Computer Science and Application 计算机科学与应用, 2018, 8(12), 1823-1832
Published Online December 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.812202

Hans汉斯

Research on Automatic Extraction of Enterprise Supply Relationship Based on NLP

Chuanlong Yang, Jinlong Wang

School of Information and Control Engineering, Qingdao University of Technology, Qingdao Shandong,
Email: 1434296509@qq.com,qdwangjinlong@163.com

Received: Nov. 29th, 2018; accepted: Dec. 10th, 2018; published: Dec. 17th, 2018

Abstract

A good supply chain is indispensable for enterprises to improve competitiveness. For a supply chain, the most important part is the supply relationship between enterprises. Existing methods of extracting corporate entity relationship did not consider the role of corporate entity in the supply relationship. Therefore, these methods are not suitable for extracting enterprise supply relationship. To solve this problem, a library of relation word is constructed by combining manual construction with automatic construction. The relation word is used to judge the theme of the text, and the nearest syntactic dependent verbs are used to judge the semantic relationship between entities. Experiments on the annual report of listed companies have met the expected requirement.

Keywords

Supply Relationship, Entity-Relationship Recognition, Information Extraction

基于NLP的企业供应关系自动抽取研究

杨传龙，王金龙

青岛理工大学信息与控制工程学院，山东 青岛
Email: 1434296509@qq.com,qdwangjinlong@163.com

收稿日期：2018年11月29日；录用日期：2018年12月10日；发布日期：2018年12月17日

摘　要

供应链对企业竞争力具有巨大意义，而供应链中最重要的部分就是企业供应关系，现有的公司实体关系

文章引用: 杨传龙, 王金龙. 基于 NLP 的企业供应关系自动抽取研究[J]. 计算机科学与应用, 2018, 8(12): 1823-1832.
DOI: 10.12677/csa.2018.812202

抽取方法没有考虑关系中公司实体的角色，不适用于企业供应关系抽取。基于此，本文采用人工构建和自动构建相结合的方式构建了关系指示词库，利用关系指示词库判断文本的主题，并使用最近句法依赖动词对实体之间的语义关系进行判断。最后在上市公司年报文本上进行了测试，取得了不错的效果。

关键词

供应关系，实体关系识别，信息抽取

1. 引言

供应链管理作为企业运营的重要组成部分，能够帮助企业提升竞争力，降低成本，提高利润率，对于企业具有极为重要的作用。供应链分析的关键一环就是获取企业与企业之间的供应关系。

目前，为了服务于用户，已有一些公司推出了企业关系分析产品。财新网推出了收费服务“数据+”，提供财团企业关系图谱功能，能够展示投资和股权关系。同花顺推出了产品图谱，展现上市公司之间的上下游供求关系，帮助股票投资者进行投资分析。企业供应关系抽取日渐成为研究热点[1] [2] [3]。文献[1]基于上市公司公告抽取了持有、投资、转让、合并、收购五种关系；文献[2]将企业关系定义为合作、附属、股权、收购和建立五种，但其仅能抽取公司间的合作关系。文献[3]根据企业关系触发词不同将企业供应关系细分为客户关系、供应商关系等类别，所抽取关系仍然没有确定公司之间供应产品信息，也未区分公司在供应关系中的角色。例如，针对文本“华为已经悄悄跟京东方达成了合作，京东方今年至少要给华为供应 100 万块自主柔性 OLED 屏”，当前研究可以确定合作关系[2] (图 1(a))，客户关系[3] (图 1(b))，但无法确定两个公司在供应关系中的角色和所供应产品信息(图 1(c))。

本文针对现有企业供应关系抽取中缺少产品信息并且公司角色不明确的问题，采用实体关系抽取和依存句法分析等自然语言处理技术，通过识别产品类别词识别和判断供应关系判断，能够有效抽取企业供应关系，如图 1(c)所示，满足使用者分析的需要。

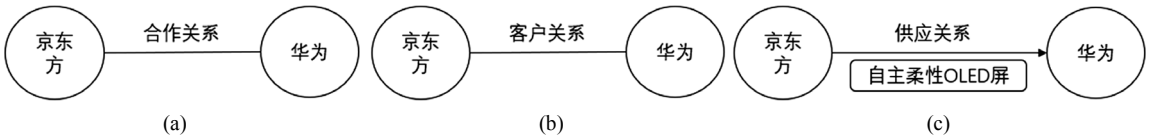

Figure 1. Comparisons of extraction effect of existing achievements
图 1. 现有成果抽取效果对比

