing as sequence-to-graph transduction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 80-94.
- [8] Li B, Wen Y, Bu L, et al. Annotating the Little Prince with Chinese AMRs[C]//Proceedings of the Linguistic Annotation Workshop, 2016: 7-15.
- [9] 李斌, 闻媛, 宋丽, 等. 融合概念对齐信息的中文 AMR 语料库的构建[J]. 中文信息学报, 2017, 31(6): 97-106.
- [10] Wang C, Li B, Xue N. Transition-based Chinese AMR parsing[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 247-252.
- [11] 顾敏. 基于转移神经网络的中文 AMR 解析研究[D]. 南京: 南京师范大学硕士学位论文, 2018.
- [12] 吴泰中, 顾敏, 周俊生, 等. 基于转移神经网络的中文 AMR 解析[J]. 中文信息学报, 2019, 33(4): 1-11.
- [13] Damonte M, Cohen S B. Cross-lingual abstract meaning representation parsing[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 1146-1155.
- [14] Blloshmi R, Tripodi R, Navigli R. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 2487-2500.
- [15] Oepen S, Abend O, Abzianidze L, et al. MRP 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing[C]//Proceedings of the CoNLL Shared Task: Cross-Framework Meaning Representation Parsing, 2020: 1-22.
- [16] Samuel D, Straka M. FAL at MRP 2020: Permutation-invariant semantic parsing in PERIN[C]//Proceedings of the CoNLL Shared Task: Cross-Framework Meaning Representation Parsing, 2020: 53-64.
- [17] 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, 2013: 748-752.
- [18] Song L, Gildea D. SemBleu: A robust metric for AMR parsing evaluation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4547-4552.
- [19] Damonte M, Cohen S B, Satta G. An incremental parser for abstract meaning representation[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017: 536-546.
- [20] Cai D, Lam W. Core semantic first: A top-down approach for AMR parsing[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 3790-3800.
- [21] Pourdamghani N, Gao Y, Hermjakob U, et al. Aligning English strings with abstract meaning representation graphs[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 425-429.
- [22] Brandt L, Grimm D, Zhou M, et al. Icl-hd at semeval-2016 task 8: Meaning representation parsing-augmenting AMR parsing with a preposition semantic role labeling neural network[C]//Proceedings of the 10th International Workshop on Semantic Evaluation, 2016: 1160-1166.

(下转第 38 页)

- [C]//Proceedings of the ACL, 2014: 55-60.
- [25] Ge D, Li J, Zhu M, et al. Modeling source syntax and semantics for neural AMR parsing[C]//Proceedings of the IJCAI, 2019: 4975-4981.
- [26] Klein G, Kim Y, Deng Y, et al. OpenNMT: open-source toolkit for neural machine translation[C]//Proceedings of the ACL, System Demonstrations, 2017: 67-72.
- [27] Kingma D, Ba J. Adam: a method for stochastic optimization[C]//Proceedings of the ICLR, 2015.
- [28] Papineni K, Roukos S, Todd W, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the ACL, 2002: 311-318.
- [29] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL, 2005: 65-72.
- [30] Popovi M. chrF ++: words helping character n-grams[C]//Proceedings of the WMT, 2017: 612-618.



朱杰(1996—), 硕士, 主要研究领域为自然语言处理, 文本生成, 语义分析。  
E-mail: zhujie951121@gmail.com



李军辉(1983—), 通信作者, 博士, 副教授, 主要研究领域为自然语言处理, 语义分析, 机器翻译。  
E-mail: lijunhui26@gmail.com

(上接第 30 页)

- [23] Barzdins G, Gosko D. Riga at semeval-2016 task 8: Impact of smatch extensions and character-level neural translation on AMR parsing accuracy[C]// Proceedings of the 10th International Workshop on Semantic Evaluation, 2016: 1143-1147.
- [24] Konstas I, Iyer S, Yatskar M, et al. Neural AMR: Sequence-to-sequence models for parsing and generation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 146-157.
- [25] Žabokrtsk Z, Zeman D, Ševíková M. Sentence meaning representations across languages: What can we learn from existing frameworks? [J]. Computational Linguistics, 2020, 46(1): 1-61.



肖力铭(1997—), 硕士研究生, 主要研究领域为计算语言学。  
E-mail: lmxiao1@foxmail.com



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



许智星(1994—), 硕士研究生, 主要研究领域为计算语言学。  
E-mail: xzx0828@live.com