汉语文本蕴涵库的设想与实现*

罗　琳　刘金凤　郭嘉伟　亢世勇　姜岚
鲁东大学汉语言文学院 烟台 264025
E-mail: luolin1007@126.com

摘　要：本文从自然语言处理的角度出发，提出在真实文本中构建符合汉语特点的文本蕴涵库，在实践中探索蕴涵产生的类型并尝试标注的难度，总结经验，以期为计算机的语义识别等相关研究提供一定的资源储备。

关键词：文本蕴涵；文本；假设；转换；类型

The Assumption and Consummation of Chinese Text Entailments Database

Luo Lin, Liu Jinfeng, Guo Jiawei, Kang Shiyong, Jiang Lan
School of Chinese Language and Literature, Lu Dong University , Yantai , 264025
E-mail: luolin1007@126.com

Abstract: On the view of natural languadge processing, this paper suggests to build a text entailments database from the real text, and tries to find the types of entailments and the difficulty of producing in practice. Certain sematic resources is looked forward to reserving for the computer and other related research..

Keywards: Text Entailment; Text; Hypothesis; Transform; Type

1　引言

目前，蕴涵问题已成为语义学、语言哲学、语言逻辑学、语用学及计算机人工智能系统等学科的共同课题。国家 863 项目“基于人类认知的语义知识融合、学习与计算技术”正是以蕴涵知识为基础理论之一，试图通过为不同的语言表达形式之间的蕴涵关系建立通用的模型，指定一个语言表达形式的意义可以从另一个语言表达形式推断出来的条件，来发展一种识别语言表达多样性的技术路线，从而为语言处理应用服务。结合项目开展的需要，我们构建了汉语文本蕴涵库，分综合文本蕴涵库和一个单独的蕴涵反例库，以便计算机用来学习。本文拟在此基础上介绍相关方面的工作。

2　文本蕴涵的定义

“蕴涵” (implication)本是逻辑学中连接命题的五种逻辑连接词中的一种。语义学引入这一概念后，学界对它的定名有“蕴含”、“语义蕴含”、“衍推”等。语义学中有“蕴涵”关系的两个命题之间在语义上是有联系的，而逻辑学中这两个命题之间却不一定有关系。

本文所提到的文本蕴涵（textual entailment）是指文本和假设之间的一种广义关系。在自然语言中，有的两个语言形式表达基本相同的意义，它们之间是同义互释关系；有的意思是从另

*本文得到国家 863 计划项目（2007AA01Z173）的资助。

一个中推断出来的，它们之间是蕴涵关系。本文所称的文本蕴涵是从自然语言处理的角度对文本与其假设之间的同义关系和蕴涵关系的统称。它存在于一定语境中，允许语句以外的参照物的参与。原文本和蕴涵假设分别用“T”“H”表示。

国外已举行过多次专门的文本蕴涵识别邀请赛，利用计算机技术识别文本蕴涵。但汉语是一种缺乏形态变化的语言，我们必须立足实际，进一步探索适合汉语文本蕴涵识别的道路。

3 文本蕴涵的层面类型规范

文本蕴涵的层面类型指的是在人工识别文本蕴涵时所属的层面及其所依据的具体手段的类型。在构建汉语文本蕴涵库时，我们暂规范为四个层面的汉语文本蕴涵，分别是词汇层面、语法层面、语义层面和言语层面。现以综合文本蕴涵库为例，具体分析说明其下的蕴涵类型。库中共有 4263 个原始文本，形成 12539 个文本蕴涵假设，包括 11776 个真值蕴涵假设，763 个蕴涵假设反例。