2. 企业供应关系抽取

2.1. 问题定义

本文将某描述供应商 A 信息文本中的供应关系定义如下：

对于文本中任一自然语句，包含的公司名称集合为 $C=\{c_1,c_2,\cdots,c_n\}$，n 为句子中公司实体数量，产品集合为 $P=\{p_1,p_2,\cdots,p_m\}$，m 为句子中产品实体数量，一个供应关系可以定义为一个四元组 $\langle A,c_i,p_j;y\rangle$，其中 $i\in[1,n]$，$j\in[1,m]$，$A\neq c_i$，y 为该组合对应的标签 $y\in\{0,1\}$，0 表示 A、c_i 和 p_j 之间

存在供应关系，1 表示不具备供应关系。

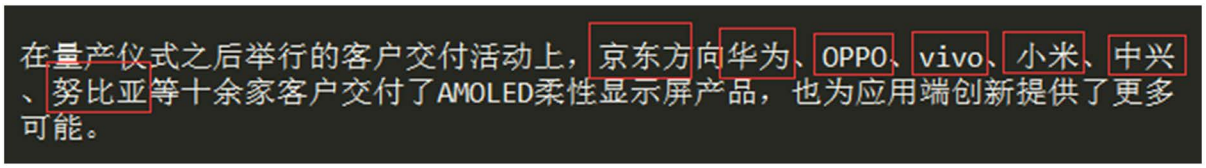

Figure 2. Text example
图 2. 文本示例图

以图 2 为例，该文本属于京东方资讯新闻，即目标供应商 A 为京东方，经过公司名称识别和产品名称识别，得到产品集合 $P=\{\text{AMOLED柔性显示屏}\}$，公司集合 $C=\{$京东方、华为、OPPO、vivo、小米、中兴、努比亚$\}$，文本中描述了京东方向华为、OPPO 和 vivo 等公司供应了 AMOLED 柔性显示膜产品，很明显，该文本中一共包含了 6 条供应关系，如下表 1 所示。本文的目标为从文本抽取出如下表所示的企业供应关系。

Table 1. Table of sample extraction results
表 1. 示例抽取结果表

供应商公司	客户公司	产品
京东方	华为	AMOLED 柔性显示屏
京东方	OPPO	AMOLED 柔性显示屏
京东方	vivo	AMOLED 柔性显示屏
京东方	小米	AMOLED 柔性显示屏
京东方	中兴	AMOLED 柔性显示屏
京东方	努比亚	AMOLED 柔性显示屏

2.2. 现有方法及问题

企业供应关系抽取属于实体关系抽取问题，主要有基于特征向量[4] [5] [6] [7] [8]，基于核函数[9]和基于模式匹配[10]的方法，也有研究者提出基于关系指示词库[11] [12]和基于依存句法分析[13]的方法。

基于特征向量的方法将关系抽取看作实体对分类问题进行处理，通过构造实体对结合上下文环境构建特征向量来判断两个实体之间是否存在预定义关系，忽略了关系中的细节，同时，需要大量预料标注，对于关系复杂问题难以进行处理。

基于模式匹配的方法需要领域专家参与，存在召回率低的问题，不适用于多个实体间关系的判断。树核函数的方法同样不适用于包含多个实体的关系抽取，并且由于汉语句法复杂，表达方式多样，不能直接应用到本研究中。现有的中文公司实体关系抽取方法中，候选实体组中实体的顺序通常按照实体在文本中出现的顺序进行排列，没有考虑到公司实体在关系中担任的角色。

3. 基于关系指示词库和句法分析的企业供应关系抽取方法

针对现有方法存在的问题，本文选择描述特定公司信息的主题文本，利用文本主题确定供应商信息。在判断供应商、客户和产品三个实体关系时，将过程拆解为两步，首先确定供应商与客户之间的关系，再将产品信息补充到结果中，从而得到完整的企业供应关系。

