因果关系抽取技术文档

慕飞腾

2018-01-08

一、因果模式分类

按照汉语的语言习惯，因果关系模式分为以下五类：

1 c1_cause_c2_effect

该模式由两个连接词表征因果关系，c1和c2为连接词。原因在前，结果在后。其中：

c1属于 {因为, 因, 由于}

c2属于 {为此, 故而, 故, 于是, 所以, 因而, 因此, 从而, 以至于, 以致于}

c1与c2的组合属于 {因为为此, 由于故, 既然所以, 因以至于, 因故而, 由于因而, 因故, 因于是, 因以致于, 因为因而,
由于为此, 由于所以, 由于以致于, 因为所以, 由于从而, 因为故, 由于以至于, 因为因此, 由于因此, 因为于是, 因所以, 因为致使, 由于于是, 既然因此, 由于致使, 因为以至于, 因为故而, 因因此, 由于故而, 因为以致于, 因因而, 因为从而}

2 c1_effect_c2_cause

该模式由两个连接词表征因果关系，c1和c2为连接词。结果在前，原因在后。其中：

c1属于 {之所以}

c2属于 {由于, 因为, 基于, 在于, 缘于}

c1与c2的组合属于 {之所以在于, 之所以缘于, 之所以由于, 之所以因为, 之所以基于}

3 cause_c_effect

该模式由一个连接词表征因果关系，c为连接词。原因在前，结果在后。其中：

c属于 {因此, 因而, 以致于, 以至于, 所以, 故, 致使}

4 effect_c_cause

该模式由一个连接词表征因果关系，c为连接词。结果在前，原因在后。其中：

c属于 {源于, 在于, 因, 由于, 缘于, 因为}

5 cause_v_effect

该模式由动词表征因果关系，v为动词。原因在前，结果在后。其中：

v属于 {促使, 引来, 引起, 引发, 导致, 招致, 使得, 造成, 使}

模式1-4的因果关系为句间因果，模式5为句内因果。

注意到，所用的因果线索词并不多。原因在于最终的目的是得到仅有少部分噪音的因果关系。所以我们仅使用了能够明确表征因果关系的线索词。同时其他的句法、语法等约束也被应用，以此来提高抽取结果的准确性，得到可信的结果。