3.1 词汇层面的文本蕴涵

词汇层面的文本蕴涵指的是词汇本体上通过词汇的转换机制（即词语替换）而得到的文本蕴涵。这种推理机制是利用词语本身之间的关系，一般不涉及句法结构的改动。如：

T：这小狗还不赖，怪伶俐的，一口就咬破了这家伙的手指头！

H：这小狗还不赖，怪伶俐的，一口就咬破了这家伙的手！（部分与整体关系）

在综合文本蕴涵库中，词汇层面的蕴涵类型及其个数分布如下表：

所属层面	中 类	蕴 涵 类 型	个 数	此中类蕴涵占真值蕴涵的比例
词汇层面	词汇本体关系 1848 个	同义关系	1622	15.69%
		反义关系	111	
		上下义关系	81	
		部分与整体关系	34	

3.2 语法层面的文本蕴涵

语法层面的文本蕴涵主要是通过相应语法规则转换机制来获取。通过标注，我们又总结为句法转换、语用转换、同指互参三个中类，其下的类型设置理论上又是开放式的，随着标注的深入而不断增加。具体介绍如下：

句法转换方面，是指具有相同的词汇元素的同义句法结构之间的转换。如：

① T：鲁镇的酒店的格局，是和别处不同的：都是当街一个曲尺形的大柜台，柜里面预备着热水，可以随时温酒。

H：鲁镇的酒店的格局，是和别处不同的：都是当街一个曲尺形的大柜台，柜里面预备着可以随时温酒的热水。（补语一定语转换）

语用转换方面，我们认为是汉语特殊句式的运用及句类之间的转换运用等，它们大都有突出强调等方面的特殊语用，其转化体现了句与句之间的整体性，特征明显。如：

② T：这种花称作菊，看来是有道理的。

H：这种花称作菊，看来不是没有道理的。（肯定句-双重否定句转换）

同指互参类型，它提供了文本中不同的词项之间的等值关系，用所指相同的任意词项来替换文本中的某个词项。如：

③ T：棚屋后边有一个小小的窗口，由此望去，可以看到田野边缘的那片树林。

H：由棚屋后边一个小小的窗口望去，可以看到田野边缘的那片树林。

T 蕴涵了 H，其中涉及一个同指互参转换：此⇒棚屋后边一个小小的窗口。

为了后续工作的便利，我们将语法转换的类型做了细致的区分，归并得到 62 种，限于篇幅，我们仅将各层面下出现个数超过 20 的类型列举如下：

所属层面	中 类	蕴 涵 类 型	个 数	此中类蕴涵占真值蕴涵比例
语法层面	句法层面 共1768个	宾语-主语转换	226	15.01%
		补语-定语转换	30	
		补语-状语转换	61	
		定中式-主宾式转换	24	
		定中式-主谓式转换	435	
		同位式-主宾式转换	36	
		同位式转换	33	
		谓语-定语转换	27	
		主谓式-定中式转换	98	
		主语-宾语转换	350	
		状语-定语转换	28	
	语用层面 共2233个	“把”字句-“被”字句转换	151	18.96%
		“把”字句-主动式转换	27	
		“被”字句-“把”字句转换	52	
		“被”字句-主动式转换	56	
		被动式-“被”字句转换	57	
		被动式-主动式转换	27	
		陈述句-感叹句转换	39	
		单句-复句转换	120	
		果因句式-因果句式转换	39	
		反问句-陈述句转换	57	
		复句-单句转换	181	
		感叹句-陈述句转换	36	
		双重否定-肯定式转换	34	
		疑问句-陈述句转换	64	
		致使式转换	222	
		主动式-“把”字句转换	368	
		主动式-“被”字句转换	526	
		主动式-被动式转换	83	
		去除数量限定	22	
	同指互参	同指互参	827	7.02%

