ZYLXGGTQL Chinese Propbank 简介

张誉泷 1800013030 张弋丰 1800013041

2020 年 11 月

1 摘要

2 Chinese Propbank 简介

Chinese Propbank 数据集是基于 Chinese Treebank 数据集和 Propbank 数据集建立的中文语料库。

在语言知识资源的两种分类中，Chinese Propbank 属于第二种，即标注语料库。对于词汇表中的每个词，Chinese Propbank 给出了它的每一次具体用例的性质，即它的 Token（实例）层的知识表示。

对于词汇表中的每个词，Chinese Propbank 以该词在句子中的不同含义为标准，将包含该词的例句分为若干类，每一类包含该词用法相同的若干个例句。Chinese Propbank 会对每一个例句的句法结构和句子中各个成分的语义进行标注。

3 Chinese Propbank 标注规范与组织形式

Chinese Propbank 数据集使用 Chinese Treebank 标注规范表示中文句子的句法结构，并用 Propbank 标注规范对各个成分进行语义角色标注。

3.1 Chinese Treebank 标注规范

Chinese Treebank 数据集以句法树的结构表示句子的句法结构，子节点对应的结构是父节点对应的结构的子结构。句法树中的每个结点均会被标记，用来表示该结点对应的结构的种类或该节点对应的词的词性。对于句法树中的结点，Chinese Treebank 提供了四大类标记：词性标记、句法标记、功能标记和空范畴标记。

3.1.1 词性标记

Chinese Treebank 提供了 33 种标签对词性进行标记。这些标签会被用来标注句法树中的叶结点，用来表示每个词的性质。部分标签及含义见下表：

标签	英文含义	中文含义
AD	adverbs	副词
AS	Aspect marker	体态词
CD	Cardinal numbers	数词
DEG	Associative “的”	连接词 “的”
FW	Foreign words	外来词
IJ	interjection	感叹词
M	Measure word	量词
OD	Ordinal numbers	序数词
P	Prepositions	介词
PN	pronouns	代词

3.1.2 句法标记

Chinese Treebank 提供了 23 种标签对句法进行标记。这些标签会被用来标注句法树的非叶结点，用来表示子结构组成该结构的方式。部分标签及含义见下表：

标签	英文含义	中文含义
ADJP	Adjective phrase	形容词短语
ADVP	Adverbial phrase headed by AD	由副词开头的副词短语
CP	Clause headed by C	由补语引导的补语从句
DP	Determiner phrase	限定词短语
NP	Noun phrase	名词短语
PP	Preposition phrase	介词短语
PRN	Parenthetical	插入语
QP	Quantifier phrase	量词短语
VP	Verb phrase	动词短语
VCD	Coordinated verb compound	并列动词复合

3.1.3 功能标记

Chinese Treebank 提供了 26 种标签对成分的功能进行标记。这些标签会被用来标记句法树中的任何结点，用来表示该结点对应的结构的功能。部分标签及含义见下表：

标签	英文含义	中文含义
ADV	Adverbial	副词
CND	Condition	条件
DIR	Direction	方向
EXT	Extent	范围
IJ	Interjective	插入语
IO	Indirect object	间接宾语
LOC	Locative	处所
OBJ	Direct object	直接宾语
SBJ	Subject	主语
TMP	Temporal	时间