二、因果关系抽取

1 模式优先级确定

总的来说，约束越多的模式越先被匹配。约束越多，说明模式更确切。

现有句子：

他如今之所以能取得这么大的成就，因为他一直都很努力。

和模式：

a：之所以……，因为……

b：……，因为……

在匹配时，a与b都能成功匹配，但是因为模式a的约束较多，所以先匹配。这样就能保证匹配的正确性。最终的匹配优先级为：模式1=模式2>模式3=模式4

句间因果与句内因果分别处理。抽取流程：

1. 根据优先级及关键词判定模式类型。
2. 确定确切模式后转到该模式的抽取函数下处理。
3. 返回结果为空说明该模式下的其他约束没有满足。

2 各模式下的抽取规则

2.1 c1_cause_c2_effect

为了尽可能准确的确定因果边界，限定如下规则：

规则1：c1和c2之间有且只有一个‘，’，而且必须在c2的前一位置。考虑如下句子：

句子1：「有一个问题，因为整句话都没有逗号真的好烦啊所以没有办法确定因果边界」

句子2：「有一个问题，因为整句话都没有逗号所以好烦啊，没有办法确定因果边界」

句子3：「有一个问题，因为这些人写句子都不断句，好烦啊，所以结果都是错的，并没有办法确定因果边界」

在这些情况下，没有办法确定原因结果的边界。基于此规则可以准确的抽取原因部分。

规则2：句子中逗号的数量大于4个，则舍弃。原因在于这种句子往往很复杂，原因或结果部分并没有紧跟线索词，所以很难确定因果边界。

经过上述过滤，句子情况如下：

1）一个逗号：「因为要经常面对风吹雨打太阳晒，所以最让人担心的是塑料窗的老化问题。」

2）两个逗号：

「在冰层中，由于下层结冰的速度比上层要慢，故盐度随深度的加大而降低」

或者：

「更由于他长期旅居日本，所以在中国很少露面，记载很少」

这两种情况因果边界很好确定。

3）三个逗号：

「因为...，所以...，...，...」

此情况下统一取‘所以’到第二个‘，’之间的内容作为结果部分

「...，因为...，所以...，...」和 「...，...，因为...，所以...」

这两种情况统一取‘所以’之后的部分作为结果部分。

抽取到原因结果部分后，去除标点 {“, ”, 《, 》, ：, ；, 、, (,), 〈, 〉},

然后如果原因与结果长度都在[2, 12]之间时，返回抽取结果。否则返回空。

2.2 c1_effect_c2_cause

为了尽可能准确的确定因果边界，限定如下规则：

规则1：c1和c2之间有且只有一个‘，’，而且必须在c2的前一位置。

基于此规则可以准确的抽取结果部分。

至于原因部分的确定，较为复杂。

考虑句子「在这些文章中，鲁迅多次提及杨荫榆，对她的所作所为给予冷嘲热讽，就如后来人们所知道的，杨荫榆之所以能够出名， 不是因为她早年大胆的抗婚之举，也不是因为她是中国近现代历史上第一位女大学校长，而是因为女师大风潮更准确地说，是因为鲁迅对她在女师大的所作所为进行的讥讽嘲骂」

简单来说出现了句式：「之所以..., 不是因为..., 而是因为...」。句子中出现了：否定词+线索词。这种情况下需要找出真的原因部分。此时进行如下步骤：

步骤1: 判断原因线索词的真假。

修饰词集合：modifier = ['不', '非', '并非', '不是', '并']

判定规则：c2前一个词在modifier中，或者 c2前一个词为‘是’且c2之前第二个词在modifier中。

否定原因候选词枚举如下：{并非[因为/由于/在于/基于/源于]，不[在于/基于/源于/由于]，不是[因为/由于/基于/在于/源于]，并非是[因为/由于]，并不是[因为/由于]，.....}。

步骤2: 遍历句子寻找真正的原因线索词，如果没找到，返回空。如果找到，进行步骤3。

步骤3: 确定原因线索词后的逗号‘，’的数量。根据逗号数量的不同，确定原因部分的边界。满足原因部分的词数量少于13个。

举例说明：

句子：「妓女是一种毫无尊严的职业，而且这种职业之所以能够长盛不衰，并不是因为有人喜欢做，而是因为有人喜欢捧场」

抽取结果：「有人 喜欢 捧场----能够 长盛不衰」

抽取到原因结果部分后，去除标点 {“, ”, 《, 》, ：, ；, 、, (,), 〈, 〉}，

然后如果原因与结果长度都在[2, 12]之间时，返回抽取结果。否则返回空。

2.3 cause_c_effect

为了尽可能准确的确定因果边界，限定如下规则：

规则1：c的前一个词必须是‘，’。

规则2：整个句子中逗号的数量不能超出2个。

经过以上过滤，句子情况如下：

1）只有一个逗号：……，因此……

此时可以准确地确定因果边界

2）有两个逗号：

分为：

「……，……，因此……」，取两个‘，’之间的内容作为原因部分。

「……，因此……，……」，取‘因此’到‘，’之间的内容为结果部分。

抽取到原因结果部分后，去除标点 {“, ”, 《, 》, ：, ；, 、, （, ）, 〈, 〉}，然后如果原因与结果长度都在[2, 12]之间时，返回抽取结果。否则返回空。