整体方案如图 3 所示，包括三个步骤：1) 语句筛选及实体识别，利用关系指示词库筛选与供应关系主题相关的句子，并进行相关的实体识别；2) 公司实体对生成，利用规则将文本描述的目标公司与句子

中的公司实体组成公司实体对，确定供应商和客户两个公司的信息；3) 利用句法分析确定产品信息，分析寻找最近句法依赖动词处理句子中出现多个公司或者多个产品时的情况，解决客户与产品之间的对应问题，最终确定一条完整的，包含供应商，客户和产品三个实体的企业供应关系。

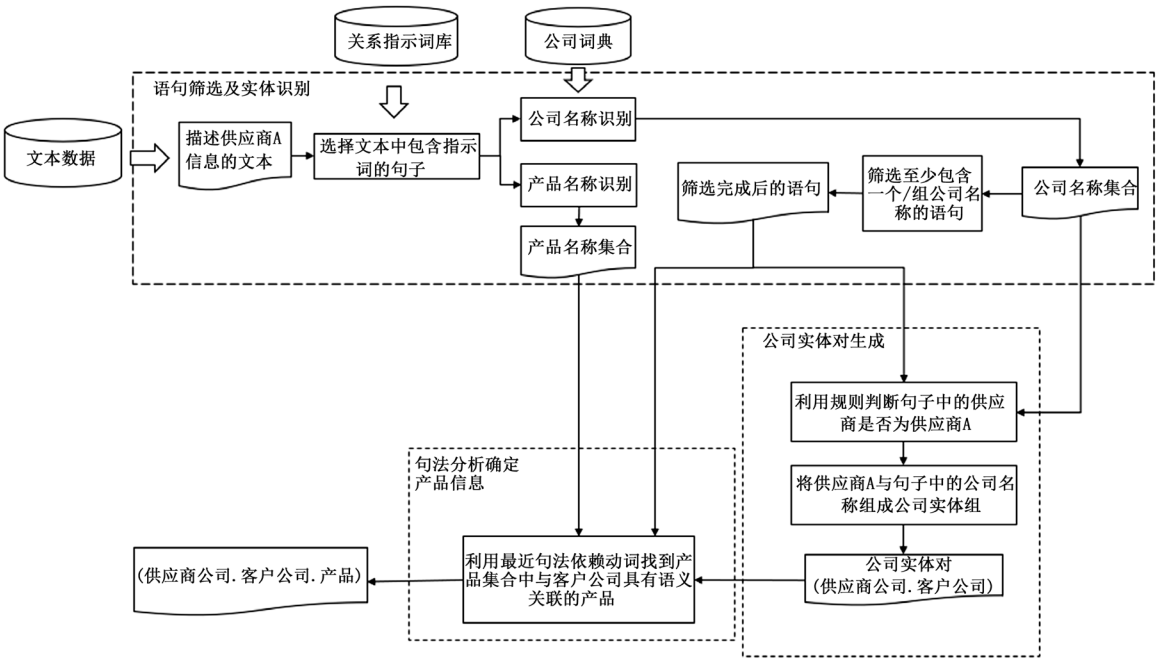

Figure 3. Flow chart of system integral processing
图 3. 系统整体处理流程图

3.1. 语句筛选及实体识别

语句筛选及实体识别阶段选择主题与企业供应关系相关并且满足包含供应关系最低要求的句子，进行公司名称和产品名称的识别。

1) 关系指示词库过滤无关句子。由于绝大部分存在关系的实例都能在文本中找到一个关系指示词来标识实体之间的关系[14]。因此，本文通过构建企业供应关系指示词进行实体关系类型判断，当句子中包含关键词库中的词时，认为句子主题与供应关系相关。

2) 公司名称识别。公司名称识别模块进行句子中公司名称识别，并在结果上进行标注。本文需要识别的主要是句子中的公司简称，在比较了现有工具后，选择了 Stanford NER 进行公司名称识别，并利用上市年报释义信息构建了公司词典辅助进行识别。

