doi:10.3969/j.issn.1673-6060.2017.07.015

面向汉语文本推理的语言现象标注规范研究

任函

(广东外语外贸大学 语言工程与计算实验室,广东 广州 510006)

摘 要:面向汉语文本推理的语言现象标注规范的方案包含两个阶段:第一,语言推理基本单元对分析,即确定两个文本片断中存在推理关系的文本对;第二,语言现象类别确定,即为语言推理基本单元对指派合适的类别。为此制定了一个包含20个类别的语言现象类别体系,探讨了语言推理基本单元对及其语言现象的判定原则和方法,说明了标注的实施流程、标注结果以及标注评估方案。

关键词:文本推理;语言现象;语义单元;推理关系

中图分类号:TP391.1 **文献标志码**:A **文章编号**:1673-6060(2017)07-0075-04

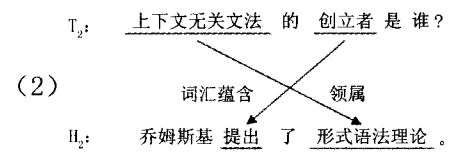
随着人工智能技术的不断发展,人们对计算机理解自然语言的能力提出了更高的要求。为了更好地理解自然语言,学界提出了文本推理的概念。文本推理是指文本表示的命题之间的逻辑推导过程,具体描述为:给定两个文本片断T和H,若H的意思可以由T推断出来,则认为T蕴含了H[1]。例如:

(1) ①T_1:鲁迅写了《狂人日记》。
②H_1:鲁迅是《狂人日记》的作者。

显然,H_1的意思包含于T_1中,或者说T_1蕴含了H_1。这一方法对如何进行语义分析给出了具体的定义和目标,计算机可以利用这一方法将复杂的语义分析问题转化为文本推理问题,从而明确分析任务,改进分析性能。例如,对于自动问答系统,若用户提问"谁是《狂人日记》的作者?",系统可利用文本推理找出与之相关的文本T_1,并抽取其中的实体名"鲁迅"作为答案。

关于实现文本推理的方法,学者们提出了许多,或从词义上分析蕴涵关系,或从句法上比较同义表达,或利用逻辑式描述推理实例,等等。然而,这些方法往往集中于针对某些特定类型的语义关系设计精确的分析方案,这种方式虽然能够提高这类问题的推理能力,然而由于推理涉及的关系众多,使得这种方式对于推理系统的整体性能提升非常有限。

为此,一些文本推理研究尝试对推理中涉及的语言现象进行分类,并据此建立语言现象的标注方法和资源[2-5]。例如:

(2)


其中,"创立者"和"提出"为"词汇蕴含"现象,即"创立者"蕴含了"提出上下文无关文法"和"形式语法理论"为"领属"现象,即"上下文无关文法"属于"形式语法理论"。上述关于语言现象标注的工作使得推理关系的标注资源在推理中的作用显得更为重要,并形成了一些可供参考的标注资源。但从关系体系上来讲,这些语言现象类别基本上是基于英语的分类标准,难以适用于其他语言。为此,有必要考察汉语文本推理中涉及的推理关系,并据此建立相应的面向汉语文本推理的语言现象标注规范。

一、汉语语言现象标注框架

为标注推理文本中的语言现象,首先需要找出不同文本内具有推理关系的片断对,然后再赋予两个片断以合适的语言现象类别。基于此,本文的汉

收稿日期:2017-04-06
基金项目:2014年国家自然科学基金青年项目"基于推理现象的中文文本推理资源建设和自动分析研究"(61402341)
作者简介:任函(1980—),男,湖北荆州人,助理研究员,博士,主要人事自然语言处理研究。

语语言标注框架包括语言推理基本单元对分析和基本推理单元对间推理关系类别的确定两部分。

(一)语言推理基本单元对分析

语言推理基本单元对(以下简称推理单元对)是不同语段(T-H)间具有推理关系的最小片段对。结合 Miller 的组块理论和高庆狮的语义单元理论[6],确定 T 和 H 间的推理单元是一种语义单元,即在一个句子或文本片段中,具有完整独立意义的单元及其组合,语法上可以是一个词、短语或结构。例如:

(3) T_3:英国作家多丽丝·莱辛获得了 2007 年诺贝尔文学奖。
H_3:英国作家多丽丝·莱辛赢得了 2007 年诺贝尔文学奖。

T_3 中的"获得"与 H_3 中的"赢得"为意义独立的词汇,且两者具有语义关系,因此可作为语义单元。例如:

(4) T_4:即便制造出健康的克隆人胚胎细胞,也必须植入代理孕母的体内。
H_4:即使是一个健康的克隆人胚胎,它还是得植入代理孕母体内。

这里,T_4 中的"即便……,也"和 H_4 中的"即使……,还是"均表示让步关系,其在 T_4 和 H_4 的上下文中也表现出相同的构式义,因此两者可作为具有推理关系的语义单元。事实上,汉语中的逻辑关系,如转折、因果、让步等,往往由特定的一系列连接词来表现。这些连接词能客观反映两个子句或语块的语义关系,因此也应作为一种语义单元存在。另一方面,汉语中还普遍存在省略连接词的现象,在这种情况下,其语义关系往往是通过标点符号来显现。例如:

(5) T_5:卡特里娜飓风侵袭美国墨西哥湾岸产油重镇,炼油厂设施受损严重。
H_5:卡特里娜飓风造成美国石油生产重镇墨西哥湾一带的产油设施重大损害。

上例中,逗号隐含了一种致使关系,而这种隐性的语义关系决定了两个子句间的逻辑联系。因此,该例中逗号也应作为语义单元。

语言现象标注工作的主要难点在于如何确定推理单元对。事实上,一旦推理单元对确定,其对应的语言现象类别也就基本确定。然而,若推理单元字数过多,则内部包含的语义关系可能越多,其结构越复杂,不仅会使分析变得困难,而且还可能对判断其它语言现象产生不利影响。为此,标注工作采用"宜少不宜多"的原则确定推理单元,即在保证能够体现推理关系的前提下,推理单元应尽可能字数少。具体而言,首先找出 T 和 H 中包含不一致文本的片断,然后逐步删除其中相同的部分,最后将可能具有推理关系的部分进行对应,构成推理单元对。例如,对于例(2),首先找出 T 和 H 中不一致的文本片断,分别是 T:"上下文无关文法"和"创立者",以及"提出"和"形式语法理论"。显然,可以发现:"上下文无关文法"与"形式语法理论"存在推理关系,"创立者"与"提出"存在推理关系。通过将上述文本片断进行对应,即完成推理单元对的确定。需要注意的是,对于结构类推理单元,利用功能词或标点而非短语作为推理单元对,从而减少推理单元对内的语义关系。

上述原则具有操作上的可行性:不仅使得标注过程中的人工分析相对简单,而且易于机器进行学习。

(二)语言现象推理关系类别的确定

语言现象推理关系类别指推理单元间具体存在何种推理关系。推理单元间语言现象的确定是分析整体文本片断间存在何种推理关系的基础。为此,需要首先制定语言现象推理关系类别体系。本文参考现有面向英语的语言现象类别研究结果,结合语料分析实践,同时考虑汉语特点,制定了面向汉语的推理关系类别体系(见表 1)。

上述语言现象推理类别与现有面向英语的语言现象推理类别的异同在于:一是部分语言现象推理类别一致,如同义、指代、上位等,这些现象普遍存在于各个语言中。二是部分语言现象推理类别包含了汉语特定的语言现象。如"结构变化"中,汉语有"把字句""被字句"等特有结构现象,这些不同结构之间具有一定推理关系。三是部分语言现象推理类别尽管在名称上一致,但其具体表现形式存在较大差异。如轻动词现象在英语中往往具有比较固定的搭配形式(如 do、have、take 等词与实义动词结合),即使省略也不会对句子造成较大影响。而汉语中,轻动词现象少有固定表现形式,且轻动词省略可能改变句子整体结构。例如:"一张炕睡三个人",这里省略了轻动词"供",将其添加到原句中则为"一张炕供三个人睡"。由于轻动词省略,为满足句法约束,动词"睡"被提前,使得两个句子中的施事与受事互换。为此,需要将这些名称相同但意义发生变化的语言现象进行区分。

