产品供应关系抽取进展

## 任务目标

从上市公司年报中找出该公司的客户和客户购买的产品。

## 信息来源

信息来自上市公司年报，具体来说分为两部分，一部分是纯文本信息，另一部分是格式化的信息。

由于格式化的信息抽取较为简单，目前主要研究是从纯文本信息中抽取供应关系。

## 整体流程图

整体处理流程如下图：

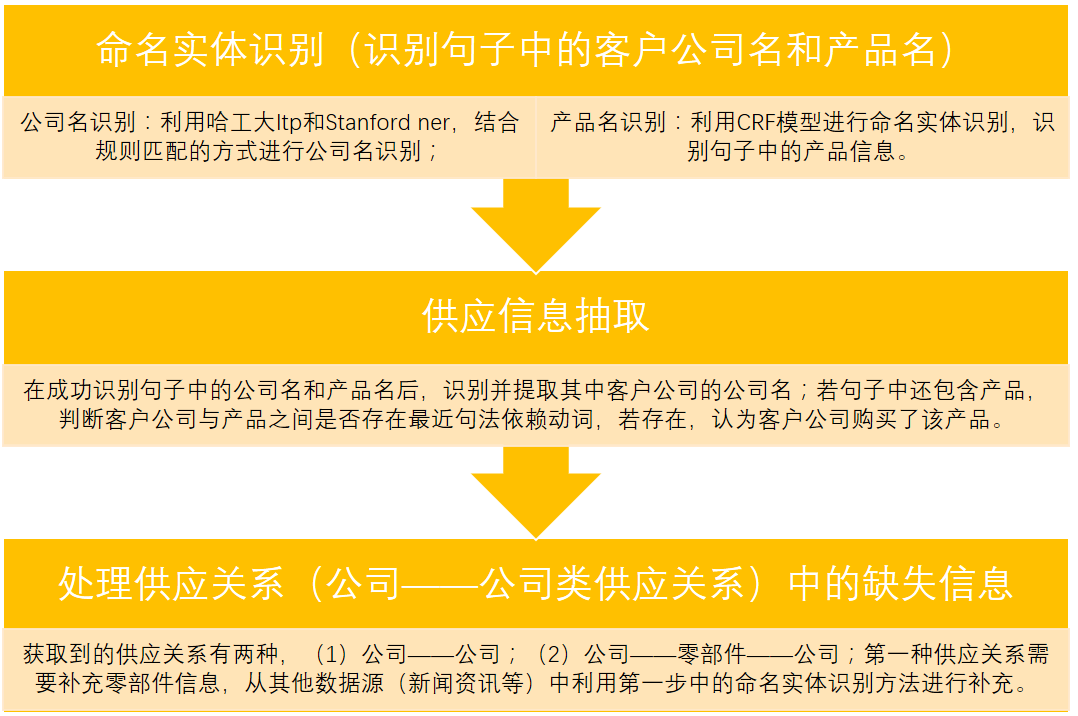


图1 整体流程图

目前已经完成了第一部分命名实体识别和第二部分实体关系抽取的任务。其中命名实体识别主要采用了机器学习的方法，第二部分实体关系抽取采用的是句法分析的方法，具体依赖的是哈工大的自然语言处理工具LTP。