3) 产品名称识别。金融文本中的产品名称属于产品类别词，针对该特点，本文选择词特征，词性特征，边界词特征和词典特征训练了条件随机场模型进行句子中的产品名称识别。

4) 筛选出至少含有一个或一组公司名称的句子。由上文的定义可知，一条企业供应关系包括供应商公司、客户公司和产品三项信息，其中供应商和客户信息是必不可少的，即至少包含两个公司名称。由于本文抽取文本中供应商信息常常不显式出现，而是以代词代指，而客户名称必定以显式出现。因此本文把包含企业供应关系基本条件设定为至少包含一个或一组公司名称。

3.2. 公司实体对生成

公司实体对生成需要确定供应商和客户，并形成公司实体对。由于本文处理对象是描述特定公司信

息的文本，因此供应商是固定的。本阶段主要任务是判断句子是否描述该特定公司的供货信息，若是，则形成(供应商，客户)实体对。以图 2 中的句子为例，该句子所属文本描述公司为京东方，因此进行公司实体对提取时需要判断“京东方”与文本中其他各公司(华为、OPPO、vivo、小米、中兴、努比亚)是否具备供应关系。

具体上，基于以下规则判断：

1) 句子中显式或隐式出现目标公司信息。句子中描述目标公司的供应关系时，公司名称需要显式或隐式出现，如“公司”，“本公司”等等。以图 4 文本为例，在描述京东方供应关系文本中，出现了代词“公司”和公司名称“京东方”。

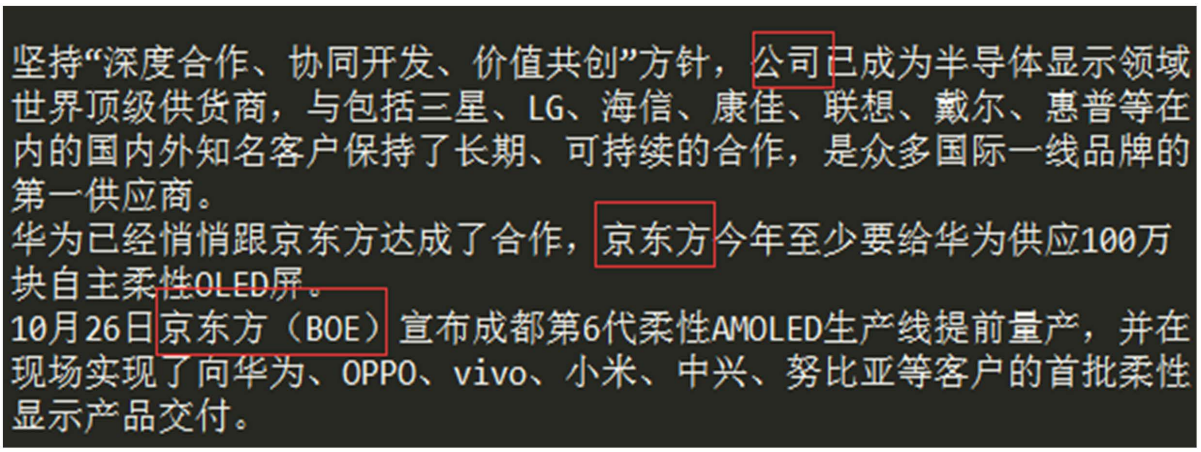

Figure 4. Examples of the appearance of the target company in the text
图 4. 文本中目标公司出现形式示例

2) 目标公司的信息在句子中作为独立成分存在，不能与其他公司名称存在并列关系。本文寻找的是目标公司的供应关系，即目标公司的供货关系，而供货关系是一对多的关系，因此文本中目标公司都是独立出现，如上图 4 中，京东方需要作为独立成分出现，而客户公司名称常常会出现多个并列的情况。

公司实体对抽取如下图 5 所示，首先确定句子来源的文档描述是哪家公司的信息，然后判断句子中是否显示或隐式的带有文档主题公司的信息，若携带目标公司的信息且该信息作为独立成分存在，说明该句子中的供应商就是文档描述的目标公司，将目标公司和句子中已有的公司名称构成公司实体对。