3.3 语义层面的文本蕴涵

语义层面的文本蕴涵指的是预设和语义蕴含。

预设原是哲学概念，这里是指话语的非断言部分表达的意义，在否定命题中预设仍具有恒定性，如果命题X和命题‘非X’都可以推导出Y，那么Y就是X的预设。

语义蕴含，既包括命题之间的单向推导关系，又包括双向推导关系。它是两个语句的整个陈述内容之间的关系。

值得注意的是，语义蕴含与预设二者都具有主观性，它们本身并不必然是客观真实的或正确的。在标注信息时，我们一般用“否定测试法”来区分预设和语义蕴含：

T：老张的儿子很老实。 → H：老张有儿子。（预设）

“老张的儿子很老实”和“老张的儿子不老实”，都包含着预设“老张有儿子”。

T：每年春二三月，粉红的桃杏花开罢，不久就开绿叶衬托的艳丽的海棠花，很热闹。

H：艳丽的海棠花每年春天二三月开。（语义蕴含）

如果T变为否定句，那H就不存在了。

在综合文本蕴涵库中，词汇层面的蕴涵类型及其个数分布如下表：

所属层面	中 类	蕴涵类型	个 数	此中类蕴涵占真值蕴涵的比例
语义层面	语义蕴含	语义蕴含	3402	28.87%
	预 设	预 设	1388	11.79%

3.4 言语层面的文本蕴涵

言语层面的文本蕴涵指的是会话含义。在具体的会话过程中，仅理解言语形式的“字面意义”是不够的，还必须依据当时的语境推导出言语形式的“言外之意”。会话含义就是字面意义同语境结合，通过语用规则而推导的一种间接的隐性的意义。就书面材料而言，我们可以简单地理解为话题的上下文。例如：

T：“我是一个美国兵。”伞兵说，“你们愿意把我藏起来吗？”“哦，当然啦。”法国女人说着便把他带进屋里。→ H：法国女人同意把他藏起来。

在进行会话含义的获取时，人工判断难度尚不太大，计算机的自动推理应该是很难的，所以不是我们开发的重点。此层面获得的蕴涵总数为308个，占库内真值蕴涵的2.62%。

4 汉语文本蕴涵库的构建与规范

4.1 语料来源

目前，我们处理的原始语料是人民教育出版社九年义务教育三年制初中语文(1-6 册)、四年制初中语文(6-7 册)及义务教育课程标准实验教材小学语文(三年级下册)的课文。汉语文本蕴涵库以句子为基本单位，多数是一个文本即指一个句子，但少数文本是句群乃至段落。

4.2 汉语文本蕴涵库标注的信息

我们利用清华大学开发的文本蕴涵辅助标注工具进行信息的标注,主要包含以下方面：

（1）数据源信息。T和H的相关文章的（版本、册数、课文序号、标题、作者）。

（2）文本-假设对。原始文本由工具直接导入，蕴涵假设需在界面中人工自行录入。

（3）蕴涵关系的真假值。“真”“假”分别用“T”“F”表示，说明蕴涵关系是否成立。

（4）蕴涵关系的类型及难度等级。文本蕴涵类型分为大中小三类，但直接标小类名称。难

度等级也分为难中易三级，表示蕴涵假设在某一种类型中获得时所处的难易等级。

（5）文本蕴涵的采集方法。可利用的是阅读理解法、问题回答法、互释习得法三种。

（6）信心度。标注信息中设有“信心度1”和“信心度2”，分别代表两个标注者对同一假设对的不同看法。信心度分为5、4、3、2、1、0六个等级，根据两位标注者信心度的平均值来确定这个蕴涵假设的典型性。

