1. 实体识别阶段

这一阶段的任务是将文本中的显式和隐式短语同时识别出来，而这些短语在本发明也被统一定义为实体，注意这与自然语言处理领域中的命名实体识别方法有所不同。与以前的方法倾向于提取具有长参数短语的关系三元组不同，本发明可以进一步将其分解为细粒度更高的三元组，其中提出的隐式短语扩展旨在通过设计的扩展规则识别不连续的隐式短语，这将使最后的提取结果更加简洁、紧凑，并提高下游任务的性能。

1.1 基于规则的显式短语识别

如表2所示，这一步中的显式短语通过建立关于词性标注信息的正则表达式识别出来，短语类型包括动词短语（VP）、名词短语（NP）和带有补足语的动词短语（VVP）。在这些正则表达式中，（MD）代表“动词，情态动词”，（VB-）代表动词的不同类别，例如VB是“动词原形”，VBD是“动词过去时”，VBG是“动名词或现在分词”，VBZ 是“动词，第三人称单数”；(NN-) 代表不同类别的名词，例如 NN 是“名词单数”，NNP 是“专有名词”，NNS 是“名词复数”；(JJ) 代表形容词；(RB) 代表副词；(DT) 代表限定词；(CD) 代表基数；(IN) 代表任何连词或介词；(TO) 代表不定式“to”。

|  |  |  |
| --- | --- | --- |
| 类型 | 正则表达式 | 示例 |
| VVP | (MD)\*(VB-)+(JJ)\*(RB)\*(JJ)\*(VB-)?(DT)?(TO)+(VB)+ | was able to start |
| (MD)\*(VB-)+(JJ)\*(RB)\*(JJ)\*(VB-)?(DT)?(IN)+(VBG)+ | has been blended without sacrificing |
| VP | (MD)\*(VB-)+(CD)\*(JJ)\*(RB)\*(JJ)\*(VB-)?(DT)?(IN|TO)+ | were released in |
| (MD)\*(VB-)+(JJ)\*(RB)\*(JJ)\*(VB-)?(DT)?(IN|TO)+ | do not contribute significantly to |
| (MD)\*(VB-)+(JJ)\*(RB)\*(JJ)\*(VB-)+ | were readily proven |
| (VB-)+(IN|TO)+(DT)?(NN-)\*(IN|TO)+ | applying in a total of |
| (DT)?(JJ)\*(NN-)+(IN|TO)+ | the director of |
| (MD)\*(VB-)+ | would work |
| NP | (CD)\*(DT)?(CD)\*(JJ)\*(CD)\*(VBD|VBG)\*(NN-)\*(POS)\*-(CD)\*(VBD|VBG)\*(NN-)\*(VBD|VBG)\*(NN-)\*(POS)\*(CD)\*(NN-)+ | Mrs. Marcos’s trial |
| (CD)\*(DT)?(CD)\*(JJ)\*(CD)\*(VBD|VBG)\*(NN-)-\*(POS)\*(CD)\*(VBD|VBG)\*(NN-)\*(VBD|VBG)\*(NN-)-\*(POS)\*(CD)\*(NN-)+ | a federation official |
| (PRP)+ | them |

表2. 显式短语的正则表达式

以“Huawei, a leading global provider of informational and communicational technology and infrastructure, which is located in Shenzhen, is committed to building an intelligent and fully-connected world”为例，使用基于规则的识别方法后，我们可以得到“communicational technology”、“infrastructure”和“intelligent and fully-connected world”这样的名词短语和“a leading global provider of”，“is located in”和“is committed to building”这样的动词短语。

1.2基于深度学习的显式短语识别

这一步使用自然语言处理工具（例如 Spacy 和 Stanford core NLP）进行命名实体识别，在本发明中我们也将它们归类为“显式短语”。图5展示了对例句“Over the last quarter Apple sold nearly 20 thousand iPods for a profit of $6 million. By contrast, Sony sold only 7 thousand Walkman music players.”的识别结果，识别的结果形如（Apple, ORG），其中“Apple”是识别出的结果，“ORG”代表识别结果的类型。