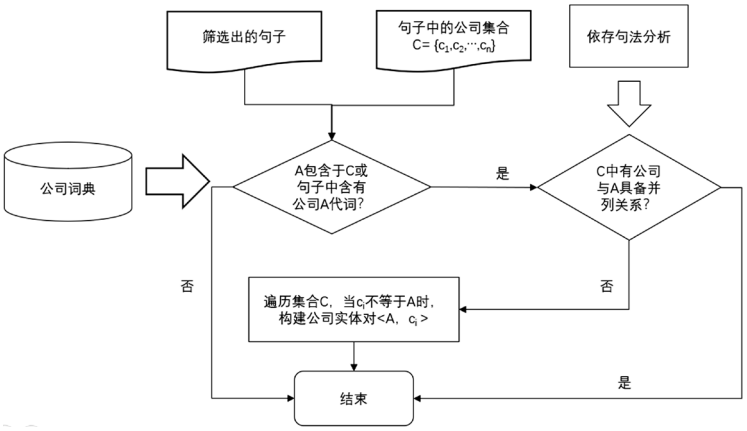

Figure 5. Corporate entity pair extraction flow chart
图 5. 公司实体对抽取整体流程图

图 2 中的示例文本为例，该句子来自京东方的企业新闻，可以从中构建(京东方，华为)、(京东方，OPPO)、(京东方，vivo)、(京东方，小米)、(京东方，中兴)和(京东方，努比亚)六组公司实体对。

3.3. 句法分析确定产品信息

由于供应关系实际上为一种供货关系，通常一个供应商对应多个客户信息，而句子中也常常包含多个产品，需要解决哪个产品供应给了哪个客户的问题，即客户与产品之间的对应问题。为此，引入最近句法依赖动词判断法[15]对客户和产品间的关系进行准确判断。

最近依赖动词可以判断实体之间的语义关联，本文通过抽取客户和产品之间最近依存动词来确定客户和产品之间是否存在对应关系，主要用于处理文本中包含多个产品名称或多个客户公司名称的情况。由于具有并列关系的实体在句子中语法角色相同，因此在判断产品信息时，将多个并列的公司或者产品实体作为整体处理。具体可以分为以下两种情况：

1) 句子中包含多组产品名称。以下图 6 为例，该句子中包含了多个产品名称{锂离子电池，六氟磷酸锂产品，电解液}和一组公司名称{比亚迪，杉杉股份，新宙邦}，需要判断供应商具体给客户供应了哪种产品。

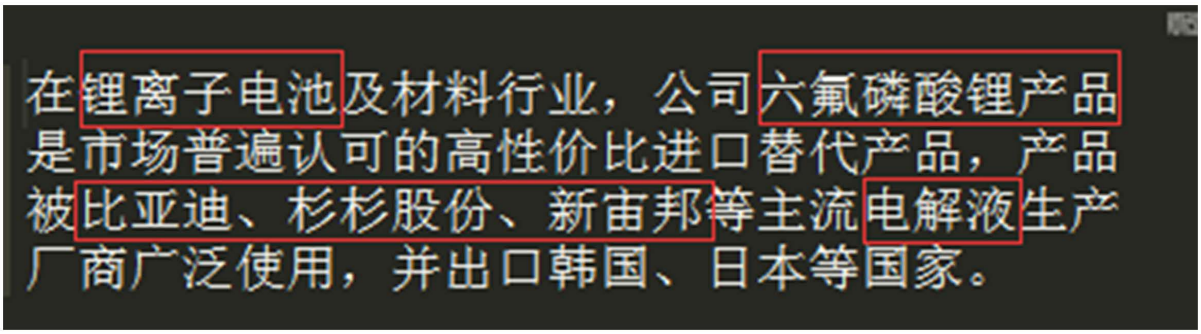

Figure 6. Multiproduct text sample
图 6. 多产品文本示例图

2) 句子中包含多组客户公司名称。例如下图 7 中的文本中包含了两组客户公司名称和一个产品名称“激光焊接设备”，需要判断该产品具体供应给了哪组公司。

Figure 7. Multiple groups of company names sample
图 7. 多组公司名称文本示例图

以图 6 为例对本方法进行说明，该文本来自多氟多 2017 年年报，其中客户公司集合 C = {比亚迪，杉杉股份，新宙邦}，产品集合为 P = {锂离子电池，六氟磷酸锂产品，电解液}，抽取三个公司实体对(多氟多，比亚迪)，(多氟多，杉杉股份)，(多氟多，新宙邦)。