表1　汉语语言现象类别

类别	实例	说明
同义	包括→包含	判断该类的标准就是两者可互换而不改变原意
上位	苹果→水果	两者是IsA的关系
比喻	办公室来了几个新面孔→办公室来了几个新成员	用一事物名称代替另一事物或用事物的部分代整体,例如转喻或可以是提喻
反义	接受→拒绝	两者意义相反
对义	东边→西边	两者意义相对
整体—部分	汽车→车轮	反映部件与整体的关系,是一种partof关系
成员	国务院下属各部委→工信部	两者是包含关系或隶属关系,即memberof的关系
缩略语	NBA→美国职篮	包括一些缩略词或共指词
空间	他出生在上海→他出生在中国	反映两者的空间或地理位置关系,通常判断这一类别需要利用背景知识
数量	地震中死亡5人,受伤10人→这次地震造成伤亡15人	反映两者数量和、差、积、商等的数量关系
结构变化	苹果我吃了→我把苹果吃了	“把字句”“被字句”等
词类活用	这本书影响了我→这本书对我有影响	某一词性成分活用为另一词性成分
中心词省略	安装过程→安装	中心词的意义隐含于修饰语中
修饰语省略	我买了双红皮鞋→我买了双皮鞋	修饰成分省略
预设	美国总统奥巴马→奥巴马是美国总统	所推导的结果是命题的前提或假设
涵义	购买→得到	表示言外之意或所蕴涵的信息
领属	鲁迅写了《狂人日记》→鲁迅的《狂人日记》	汉语中的“人名/称谓/代词+的”结构通常表示一种领属关系
等级	他得分最高→他得分比其他人高	表达级别的比较
指代	小明买了本书后他就回家了→小明回家了	代词指向判断
轻动词	同意进行谈判→同意谈判	可以省略或补充轻动词

为确定推理单元间的语言现象类别,我们将语言现象类别分为两类,包括“结构”类及“非结构”类,其中,结构类语言现象包括“结构变化”“修饰语省略”“中心词省略”“预设”和“领属”,这些现象主要体现在推理单元对在句法或语义结构上的不一致,例如“苹果被我吃了”和“我把苹果吃了”中的推理单元“被”和“把”就属于结构上的变化;非结构类语言现象包括上述类别之外的其他语言现象类别,这些语言现象主要体现在词汇或短语上的不一致,例如“他接受了这个建议”和“他拒绝了这个建议”中的“接受”和“拒绝”就属于词汇上的变化。根据这一类别划分,可以制定语言现象判定规则:即首先确定推理单元对为结构类语言现象或非结构类语言现象,再根据其中的具体类别进行选择。

二、标注实施

(一)标注流程

根据本文所提出的标注框架,其标注工作主要分为两部分:一是确定推理单元;二是确定语言现象类别。标注方案见图1。

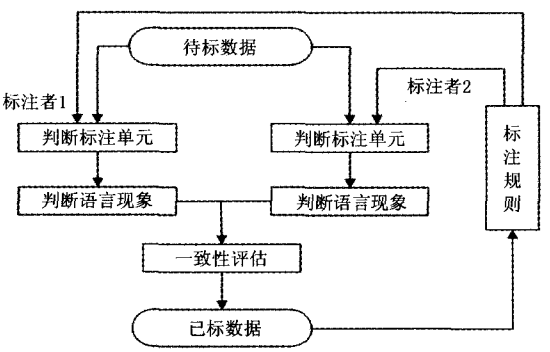


图1　汉语语言现象标注流程

在标注过程中,每轮标注从数据中选取一部分作为待标集合,由2个标注者分别进行标注,包括推理单元和语言现象的标注。标注的结果由机器进行一致性评估,并选取那些一致性程度高的标注数据放入标注数据集中。在此过程中,标注人员根据已标数据总结并更新标注规则,以此指导下一步的标注工作。该过程迭代进行,一直到标注工作全部完成。