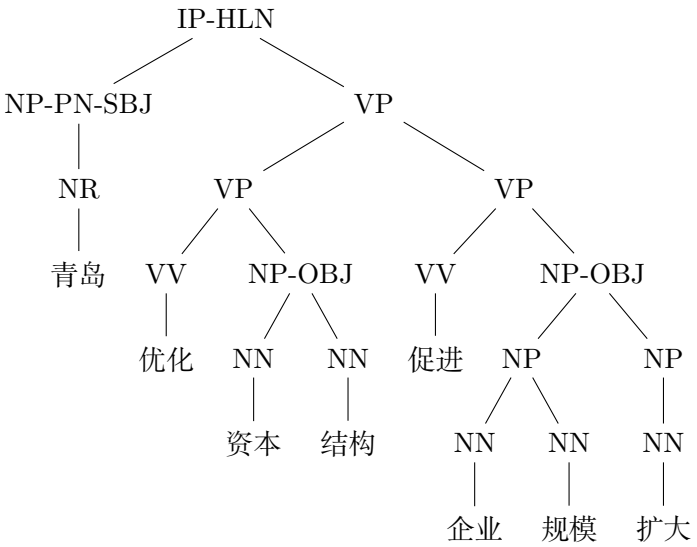
3.1.4 空范畴标记

Chinese Treebank 提供了 7 种标签对空范畴成分进行标记。空范畴成分是指不在句子中出现，但可以起到语法、语义作用的成分。标签及含义见下表:

标签	英文含义	中文含义
OP	operator	在 relative constructions 相关结构中的操作符
pro	dropped argument	丢掉的论元
PRO	used in control structures	在受控结构中使用
RNR	right node raising	右部节点提升的空范畴
T	trace of A’-movement	A’移动的虚迹，话题化
*	trace of A-movement	A 移动的虚迹
?	other unknown empty categories	其他未知的空范畴

3.1.5 Chinese Treebank 标注举例

以句子“青岛优化资本结构促进企业规模扩大”为例，该句子在 Chinese Treebank 中的表示形式如下:



该句子由“青岛”（NP-PN-SBJ，专有名词短语作主语）和“优化资本结构促进企业规模扩大”（VP，动词短语）构成。“青岛”自身是一个专有名词（NR），“优化资本结构促进企业规模扩大”由“优化资本结构”（VP，动词短语）“促进企业规模扩大”（VP，动词短语）构成。“优化资本结构”由“优化”（VV，其他动词）和“资本结构”（NP-OBJ，名词短语作直接宾语）构成，“优化”、“资本”、“结构”自身均是普通名词（NN）。“促进企业规模扩大”由“促进”（VV，其他动词）和“企业规模扩大”（NP-OBJ，名词短语作直接宾语）构成。“企业规模扩大”由“企业规模”（NP，名词短语）和“扩大”（NP，名词短语）构成。“企业”、“规模”和“扩大”自身均是普通名词（NN）。

可见，该句法树完整地展示了该句子的句法结构。

3.2 Propbank 标注规范

在 Propbank 数据集中，一个句子中的每个动词都会被当作语义谓词，其他文本会被标注为该谓词的论元或附加角色，对应论元标记与附加角色标记。

3.2.1 论元标记

论元是与语义谓词有直接关系的文本成分。Propbank 提供了 5 种标签对论元进行标记，用来表示该论元与语义谓词之间的关系。论元标记均形如“ARG*”，标签及含义见下表：

标签	英文含义	中文含义
ARG0	proto-agent	原型施事
ARG1	proto-patient	原型受事
ARG2	benefactive, instrument, attribute, end state	工具、受益人、属性等
ARG3	start point, benefactive, instrument or attribute	起点、受益人、属性等
ARG4	end point	终点

3.2.2 附加角色标记

附加角色是与语义谓词没有直接关系的文本成分。Propbank 提供了 18 种标签对附加角色进行标记，用来表示该附加角色的作用。附加角色标记均形如 “ARGM_***”，部分标签及含义见下表:

标签	英文含义	中文含义
ARGM-TMP	Temporal	时间
ARGM-LOC	Locative	地点
ARGM-DIR	Directional	方向
ARGM-MNR	Manner	方式
ARGM-ADV	Adverbials	附加的，默认标记
ARGM-CAU	Cause	原因，起因
ARGM-PRP	Purpose	目的
ARGM-DSP	Direct Speech	直接引语

3.3 Chinese Propbank 组织形式

Chinese Propbank 中知识的组织方式如下。其中，句法结构使用 Chinese Treebank 标注规范进行标注，各个语义角色使用 Propbank 标注规范进行标注。