## 目前处理流程

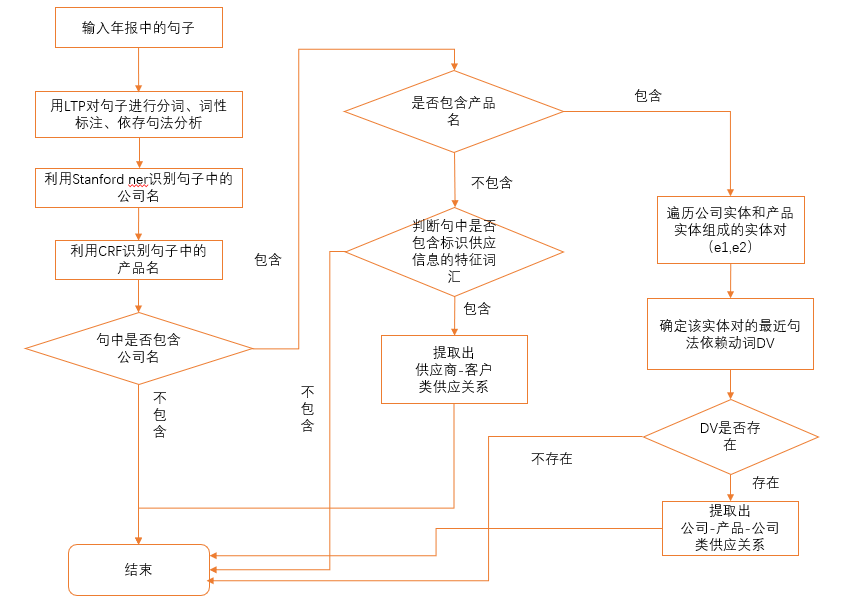


图2 整体运行流程图

## 过程描述及示例

### 步骤一：实体识别

1. 输入年报句子，利用LTP进行分词、词性标注、依存句法分析；

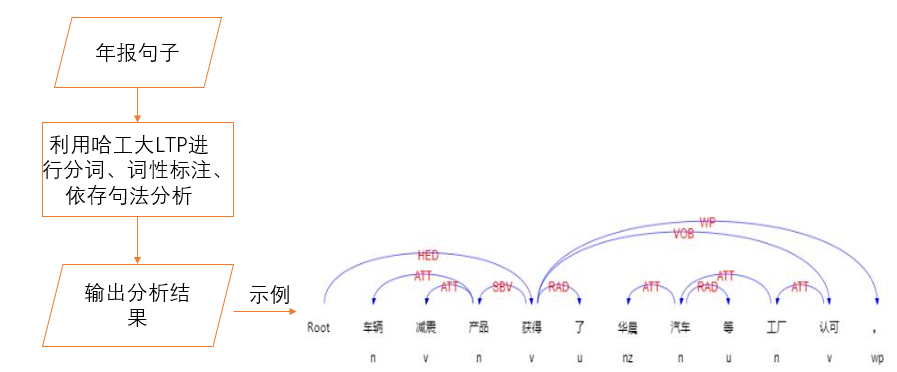


图3 LTP处理

1. 利用Stanford ner命名实体识别句子中的公司名；



图4 Stanford ner 识别公司名

1. 利用CRF进行产品名识别；



图5 CRF识别产品名

1. 根据LTP句法分析结果，利用规则对实体识别结果进行修正。

规则一：具备并列关系的词具有相同的实体类别，即如果一个词被识别为公司名/产品名，那么与该词并列的词也是公司名。

规则二：具备ATT（定中关系）的相邻词，如果其中一个词被识别为公司名/产品名，那么另外一个也是公司名/产品名。

### 步骤二：供应关系提取

#### 5.2.1 方法原理

通过步骤一筛选出内容中同时包含公司名和产品名的句子，基本上可以确定句子中包含供应关系，现在要确定哪个产品供给哪个公司，即判断公司和产品之间的语义关联。我采用的方法是判断公司名依赖的动词和产品名依赖的动词，如果这两个动词为同一个或者两个动词具备并列关系（COO），那么判断该公司和产品具备语义关联。

如果两个实体依赖的动词为同一个，那么它们具备直接语义关联，若两个动词具备并列关系，那么判断他们具备间接语义关联。

问题转换成寻找两个实体都依赖的动词，称之为最近句法依赖动词，能找到说明两个实体具备语义关联，否则不具备。

1. 直接语义关联实体间的最近句法依赖动词

如果实体对<ei,ej>中的2个实体能够通过一个最近句法依赖动词直接发生语义关联，则它们之间存在着一条满足如下图所示的依存句法路径．其中，节点表示实体或动词，有向边表示从动词节点指向实体节点，边上的内容表示实体节点与动词节点之间的依存关系。