现在需要确定每个公司实体对中的客户与产品之间对应关系，并将与客户存在对应关系的产品加入到公司实体对中去，形成完整的供应关系。利用最近句法依赖动词进行客户与产品之间对应关系的判断过程如下。

对句子进行依存句法分析，部分分析结果如下图 8 所示。

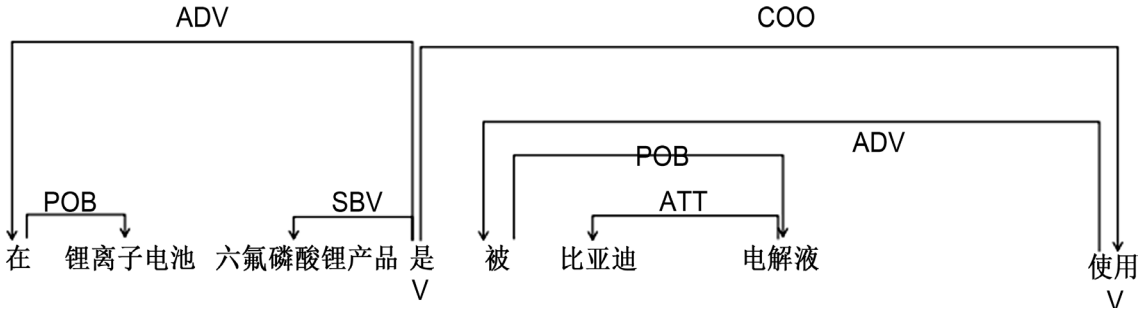

Figure 8. The analysis results of syntactic dependence of example
图 8. 示例部分句法依赖分析结果

由于比亚迪，杉杉股份和新宙邦在句子中具有并列关系，属于同一个实体组，按照最近依赖句法依赖动词判断方法，这三个词对应的最近动词相同，因此，仅对(多氟多，比亚迪)这一个公司实体对进行处理，查找其对应的产品，其它两个实体对(多氟多，杉杉股份)和(多氟多，新宙邦)的处理方式和结果相同，这里不再赘述。

现寻找客户比亚迪与三个产品锂离子电池，六氟磷酸锂产品，电解液之间的最近句法依赖动词，从而确定哪个产品与公司比亚迪对应。首先确定要寻找最近句法依赖动词的实体对<e_i, e_j>，一共包含三个(锂离子电池，比亚迪)、(六氟磷酸锂产品，比亚迪)和(电解液，比亚迪)，其中实体顺序由实体在句中出现的位置确定，寻找最近句法依赖动词过程中各节点统计结果如下表 2 所示。

Table 2. The recent dependency verb statistics table
表 2. 最近依赖动词统计表

实体对<e_i, e_j>	e_i 的依存节点	e_j 的依存节点	e_i 的最近动词	e_j 的最近动词	最近句法依赖动词
(锂离子电池，比亚迪)	锂离子电池	电解液	无	使用	无
(六氟磷酸锂产品，比亚迪)	六氟磷酸锂产品	电解液	是	使用	有
(电解液，比亚迪)	电解液	电解液	无	使用	无

由上表可知，只有六氟磷酸锂产品与比亚迪之间存在最近依赖动词，其依存路径如下图 9 所示。因此，将六氟磷酸锂产品加入到供应关系中，获得供应关系为(多氟多，比亚迪，六氟磷酸锂产品)。同理，其余两个公司实体组处理方式相同，最终获得(多氟多，比亚迪，六氟磷酸锂产品)，(多氟多，杉杉股份，六氟磷酸锂产品)和(多氟多，新宙邦，六氟磷酸锂产品)三条企业供应关系。

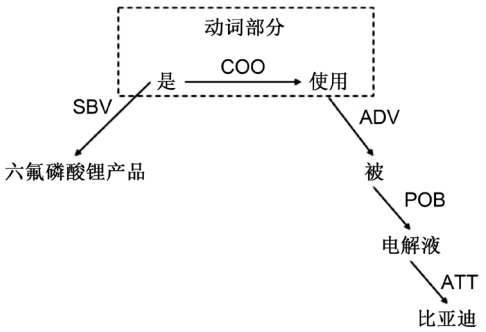

Figure 9. Entity dependency path map
图 9. 实体依存路径图

句子中包含多组公司名称时，处理方式相同，同样是判断客户与产品之间是否存在最近句法依赖动词。当句子中出现多个产品名称或多个客户公司时，对于每一个公司实体对，本文补全产品名称的整体处理流程图如下图 10 所示。