3.3.1 Chinese Propbank 数据集结构

```
Chinese propbank:
    词1
        Frameset1（用法1）
            Frame1（例句1）
                句法结构
                各个语义角色
            Frame2（例句2）
            ...
```

```
        FrameX（例句X）
    Frameset2（用法2）
    ...
    FramesetX（用法X）
词2
...
词N:
```

在 Chinese Propbank 数据集中，每个语义谓词对应一个或多个 Frameset，每个 Frameset 代表该语义谓词的一种用法。

每个 Frameset 中均包含若干个 Frame。每个 Frame 通过 Chinese Treebank 和 Propbank 标注规范描述了一个含有该语义谓词的句子的句法结构和该句子中各个成分的语义角色。位于同一个 Frameset 中的 Frame 对应的句子中语义谓词的用法是相同的。

3.3.2 Frameset 举例

以“对”为例，该词在 Chinese Propbank 中对应 6 个 Frameset。此处截取 Frameset: f5 和 Frameset: f6 进行介绍。

Frameset: f5 对应的信息如下：

```
Frameset: f5
   ARG0: person described
   ARG1: entity arg0 has done enough for
Frame:
 (ADVP (AD 这样子))
(NP-SBJ (NN 政党)
        (NN 领导人))
(VP (VP (ADVP (AD 才))
        (VP (VV 能够)
            (VP (VPT (VV 对)
                     (DER 得)
                     (VV 起))
                (NP-OBJ (NT 过去)))))
ARGM-ADV: 这样子
ARG0: 政党 领导人
ARGM-DIS: 才
ARG1: 过去
REL: 对
```

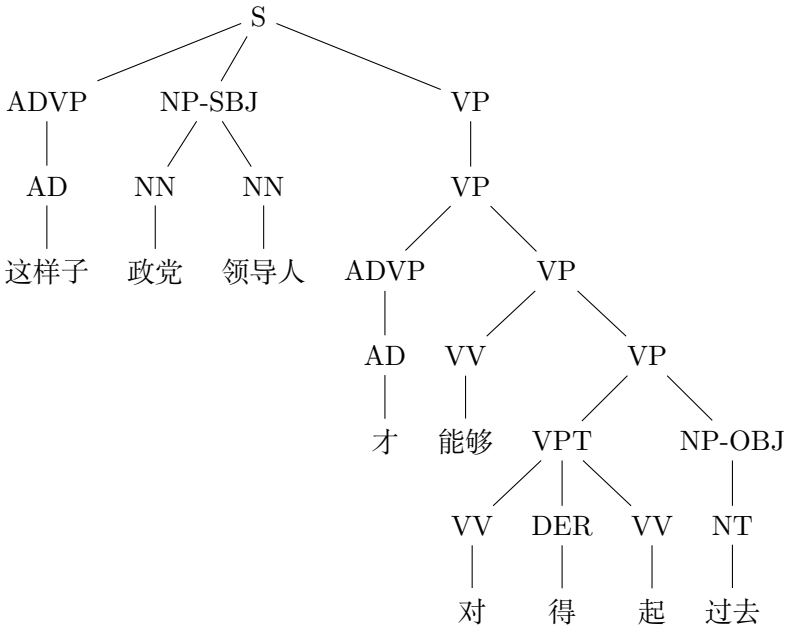
可见，Frameset: f5 对应的“对”的用法是“对得起”，即一个个体已经为另一个个体做了足够的事情。与“对”有直接关系的论元包括 ARG0 和 ARG1。ARG0 为原型施事，即做了足够的事情的个体

“person described”；ARG1 为“对”的原型受事，即“对得起”的对象“entity arg0 has done enough for”。

Frameset: f5 中共有一个 Frame，对应句子“这样子政党领导人才能够对得起过去”。该 Frame 中含有句法树的字符串表示，并记录了附加角色 ARGM-ADV、论元 ARG0、附加角色 ARGM-DIS、论元 ARG1 和语义谓词 REL 的具体含义。