1.3隐式短语扩展

仍以“Huawei, a leading global provider of informational and communicational technology and infrastructure, which is located in Shenzhen, is committed to building an intelligent and fully-connected world”为例，经过前两步之后，从子句“informational and communicational technology and infrastructure”我们只能得到“communicational technology”和 “infrastructure”，而这些不足以表达完整的信息。因此，我们需要尽可能识别出所有的短语，即便它们在原句上的位置是不连续的，比如“informational technology”，“communicational technology”，“informational infrastructure”和“communicational infrastructure”，通过尽可能多地识别这些短语，可以保证最终生成的关系三元组能完整地表达出原句信息。这一步骤具体做法如下：

首先，我们做出了如下的符号定义：假设句子为S，并且已经识别出所有显式短语。为了区分显式名词短语和隐式名词短语，我们将所有显式名词短语定义为ENP（Explicit Noun Phrase），如（1）所示，

S = { ENP1, ENP2, …, ENPi, …, ENPm; VP1, VP2, …, VPj, …, VPn; OUTSIDE } (1)

其中ENPi 代表句中第i个显式名词短语，VPj 代表句中第j个动词短语，OUTSIDE代表句中既不属于ENP也不属于VP的部分，m、n分别代表句中显式名词短语、动词短语个数。

ENPi = { ni,1 , ni,2 , …, ni,p , …, ni,u } (2)

VPj = { vj,1 , vj,2 , …, vj,q , …, vj,v } (3)

OUTSIDE = { o1, o2, …, or, …, ow } (4)

其中，ni,p 代表ENPi的第p个词项，vj,q 代表VPj 的第q个词项，or 代表OUTSIDE部分的第r个词项，u、v、w分别代表词数。

然后，我们制定了两种不同的扩展规则（识别名词短语中形容词修饰语的第一规则，识别名词短语中核心名词的第二规则）来识别句中的隐式名词短语。其中，第一规则用于识别名词短语中的形容词修饰语，第二规则用于识别名词短语中的核心名词或词块，其中形容词修饰语识别旨在尽可能全面地找到名词短语中的修饰语，而核心名词或词块识别旨在尽可能地在名词短语中找到核心名词。

**识别名词短语中形容词修饰语的第一规则**：该规则旨在尽可能全面地找出名词短语中的形容词修饰语。在同一个名词短语中，形容词修饰语会对该名词短语中的所有名词成分起共同修饰的作用，所以识别出形容词修饰语的作用在于对现有名词短语作细粒度更高且不影响语义的拆分，以提高信息抽取方法的原子性抽取能力。具体地，如果我们通过依存解析从S中提取到conj(ni,p, or)，amod(ni,q, ni,p) (p<q)的关系，那么我们就可以用or 替换ENPi中的ni,p 并且将（ni,1, …, ni,p-1, or, …, ni,q, …）扩展为一个隐式短语，其中conj表示并列关系，amod表示形容词修饰关系。比如，在例句“is committed to building an intelligent and fully-connected world”中，ENP是“fully-connected world”并且“fully-connected”是n1,1 “world”是n1,2 ，“intelligent”是OUTSIDE部分的o2。通过依存解析，我们可以得到关系conj:and(fully-connected, intelligent)和amod(world, fully-connected)，因此我们可以将“intelligent world”扩展为一个隐式短语。

**识别名词短语中核心名词的第二规则：**该规则旨在确定并划分名词短语中的核心名词成分。句中不同的核心名词往往会关联不同的关系，如果一个名词短语中存在多个不同的核心名词，会造成关系抽取混乱和细粒度不高的问题，严重影响信息抽取的效果，影响下游任务的使用，比如知识库构建。所以识别核心名词的作用在于对现有名词短语中的核心名词进行区分，以提高信息抽取方法的跨子句抽取和原子性抽取能力。具体地，如果我们通过依存解析从S中提取到conj(ni,p, nj,q)(i≠j)或者conj(ni,p, or)的关系并且ni,p，nj,q和or的词性标签都是“NN”(即名词)，那么我们可以用ENPj中的nj,q或者or替换ni,p，并且将(ni,1, …, ni,p-1, nj,q, …)或(ni,1, …, ni,p-1, or, …)扩展为一个隐式短语。比如，在例句“a leading global provider of informational and communicational technology and infrastructure”中，ENP1是“communicational technology”并且“technology”是n1,2，ENP2是“infrastructure”同时也是n2,1。通过依存解析，我们可以得到关系conj:and(technology, infrastructure)，同时，它们的词性都是“NN”。另外，从第一规则我们可以事先得到隐式短语“informational technology”，因此我们可以额外扩展两个隐式短语“communicational infrastructure”和“informational infrastructure”。