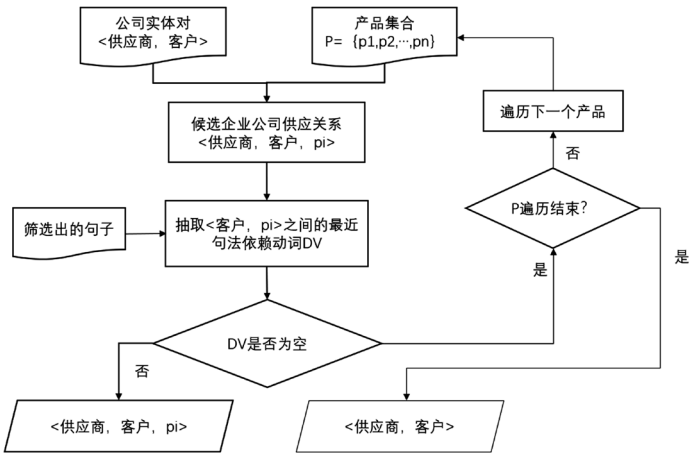

Figure 10. Processing flow chart of semantic relation judgment using the nearest syntactic dependent verbs
图 10. 利用最近句法依赖动词判断语义关系处理流程图

当句子中只出现一组产品名和一组公司名时，本文默认该产品名为供应关系中对应的产品或者产品组，不再进行额外处理，直接加入到供应关系中去。

3.4. 关系指示词库构建

构建关系指示词库主要基于两点考虑：1) 关系指示词的词频，上下文中某个关系指示词的词频越高，说明该词是一种常用词，越有可能作为实体对之间正确关系的表述；2) 实体与关系指示词之间的距离，具体值的是关系指示词和候选实体对之间在文本中间隔的字词数量，它们之间的距离越接近，说明关系指示词与实体对之间的联系越紧密。

基于以上的考虑，关系关键词库采用人工构建和自动构建相结合的方式。首先通过对上市公司年报语料的观察，筛选出 9 个关系关键词如表 3 所示。

Table 3. Examples of initial keywords
表 3. 初始关键词示例

编号	词
1	供应商
2	业务
3	合作
4	客户
5	中标
6	销往
7	签订
8	签署
9	承揽

关键词库的扩展借鉴了拔靴法的思想，通过抽取两个句子间的公共词汇来进行词库的扩展，同时考虑了关键词的出现频率，兼顾了关键词提取的覆盖率和准确率，扩展流程如表 4 所示。

Table 4. Extension method of relation word
表 4. 关系指示词库扩展方法

输入：包含现有关系指示词库中关键词的语句集合，
输出：经过扩充以后的关系指示词库
过程：
For each 语句集合 set
2. For each 关键词 k ∈ 关键词词库
3. For each 两个包含了同一个 k 且包含两个以上公司名称的不同语句 S1, S2
4. 寻找两个句子中的公共词组
5. 进行去停用词操作
将去停用词以后的关键词词组保存进备选词库，并计算出现频率
7. For each 备选词 word
8. IF word 出现频率大于 0.3 THEN
9. 把该词添加到关系指示词库
10. END IF
1 End for
12. End for
13. End for
14. End for

4. 实验结果以及结果分析

4.1. 实验数据

测试语料主要来自各上市制造业公司 2017 年年报，共 1483 篇，本文随机选取了其中 100 篇年报作为测试数据。

4.2. 实验评价以及指标

为了验证本文方法有效性，采用准确率(P)、召回率(R)和 F 值三个值作为效果评价的标准，计算公式如(1)、(2)、(3)：

$$P=\frac{\text{识别出正确的供应关系个数}}{\text{识别出的所有供应关系个数}}\times 100 \tag{1}$$

$$R=\frac{\text{识别正确的供应关系个数}}{\text{文本中所有的供应关系的个数}}\times 100\% \tag{2}$$

$$F\text{值}=\frac{2\times P\times R}{P+R}\times 100\% \tag{3}$$

4.3. 实验结果以及对比分析

为了更准确测试方法有效性，当出现多个公司名或产品名并列的情况时，将产品实体组和公司实体组作为整体进行计数，这是因为同一个实体组中的各实体在句中句法角色相同，依赖的都是同一条依存路径，重复对其计数反而会影响对方法效果的判断，例如 3.3 中的示例抽取出的三条供应关系在本实验中只计算为一条，即把比亚迪、杉杉股份、新宙邦作为一个整体进行统计。