附加角色 ARGM-ADV 对应成分“这样子”，作为附加标记，表示对“对”的补充；论元 ARG0 对应成分“政党领导人”，作为原型施事；附加角色 ARGM-DIS 对应成分“才”，作为连接词，表示句子中前后两部分之间的逻辑关系；论元 ARG1 对应成分“过去”，作为原型受事；语义谓词 REL 对应动词“对”。

可得到句法树为：



从中可以得到该句子的句法结构，例如“对得起过去”构成一个动词短语（VP），该短语由“对得起”（VPT，V 得 R 结构）和“过去”（NP-OBJ，名词短语构成直接宾语）构成。“对得起”由“对”（VV，其他动词）、“得”（DER，得）和“起”（VV，其他动词）构成，“过去”本身是一个时序词（NT，表示时间的名词）。

Frameset: f6 对应的信息如下：

```
Frameset: f6
  ARG0: agent
  ARG1: entity arg0 treats
Frame:
 ( (IP (IP (NP-SBJ (-NONE- *pro*))
           (VP (ADVP (AD 不))
               (ADVP (AD 光是))
```

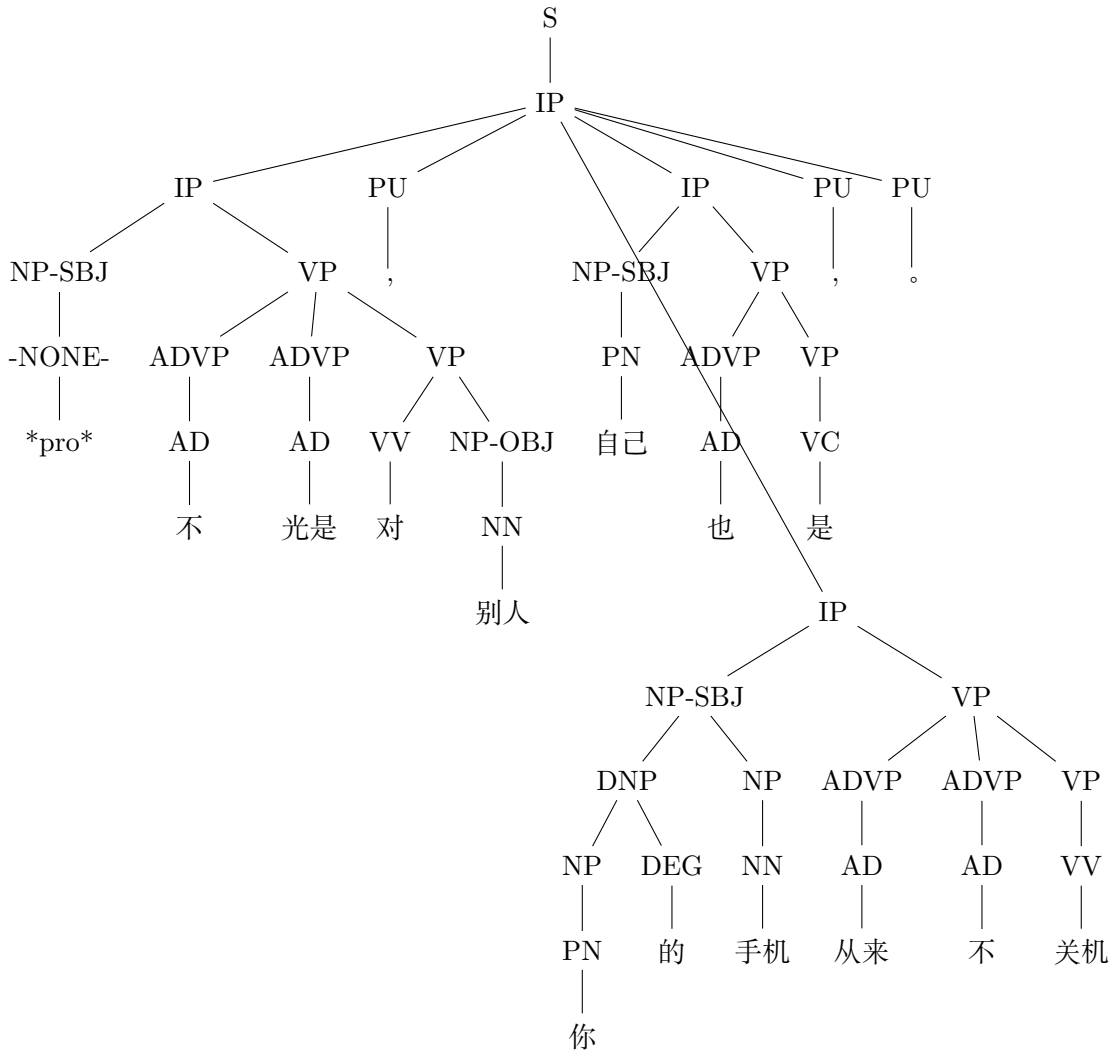
```
            (VP (VV 对)
                (NP-OBJ (NN 别人 )))))
        (PU ， )
        (IP (NP-SBJ (PN 自己 ))
            (VP (ADVP (AD 也 ))
                (VP (VC 是 ))))
        (PU ， )
        (IP (NP-SBJ (DNP (NP (PN 你 ))
                         (DEG 的 ))
                    (NP (NN 手机 )))
            (VP (ADVP (AD 从来 ))
                (ADVP (AD 不 ))
                (VP (VV 关机 ))))
        (PU 。)))
ARG0: *pro*
ARGM-DIS: 不光是
ARG1: 别人
REL: 对
```