经过隐式短语扩展后，我们可以得到细粒度更高的名词短语，如“intelligent world”、“informational technology”、“communicational infrastructure”和“informational infrastructure”。 仍以“Huawei, a leading global provider of informational and communicational technology and infrastructure, which is located in Shenzhen, is committed to building an intelligent and fully-connected world”为例，使用隐式短语扩展和不使用隐式短语扩展时生成的名词短语的差异如表3所示。

|  |  |  |
| --- | --- | --- |
| 类别 | 包含步骤 | 名词短语 |
| 不使用隐式短语扩展 | 基于规则识别  基于深度学习识别 | Huawei |
| Shenzhen |
| communicational technology |
| infrastructure |
| intelligent and fully-connected world |
| 使用隐式短语扩展 | 基于规则识别  基于深度学习识别  隐式短语扩展 | Huawei |
| Shenzhen |
| communicational technology |
| **informational technology** |
| **communicational infrastructure** |
| **informational infrastructure** |
| **intelligent world** |
| **fully-connected world** |

表3. 使用/不使用隐式短语扩展生成名词短语的比较

2.关系抽取阶段

关系抽取阶段通过制定一组结合语言场景与依存解析的规则来过滤掉实体识别阶段中无用的短语，并进一步利用保留的短语生成关系三元组。

2.1关系三元组生成

由于自然语言的特性，同一语义会有多种不同的表达方式，这也进一步导致了自然语言语句的冗长性和结构复杂性。 通过大量的实证研究，我们总结出六种常见的、不同的语言场景来处理这一特性，这六种场景并不是完全独立的，通过对这六种场景的不同组合，我们可以对一些结构复杂的长句进行有效的处理。下面是对这六种场景的具体说明。

**场景-1**：场景-1的目标是处理所有包含主谓宾结构的句子。主谓宾结构是英语简单句中一种常见的基本构造，句子中的主要成分是主语、谓语动词和宾语。其中主语表示句子主要说明的人或事物，谓语说明主语的动作，宾语表示动作行为的对象。在本规则中，我们除了抽取出句子中主谓宾结构的主要成分，同时还根据谓语动词的形式来确定时态，包括主动时态和被动时态两种，并生成相应的关系三元组。具体主谓宾结构规则如下：

1）带有主动时态的主谓宾结构语句：通过依存解析，如果我们提取到nsubj(ni,p, vk,r)和dobj(vk,r, nj,q)（i≠j）这样的关系（其中nsubj表示主动时态中的句法主语关系，dobj表示直接宾语关系），那么<NPi; VPk; NPj>就可以作为一个关系三元组被提取出来。以“The pictures provide irrefutable evidence.”为例，“The pictures”是NP1并且“pictures”是n1,1，“irrefutable evidence”是NP2且“evidence”是n2,2，“provide”既是VP1又是v1,1。从依存解析结果中，我们可以得到nsubj(pictures, provide)和dobj(provide, evidence)，所以我们可以抽取<The pictures; provide; irrefutable evidence>作为关系三元组。

2）带有被动时态的主谓宾结构语句：通过依存解析，如果我们抽取到nsubjpass(ni,p, vk,r)和nmod(vk,r, nj,q)（i≠j）这样的关系（其中nsubjpass代表被动时态中的句法主语关系，nmod代表名词性修饰语关系），那么<NPi; VPk; NPj>就可以作为一个关系三元组被提取出来。以“The temple is supported by marble columns.”为例，“The temple”是NP1并且“temple”是n1,1，“is supported by”是VP1且“supported”是 v1,2 ，“marble columns”是NP2且“columns”是n2,2，另外我们可以得到依赖关系nsubjpass(temple, supported)和nmod(supported, columns)，所以<The temple; is supported by; marble columns>会作为关系三元组被抽取出来。