5 标注结果样例及存在的问题分析

5.1 标注结果样例

每一篇标注的课文形成一个ent文件，文件中蕴涵信息以如下的方式呈现：

```
<数据源信息>
 <版本>人民教育出版社九年义务教育三年制初中语文</版本>
 <册数>第六册</册数>
 <课文序号>8</课文序号>
 <标题>菜园小记</标题>
 <作者>吴伯箫</作者>
</数据源信息>
<蕴涵对信息>
 <编号>0</编号>
 <文本 T>
      <原始文本>
      种花好，种菜更好。
      </原始文本>
      <句法语义标注>
      [S 种花/v]D1[P 好/a]V1，[S[P 种/v]V[O 菜/n]O]U2[D 更/d[P 好/a]V2。
      </句法语义标注>
 </文本 T>
 <假设 H>
      <文本假设 1>
           <原始文本>
           种菜比种花好。
           </原始文本>
           <句法语义标注>

           </句法语义标注>
           <逻辑真假>T</逻辑真假>
           <采集方法>互释习得法</采集方法>
           <蕴涵类型>句法转换</蕴涵类型>
           <难度等级>易</难度等级>
           <信心度 1>5</信心度 1>
```

<信心度 2>0</信心度 2>

</文本假设 1>

</假设 H>

</蕴涵对信息>

5.2 问题与发现

（1）在综合文本蕴涵库中，语法层面获得的文本蕴涵占到 40.99%，其次是语义层面 40.66%、词汇层面 15.69%，言语层面 2.62%。笔者认为，语法层面也许是最具显性和可操作性的。当读者接触到文本时，往往会受原始形式的一定影响，而利用不同的语法形式表达相同的意义，所以语法层面获得的文本蕴涵最多。言语层面推理性更强所以获取的难度最大。

（2）这一阶段我们处理的文本是以简单单句为基本，限于目前的需要和处理难度，对于复句没有做深入规范，但有关复句类型之间转换，尤其是因果关系复句转换的问题在标注过程中也有出现。

（3）要把涉及到的文本引用全面，避免缺失，才能便于观察转换机制。尤其是对待同指互参类型的假设时，对应的两部分内容应体现全面。

（4）有些蕴涵类型的命名有待进一步规范和实践的检验。首先，涉及可逆向的双项转换类型，定名时应将两项位置互换，如“‘把’字句-‘被’字句转换”、“‘被’字句-‘把’字句转换”分列为两种。对于没有提到的新类型，能归类定名的尽量定名，不能定名的暂标其中类名称。其次，有些类型的设置也许有待商榷，如暂归在语法层面的“去除冗余信息”“ 去除修饰成分”等。再次，对于同一文本，不同的人可能会得到不同的蕴涵，也可能定义成不同的蕴涵类型，但由于其开放性的特性，难免需要后续整理。

6 结语

识别文本蕴涵是一个十分复杂的工作，需要综合运用语言知识、世界知识和逻辑推理等多方面的知识。本文只是一次在自拟的理论规范基础上结合实践的尝试，所得到的语料尚需进一步整理和规范。希望通过建立相对集中的汉语文本蕴涵语料库，探寻蕴涵内部的规律性，对未来相关方面的深入研究提供有利的支持与帮助。

致谢：本项研究是在亢世勇教授主持下完成的，除了署名作者之外，还有李敏、胡晓清、徐德宽、李海英、徐艳华等老师和王莉、赵文、杨慧丽、张超男、翟蕾艳、李文玲、王磊、闫君、周明海等同学。谨致谢忱。

参 考 文 献

[1] 利奇. 语义学[M]. 上海外语教育出版社, 1982.

[2] 尼尔·史密斯, 达埃德尔·威尔逊. 现代语言学[M]. 外语教学与研究出版社, 1983.

[3] 石安石. 语义研究[M]. 语文出版社, 1994.

[4] 袁毓林. 文本蕴涵的类型层级和推理机制与识别模型. (尚未公开发表)

[5] 胡裕树, 范晓. 试论语法研究的三个平面[J]. 新疆师范大学学报(哲学社会科学版), 1985, (2).

[6] 郭幸楷. 句子间的蕴涵关系[J]. 中国俄语教学, 1998, (3).

[7] 王跃平. 试析语义蕴涵的基本特征[J]. 徐州师范人学学报(析学社会科学版), 2005, (5).

[8] 江结宝. 话语隐性意图的理解规则. 语言文字应用, 2006, (3).