2.4 effect_c_cause

为了尽可能准确的确定因果边界，限定如下规则：

规则1: c的前一个词为‘是’，进行如下步骤：

步骤1: 遍历c之后的部分，如果有‘，’则返回空。

步骤2: 如果整个句子中逗号数量为0或者大于2，返回空。

经过以上过滤，句子句式为：「……，……是由于……」或者「……，……，……是由于……」。

当逗号数量为1时，逗号与线索词之间可能会存在一些词，这些词可能是指代内容，指代的内容在上一部分出现。如果这部分的词数量小于等于3，则认为是指代情况，指代内容（结果部分）为逗号之前的内容。如果大于三个，无法判断结果是哪部分，返回空。

逗号数量为2时，忽略第一个逗号之前的内容，转成一个逗号的情况下处理。

规则2: c的前一个词不为‘是’时，进行如下步骤：

步骤1: 如果整个句子中逗号数量为0或者大于2，返回空。

步骤2: 一个逗号情况下，如果线索词出现在句首，句式变为：「由于……，……」，进而抽取因果部分。如果逗号是线索词前一个词，句式变为：「……，由于……」，进而抽取因果部分。

步骤3: 两个逗号情况下，如果两个逗号不是都在线索词之前，返回空。此时句式为「……，……，由于……」。规定结果部分长度小于13，以此确定结果部分是否包含线索词之前的所有内容。

抽取到原因结果部分后，去除标点 {“, ”, 《, 》, ：, ；, 、, (,), 〈, 〉}，然后如果原因与结果长度都在[2, 12]之间时，返回抽取结果。否则返回空。

2.5 cause_v_effect

此模式下线索词集合为 {促使, 引来, 引起, 引发, 导致, 招致, 使得, 造成, 使}。根据汉语使用习惯，又可以分为如下几种情况，每种情况其处理方式又有所不同。

第一种：「使得，使」

这种情况下要求：线索词前一个词必须为‘，’。然后进行如下步骤：

步骤1： 对句子进行句法分析，如果线索词没有‘DBL’，返回空。最终结果部分为线索词的‘DBL’和‘VOB’的集合。

步骤2： 从线索词前的逗号开始向前遍历，找到第一个‘，’或‘：’或‘；’，这部分之间的内容为候选原因部分。如果候选原因中出现了「因，因为，由于，由」中的词，则真正的原因部分为这个词到线索词之间的部分。

举例说明：

句子1:「近年试作红日、雪林等，具现代气息，使人耳目一新」

抽取结果：「具 现代 气息----人 耳目一新」

句子2:「与水性好的沐童是死对头，两人在日后的生活中因为麻烦不断，使他们的矛盾升级」

抽取结果：「麻烦 不 断----他们 的 矛盾 升级」

第二种：线索词的后一个词为‘的’，且线索词不是「使得，使」

这种情况下，为了简化问题，只关注特定的句法结构：

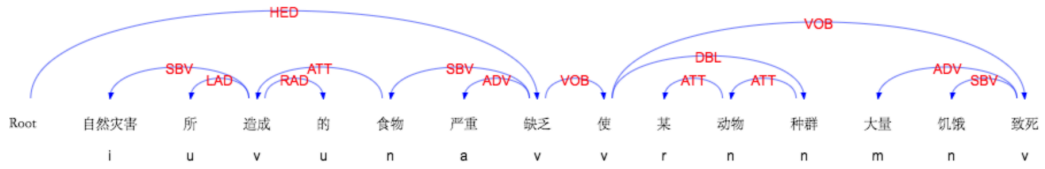

图5.1 句法树举例

如图5.1所示，原因部分为线索词的子节点，结果部分从线索词开始沿‘ATT’关系向上遍历。下面具体说明。

原因部分的确定：

遍历以线索词为根结点的子树的所有节点，去除**线索词**及**‘的’**后，得到候选原因部分。接着从线索词的前一个词开始向前遍历，找到第一个‘，’或‘：’或‘；’为止，如果这部分中出现了「因，因为，由于，由」中的词，则真正的原因部分为这个词到线索词之间的部分。同时，**线索词前一个词如果是‘所’或者‘而’，则去除。**