(二)标注结果

最终标注结果是一个XML文件,例如,对于例(3),其标注结果如图2所示。

```
<corpus>
  <pair id="1" T="英国作家多丽丝·莱辛获得了2007年诺贝尔文学奖"
    H="英国作家多丽丝·莱辛赢得了2007年诺贝尔文学奖">
    <annotation>
      <unit id="1" t="获得" pos_t="10" h="赢得" pos_h="10"
        rel="synonym"/>
    </annotation>
  </pair>
</corpus>
```

图2　例(3)的标注结果文本

(三)标注评估

标注评估工作主要是一致性评估,其不仅决定了标注数据的质量,而且为下一步标注工作提供了重要的规则指导。为此,标注工作采用迭代方案,即将语料分为若干部分,每部分由两个以上的标注者进行标注,标注完成后对标注结果进行一致性评估,并组织分析讨论,确定最合适的标注结果,并修订标注规则,在下一轮标注中即以此标准进行标注。这种方案的可行性在于,标注者在每个阶段都能够纠正标注过程中的错误,保证了标注结果的可靠性。

三、结论

本文以分析现有语言现象标注规范的现状及存在的问题为切入点,提出了一种面向汉语文本推理的语言现象标注规范。该方案建立了一个两阶段标注方案,即首先确定推理文本对中的推理单元,然后在此基础上确定其包含的语言现象;提出了语言现象标注的具体方案,并进行了可行性分析。这一工作将有助于推动汉语文本推理资源的建设,进而改进汉语文本推理系统的性能。

参考文献:

[1]DAGAN I, GLICKMAN O. Probabilistic textual entailment: generic applied modeling of language variability[C]. In proceedings of pascal workshop on learning methods for text understanding and mining, 2004:1 - 6.

[2]BENTIVOGLI L, CABRIO E, DAGAN I, et al. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference[C]//Proceedings of the international conference on language resources and evaluation, 2010:3542 - 3549.

[3]GAROUFI K. Towards a better understanding of applied textual entailment: annotation and evaluation of the RTE - 2 dataset[D]. Germany: Saarland University, 2007:29.

[4]SAMMONS M, VYDISWARAN V G V, ROTH D. "Ask not what Textual Entailment can do for you…"[C]//Proceedings of the annual meeting of the association for computational linguistics, 2010:1119 - 1208.

[5]KANEKO K, MIYAO Y, BEKKI D. Building Japanese textual entailment specialized data sets for inference of basic sentence relations[C]//In proceedings of the 51st annual meeting of the association of computational linguistics, 2013:273 - 277.

[6]高庆狮. 统一语言学基础[M]. 北京:科学出版社, 2009:24 - 30

(责任编辑:王凤玲)

Research on Annotation of Linguistic Phenomena for Chinese Text Reasoning

REN Han

(Language Engineering and Computation Laboratory, Guangdong University of Foreign Studies, Guangzhou 510006, China)

Abstract: This paper proposes a Chinese language annotation specification for Chinese text reasoning. The program consists of two stages: First, the basic unit of linguistic reasoning for analysis, that is, the text pairs that determine the inference relations in two text fragments; Second, the category of linguistic phenomena is defined as the basic unit of linguistic reasoning and the appropriate categories are assigned. For this reason, this paper develops a system of linguistic phenomena containing 20 categories, at the same time, the principle and method of judging the basic unit of linguistic inference and its linguistic phenomena are discussed. Finally, the implementation process, annotation results and annotation evaluation scheme are illustrated.

Key words: text reasoning; language phenomenon; semantic unit; reasoning relation