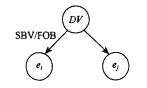


图6 实体对直接语义关联的依存句法路径

提取使2个实体发生直接语义关联的最近句法依赖动词的步骤为：

1)找出与第2个实体ej，直接发生依存关系的动词V，；

2）找出与第1个实体ei，直接发生SBV（主谓结构）或FOB（前置宾语）依存关系的动词V；

3)判断V，与V，是否为同一动词，若相同则实体对( ei，ej )的最近句法依赖动词DV为V，否则置为空值Null。

示例：

“华晨汽车购买了我公司的车辆减震产品。”的依存句法分析和词性标注如下图所示。

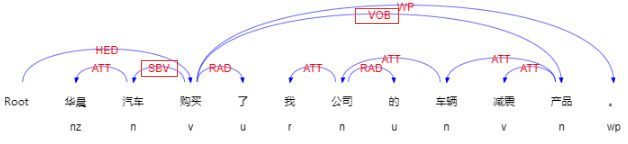


图7 句法分析结果

对于其中的实体对<华晨汽车，车辆减震产品>，第2个实体“车辆减震产品”是动词“购买”的宾语，与“购买”构成动宾关系，第1个实体“华晨汽车”是动词“购买”的主语，即SBV关系，因此，实体对<华晨汽车，车辆减震产品>的最近句法依赖动词为“购买”。

（2）间接语义关联实体间的最近句法依赖动词

如果实体对<ei,ej>中的2个实体能够通过一个最近句法依赖动词间接发生语义关联，则它们之间存在着一条如下图所示的依存句法路径，其中节点表示动词、实体或非实体类型的名词。图中的依存句法路径可以分为2个部分：实体部分和动词部分。

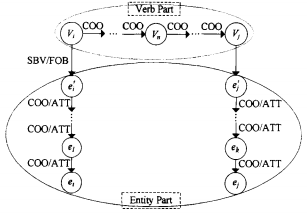


图8 实体对间接语义关联的依存句法路径

在判断间接语义关联时利用了性质：具备并列结构（COO）、定中结构（ATT）的组合词词具有相同的最近依赖动词。

**示例1：**

“车辆减震产品获得了华晨汽车的认可。”，句法分析如下：

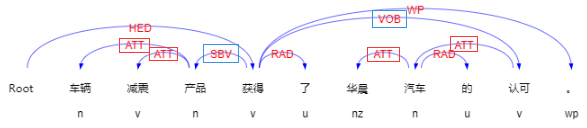


图9 句法分析图

如图9所示，实体对<车辆减震产品，华晨汽车>没有直接的动词关联，车辆减震产品与动词“获得“构成SBV（主谓关系），华晨汽车不直接与”获得“发生关联，与其具备ATT(定中结构)的词”认可“与”获得”构成VOB（动宾关系），故华晨汽车与车辆减震产品具有间接语义关联，最近依赖动词为“获得”。

**示例2：**

“公司主营产品天线和滤波器是基站建设的核心组件，主要客户包括华为和爱立信。“依赖句法分析如下：

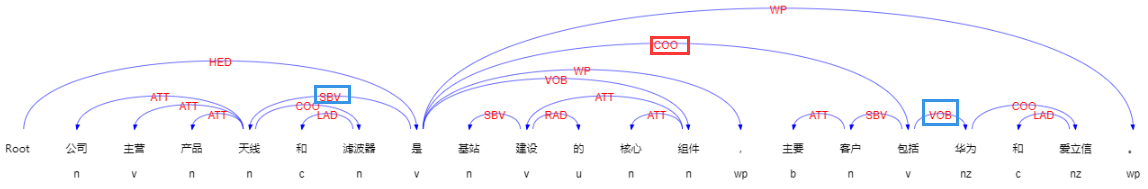


图10 句法分析图

现分析其中的实体对<天线、华为>，天线的依赖动词为“是“，与其构成主谓结构，华为的最近依赖动词为”包括“，与其构成动宾结构，而动词”是“与”包括“之间构成并列结构（COO），故实体天线与实体华为之间存在间接语义关联，最近依赖动词为”是“。

#### 5.2.2 具体实现

（1）算法步骤

步骤1．分别提取与实体ei或ej，存在COO并列结构或ATT定中结构关系的依存关联节点eei和eej;

步骤2．提取与第2个实体ej的依存关联节点eej发生依存关系的最近动词Vj;

步骤3．获取与第1个实体ei的依存关联节点eei发生SBV或FOB关系的最近动词Vi;

步骤4．通过判断动词Vi与Vj是否为同一个动词或为COO并列结构关系，确定该实体对(ei,ej>的最近句法依赖动词DV;

利用上述算法提取实体关系类型中的高频最近句法依赖动词信息。

（2）输入输出

输入：步骤一获得的分词、词性标注、句法分析、公司名产品名识别结果；

输出：经过判断具备语义关联的公司、产品对<ei,ej>