另外，在一些其他情况下，比如例句“An electric car was developed by Toyota in 2015.”中，我们会得到依存关系nmod:agent(developed, Toyota)和nmod:in(developed, 2015)，这时我们会抽取出<An electric car; was developed by; Toyota>和<An electric car; was developed in; 2015>两个关系三元组，因为我们会根据nmod判断出的介词来替换动词短语中原本的介词，来进一步提高抽取结果语义表达的准确性。

**场景-2**：该场景的目标是处理所有包含主系表结构的句子。主系表结构也是英语中简单句基本构造的一种，句子中的主要成分是主语、[系动词](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=697638&ss_c=ssc.citiao.link)（最常见的是be和become，还有感官动词如feel、look、sound等）和表语。其中主语是句子的中心；系动词是用来辅助主语的动词，它本身有词义，但不能单独用作谓语，其后必须跟表语；表语用来修饰主语，说明主语的身份、性质、特征和状态。本规则用于抽取出句子中构成主系表结构的主要成分并生成关系三元组。具体而言，如果我们从依存解析中获得关系cop(nj,q, vk,r)和nsubj(nj,q, ni,p)（i≠j）（cop表示系动词关系），我们会生成关系三元组<NPi; VPk; NPj>（通常，在该场景中出现的动词短语都是系动词）。 例如，在“Roger Federer is one of the greatest tennis players.”这句话中，“Roger Federer”是NP1且“Federer”是n2,1，“is”代表VP1和v1,1，“one of the greatest tennis players”是NP2且“player”是n2,3，我们将生成关系三元组<Roger Federer; is; one of the greatest tennis players>。

**场景-3**：场景-3的目标是处理句子中的补语成分。补语是句子中起补充说明作用的成分，其作用对象是[主语](https://baike.baidu.com/item/%E4%B8%BB%E8%AF%AD/105797)和[宾语](https://baike.baidu.com/item/%E5%AE%BE%E8%AF%AD/87544)，具有鲜明的定语性描写或限制性功能，在[句法](https://baike.baidu.com/item/%E5%8F%A5%E6%B3%95/8019868)上是不可或缺的。因为起补充说明的作用，所以补语通常会表达新的意思和关系，因此本规则用于识别句中存在的补语关系，抽取补语成分表达的新含义生成关系三元组。具体来说，如果我们从句子S抽取到依赖关系xcomp（开放补语关系），那么我们将为原句S额外添加一个依赖关系nsubjpass/nsubj (视原句具体时态而定，nsubjpass和nsubj分别代表被动时态和主动时态中的句法主语关系)。以 “The small rectangular pavilion was built on the top of the wall of the palace, using durable limestone.”为例，通过依存解析，得到关系xcomp(using, built)和nsubjpass(pavilion, built)，然后我们添加额外关系nsubjpass(pavilion, using), 这样做的原因是，作为动词或形容词的开放补语，“xcomp”通常会表达一层新的关系。最后，以上例句我们得到的关系三元组是：< The small rectangular pavilion; was built on top of; the wall of the palace >和<The small rectangular pavilion; using; durable limestone >。

**场景-4**：不同于场景-3，场景-4的目标是处理出现在VVP（带有开放补语成分的动词短语VP）中的补语成分。具体来说，如果原句S中包含了VVP，那么即便从S中提取到“xcomp”的关系，我们也不会生成像场景-3那样的两个关系三元组，这是因为VVP的两个动词共享同一主语且二者组合才能构成完整、准确的语义，因此我们只生成一个关系三元组，并将VVP中的两个动词放在一起作为关系三元组的关系动词。 例如，在“Salt is used to make glass, washing powder and paper”这句话中，VP1是“is used to make”，并且存在依赖关系xcomp(used, make)，但是它们共享主语“Salt”，并且二者组合才能表达完整准确的语义，所以我们生成的关系三元组为<Salt; is used to make; glass>、<Salt; is used to make; powder>和<Salt; is used to make; paper >。