本文通过对年报文本的分析，发现年报文本的内容有着严格的规定，章节分布和主要内容是固定的，并且包含公司名称的章节主要集中在第三节公司业务概要和第四节经营情况讨论与分析这两部分中，因此本文只抽取年报中的第三节和第四节进行处理。本文筛选出的 100 篇年报中共含企业供应关系 312 条，结果如表 5 所示。

通过实验数据分析可以看到，本文的供应关系抽取方法在上市公司年报文本上取得了不错的效果，对企业供应关系的抽取分别达到了 83.6%的 F 值，基本上达到的预期效果，证明了利用依存句法分析和关系指示词库提取文本中企业供应关系的有效性。

Table 5. Experimental results table
表 5. 实验结果表

Precision (%)	Recall (%)	F-measure (%)
79.3%	88.5%	83.6%

5. 总结与展望

本文提出的方法有效的提取了主题文本中的企业供应关系，但是在识别过程中也出现了一些识别错误的问题，具体原因有如下几点：1) 利用句法最近依赖动词判断实体之前的语义关系不总是可靠，存在判断错误的问题；2) 利用关系指示词库进行企业供应关系文本的筛选粒度太粗，出现部分错误分类的问题；3) 发现极少数包含供应关系的文本，其中不包含目标公司的信息，而是隐含在上下文中，导致抽取失败。接下来的工作将针对上述问题进行进一步的研究，以提升实体关系的抽取效果。

基金项目

本研究由国家自然科学基金资助项目(61502262)提供支持。

参考文献

[1] 孙晨, 付英男, 程文亮, 等. 面向企业知识图谱构建的中文实体关系抽取[J]. 华东师范大学学报(自然科学版), 2018(3): 55-66.

[2] 孟蕾. 基于网络数据的中文公司实体关系抽取研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2018.

[3] 代江波, 毛建华, 刘学锋, 张鸿洋. 基于特征向量与 SVO 扩展的企业生态关系抽取[J]. 计算机技术与发展, 2018(10): 1-6.

[4] 车万翔, 刘挺, 李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2): 1-6.

[5] 程文亮. 中文企业知识图谱构建与分析[D]: [硕士学位论文]. 华东师范大学, 2016.

[6] 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报, 2014, 28(6): 183-189.

[7] Guodong, Z., Jian, S., Jie, Z., *et al.* (2005) Exploring Various Knowledge in Relation Extraction. *ACL* 2005, *Meeting of the Association for Computational Linguistics*, University of Michigan, Michigan, 25-30 June 2005, 419--444. https://doi.org/10.3115/1219840.1219893

[8] 朱艳辉, 李飞, 胡骏飞, 等. 基于三支决策的两阶段实体关系抽取研究[J]. 计算机工程与应用, 2018, 54(9): 145-150.

[9] 陈鹏, 郭剑毅, 余正涛, 严馨, 张志坤, 高盛祥. 融合领域知识短语树核函数的中文领域实体关系抽取[J]. 南京大学学报(自然科学), 2015, 51(1): 181-186.

[10] 秦兵, 刘安安, 刘挺. 无指导的中文开放式实体关系抽取[J]. 计算机研究与发展, 2015, 52(5): 1029-1035.

[11] 王树伟. 面向金融文本的实体识别与关系抽取研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2014.

[12] 王月, 周刚, 南煜, 等. 基于关系指示词库的开放式实体关系抽取算法[J]. 信息工程大学学报, 2017, 18(2): 242-247, 252.

[13] 李颖, 郝晓燕, 王勇. 中文开放式多元实体关系抽取[J]. 计算机科学, 2017, 44(z1): 80-83.

[14] 刘安安. 开放式中文实体关系抽取研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2013.

[15] 甘丽新, 万常选, 刘德喜, 等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展, 2016, 53(2): 284-302.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org