可见，Frameset: f6 对应的“对”的用法是“对待”，即一个个体对另一个个体表示某种态度或施以某种行为。与“对”有直接关系的论元包括 ARG0 和 ARG1。ARG0 为原型施事，即表示态度或施加行为的个体“agent”；ARG1 为“对”的原型受事，即态度或行为的接受者“entity arg0 treats”。

Frameset: f6 中共有一个 Frame，对应句子“不光是对别人，自己也是，你的手机从来不关机。”。该 Frame 中含有句法树得字符串表示，并记录了论元 ARG0、附加角色 ARGM-DIS、论元 ARG1 和语义谓词 REL 的具体含义。

论元 ARG0 对应的成分 *pro*，代表在受控结构中使用，即原型施事并未在句子中出现；附加角色 ARGM-DIS 对应成分“不光是”，作为连接词，表示句子中前后两部分之间的逻辑关系；论元 ARG1 对应成分“别人”，作为原型受事；语义谓词 REL 对应动词“对”。

可得到句法树为：



从中可以得到该句子的句法结构，例如“不光是对别人”构成一个动词短语（VP），该短语由“不”（ADVP，由副词开头的副词短语）、“光是”（ADVP，由副词开头的副词短语）和“对别人”（VP，动词短语）构成。“过去”本身是一个副词（AD），“光是”本身是一个副词（AD），“对别人”由“对”（VV，其他动词）和“别人”（NP-OBJ，名词短语做直接宾语）构成。“别人”本身是一个普通名词（NN）。

4 Chinese Propbank 内容举例

Chinese Propbank 的最新版本 Chinese Proposition Bank 3.0 含有 173206 个涵盖了 24642 个动词和 1421 个名词。其中，1337 个动词具有多于一种含义（即对应多于一个 Frameset），使动词的 Frameset 总数达到 26467，平均每个动词对应 1.07 个 Frameset；48 个名词具有多于一种含义（即对应多于一个 Frameset），使名词的 Frameset 总数达到 1528，平均每个名词对应 1.08 个 Frameset。本次提供的 Chinese Propbank 数据集仅包括 Chinese Proposition Bank 3.0 中的部分动词。

5 Chinese Propbank 的优点与不足

6 Chinese Propbank 的应用

7 结论