**场景-5**：场景5的目标是处理句子中的并列关系。句子中表示并列关系的连词主要含有“和”、“补充”和“增加”等意思，句子中构成并列关系的成分意义同等重要，相互之间没有从属关系，是平行并列的关系。在本规则中，我们不但抽取出构成并列关系的句子成分，同时也按照句子成分性质的不同分成两种情况：名词短语之间的并列和动词短语之间的并列，并进一步生成相应的关系三元组。

1）名词短语间的并列关系：如果名词短语之间通过“and”构成并列关系（这种关系在依存解析结果中用“conj:and”表示，其中conj代表并列关系，冒号后的连词代表具体的并列关系类型），我们将提取这些带有不同主语或宾语的关系三元组。比如，在例句“This hotel offers breakfast and morning call service.”中，我们将会提取<This hotel; offers; breakfast>和<This hotel; offers; morning call service>两个关系三元组。

2）动词短语间的并列关系：如果我们从依存解析中提取类似conj:and(vm,u, vn,v)或conj:or(vm,u, vn,v)（m≠n）的关系，我们将抽取出带有相同名词短语和不同动词短语的关系三元组。比如，在例句“Bell, which is based in Los Angeles, makes and distributes electronic products.”中，“makes”是VP2也是v2,1，“distributes”是VP3也是v3,1 ,同时存在关系conj:and(makes, distributes)，那么我们会抽取<Bell; makes; electronic products>和<Bell; distributes; electronic products>作为关系三元组。

**场景-6**：场景6用于处理同位语关系。同位语指的是当两个指向同一事物的句子成分放在同等位置时，若其中一个句子成分是用于说明或解释另一个句子成分的，那么用于说明或解释作用的句子成分就叫做另一成分的同位语。本规则不但抽取到句子成分和其相应的同位语成分，同时也按照成分性质的不同分成两种情况：名词短语之间的同位语关系、名词短语与动词短语之间的同位语关系。另外，由于句子成分和其同位语之间往往不存在明显的关系短语连接，所以在本规则中，我们会添加额外动词（如“is”）来构成最终的关系三元组。

1）名词短语之间的同位语关系：如果我们通过依存解析提取到关系appos(ni,p, nj,q)（i≠j，appos表示同位语关系），我们将添加额外动词（比如“is”）来构成完整语义并生成一个关系三元组<NPi; “is”; NPj>。 例如，在“Bell, a telecommunication company, makes and distributes electronic products.”这句话中，“Bell”是NP1的n1,1，“company”为NP2的n2,3，并且二者存在同位语关系appos(Bell, company)，我们将生成关系三元组<Bell; “is”; a telecommunication company>。

2）名词短语和动词短语之间的同位语关系：如果我们通过依存解析提取到关系appos(ni,p, vk,r)和nmod(vk,r, nj,q)，我们也会添加额外动词“is”到原动词短语并生成一个关系三元组<NPi; “is” VPk; NPj >。 例如，在句子“Rome, the capital of Italy, is known For the rich history”中，NP1和n1,1都是“Rome”， “the capital of”是VP1和v1,2是“capital”，NP2和n2,1是“Italy”，并且存在依存关系appos(Rome, capital)和nmod(capital, Italy)， 我们将生成关系三元组<Rome; “is” the capital of; Italy>。

通过定义以上6个语言场景，对于例句“Huawei, a leading global provider of informational and communicational technology and infrastructure, which is located in Shenzhen, is committed to building an intelligent and fully-connected world.”，将生成如下关系三元组：

(a) <Huawei; “is” leading global provider of; informational technology>

(a) <Huawei; “is” leading global provider of; communicational technology>

(b) <Huawei; “is” leading global provider of; informational infrastructure>

(c) <Huawei; “is” leading global provider of; communicational infrastructure>

(d) <Huawei; is located in; Shenzhen>

(e) <Huawei; is committed to building; intelligent world>

(f) <Huawei; is committed to building; fully-connected world>