结果部分的确定：

从线索词开始沿‘ATT’关系向上遍历，到达最终的词。当这个词的关系为‘SBV’时，如果其父节点的词性为‘n’或‘a’，则将最终位置指向其父节点。如果其父节点的词性为‘v’且父节点有两个字符组成的词，同时满足父节点在之后的第一个‘，’之前，则将最终位置指向其父节点。结果部分是从线索词开始到最终位置之间的部分。

如图5.1所示，最终位置先是指向‘食物’，然后指向‘缺乏’。结果部分为‘食物 严重 缺乏’

举例说明：

句子1:「用途用于各种功能性心律失常、室上性及室性异位期外收缩、心房纤维颤动和麻醉引起的心律不齐等」

抽取结果：「和 麻醉----心律 不 齐」

句子2:「自然灾害所造成的食物严重缺乏使某动物种群大量饥饿致死」

抽取结果：「自然灾害----食物 严重 缺乏」
句子3:「连绵的棘林偶尔会间有小片的棕榈林、盐土乾草原和由火或砍伐造成的稀树草原」
抽取结果：「火 或 砍伐----稀树 草原」

第三种：线索词不是「使得，使」且后一个词不是'的'

这种情况下，要求线索词是整个句子的句法树的根节点。抽取线索词的宾语作为结果部分。抽取线索词的主语作为候选原因部分。从线索词开始向前遍历，如果出现「因，因为，由于，由」中的词，则真正的原因部分是这个词到线索词之间的部分。
举例说明：
句子1:「笔记本的丢失同样也可能引起数据丢失」
抽取结果：「笔记本 的 丢失----数据 丢失」
句子2:「怀孕期间胎儿父亲因他人侵权行为造成死亡的，婴儿出生后享有请求赔偿的权利王德钦诉杨德胜、泸州市汽车二队交通事故损害赔偿纠纷案」
抽取结果：「他人 侵权 行为----死亡」
句子3:「1000年前，“播种者”在火星上培育的生物由于彗星的撞击导致毁灭」
抽取结果：「彗星 的 撞击----毁灭」

抽取到原因结果部分后，去除标点 {“, ”, 《, 》, ：, ；, 、, （, ）, 〈, 〉}，然后如果原因与结果长度都在[2, 12]之间时，返回抽取结果。否则返回空。

三、负样本抽取

首先判定如果句子没有因果关系，则进行负样本抽取。负样本的抽取分为连词和动词分别处理。

1 由连词引导的负样本抽取

在汉语中，有些连词不表示因果关系，但是表示其他关系，比如转折，并列，假设等。这些词集合如下：
{只有， 不然， 不仅， 还是， 甚至， 不论， 反之， 而是， 并且， 况且， 不仅仅， 或者， 不管， 其次， 否则， 尽管，
以及， 然后， 即便， 及其， 不过， 即使， 而且， 另外， 不但， 虽说， 以此， 何况， 接着， 以便， 纵然， 可是，
此外， 只是， 虽然， 无论， 从此， 但， 然而， 一边， 除非，……}

如果句子出现两个连词，比如「不但，而且」、「虽然、但是」等，抽取其间的部分作为负样本。如果只有一个连词，根据逗号的位置，抽取出两个部分作为负样本。

2 由动词引导的负样本抽取

有些动词在某些情况下可能表示因果关系，他们的集合为candidate_verb_set：{诱使， 促使， 引来， 促成， 引发， 诱发， 诱导}。所以在判定时，当动词不属于candidate_verb_set也不属于因果动词时，抽取这个动词的主语和宾语作为负样本。

四、因果样本抽取结果及评估

从10GB预料中进行抽取，抽取结果如下：

因果动词	533944
因果连词	288745

从用因果动词抽取出的样本中随机选取了2000条进行评估，其中错误的有209个。