融入搜索引擎的启发式开放域三元组抽取

刘勇杰　姜天文　秦 兵　刘 铭　刘 挺
社会计算与信息检索研究中心，哈尔滨工业大学
{yongjieliu,twjiang,qinb,mliu,tliu}@ir.hit.edu.cn

摘要：面向开放域知识图谱构建及其补全，本文通过分析实体关系在句法依存树中的表述方式，提出了一种融入了搜索引擎的启发式实体关系三元组抽取方法。该方法首先利用句法依存工具辅以启发式规则，从新闻语料中抽取出大量的候选实体关系三元组，再利用搜索引擎返回结果计算三元组的置信度，过滤置信度较低的实体关系三元组后获得最终的结果。实验结果显示本文的方法简单而有效，获得了 80% 以上的 F1 值，并且准确率高达到 89%，满足实际要求，可以用于知识图谱的构建以及补全。

关键词：中文知识图谱，关系抽取，开放域三元组，搜索引擎

1　引言

一直以来，人工智能都是计算机科学家关心的问题，使计算机具有一定的智能的重要一步就是获取知识并将其存储，储备这些知识的仓库称其为知识库。2012 年 Google 提出知识图谱的概念，它是一种对知识库网络化的表征，网络中每个节点代表实体，而每条连边则代表实体间的关系，知识往往可以用三元组来表示，即头实体、关系指示词，以及尾实体三部分，这种三元组表征关系的形式最具代表性的就是万维网联盟发布的资源描述框架技术标准（简称 RDF） [10]。作为知识图谱构建的关键步骤，实体关系的抽取成为知识图谱相关研究工作的重点。本文主要研究如何从非结构化的文本中自动抽取实体关系，以用于知识图谱的构建以及补全。

早期的实体关系抽取任务，来自于美国国家标准与技术研究院组织开展的自动内容抽取 (简称 ACE) 测评会议，基于此所展开的研究工作可以大致分为四类：基于模式匹配 [5]、基于特征 [6]、基于神经网络 [13]，以及基于核函数 [7] [14] 的方法。其中后三种的方法把关系抽取看作分类问题，所抽取的关系属于 ACE 定义的六大类十八子类关系，这种传统关系用于扩充知识图谱仍显得不够具体，人们无法从中获得确切的语义信息。基于模式匹配 [5] 的方法抽取的关系有极高的准确率，但这种方法需要领域专家设计专门的模板，领域性较强，不能很好的移植到其他领域，局限性较为明显。另外，由于难以建立完整而准确的模式集合，基于模式识别的方法很难取得理想的召回率。模式扩展 [17] 的方法可在一定程度上缓解基于模式匹配方法的不足，其思想是定义少量的关系模式种子集合，通过自学习的方式扩充种子集合。

基于特征、基于神经网络、基于核函数的方法将关系抽取当做分类问题来做。这类方法能取得较不错的准确率。如刘克彬等人的工作 [14]，基于核函数的中文实体关系自动抽取系统，应用改进的语义序列核函数，结合 KNN 机器学习算法构造分类器来分类并标注关系的类型，以 ACE 评测定义的实体关系类型为标准，关系抽取的平均精度可以达到 88%，明显高于基于特征向量和传统的基于序列核函数的方法。但这类方法抽取的关系属于 ACE 定义的六

大类、十八子类关系，这些传统关系对于知识图谱的构建仍显得不够细腻，例如对于“中共中央”与“习近平”两个实体，这两个实体间的关系是“受雇于”，但是对于知识图谱补全来说，用更细粒度的关系指示词“总书记”去代替“受雇于”会更好。

伴随着互联网技术的发展，其数据类型更加多样化，关系抽取受到语义单元类型的限定以及关系类型的限制，难以与当下网络数据快速，多样化的增长趋势相适应。为了弥补限定域关系抽取的上述不足。Banko [1] 提出开放域三元组抽取这一任务，其任务是从大规模文本中抽取实体关系三元组，即头实体、关系指示词，以及尾实体三部分，实现不受领域限制的关系抽取。开放域三元组抽取系统更加注重文本实体对之间的语义表达, 而不再强调类别关系。

学者们大多通过开发完整的关系抽取系统来获得文本中的三元组，例如美国 Washington 大学人工智能实验室开发的 TextRunner [1]，ReVerb [4]，OLLIE [12]，Carnegie Mellon 大学实现的 NELL [2]，德国 Max Planck 研究中心的 PATTY [11] 等。这些系统中有的需要自动构造训练语料，从标注语料中提取关系模版或训练分类器，之后再抽取三元组，有的则根据语法特征直接从句法分析结果中抽取三元组。

[16] 对于开放域实体关系三元组抽取提出了一种无指导的抽取方法。首先利用自然语言处理工具标示出抽取文本中的命名实体与所有词的词性，每两个命名实体之间组合成候选实体关系三元组，关系指示词为两个命名实体之间所有的名词与动词。显然，候选实体关系三元组中必然有大量的噪音，为过滤掉噪音，其为每个三元组的关系指示词计算先验概率，将按照先验概率排序后的前 N 个关系词作为全局词表保存，之后针对每一类实体关系三元组（共五类），分别计算关系指示词对该类三元组类型的先验概率词表，最后将关系指示词不在词表的三元组去除，再辅以简单的句法规则用于过滤一些错误的三元组，得到最后的实体关系三元组。

本文从中文句子中词语之间的依存特点入手，对实体关系在语句中的表述方式进行分析，提出了一种面向开放域知识图谱构建的实体关系三元组抽取方法。该方法首先利用句法依存工具辅以启发式规则，从新闻语料中抽取出大量的候选实体关系三元组，再利用搜索引擎为每个候选实体关系三元组计算置信度，过滤掉置信度较低的三元组，获得最终的实体关系三元组。实验结果显示本文的方法简单而有效，准确率与 F1 值都达到了 80% 以上，可以应用于知识图谱构建以及知识补全的任务上。

2 方法

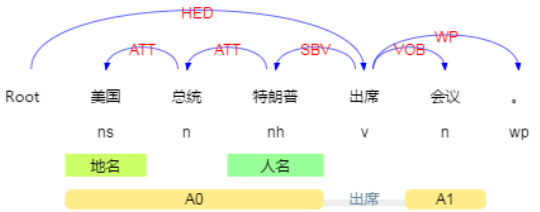
针对中文句子中词语之间的依存关系特点，首先对语句中实体关系表述的方式进行分析，获取了四类具有一定泛化能力的启发式规则。接下来，利用句法依存关系分析工具辅以启发式规则，提出了一种简单可行的开放域候选三元组抽取方法，获得一些较为准确的候选实体关系三元组，但在抽取过程不可避免地会抽取到一些错误的三元组。考虑一个人在判断一个事件的真假时，他很可能会将事件输入搜索引擎，通过搜索引擎返回的信息来判断事件的真伪。基于此，本文也将抽取到的候选关系三元组放入搜索引擎中，为每一个候选三元组获取置信度，基于置信度筛选掉错误的和不常见的关系三元组，得到最后的结果。

2.1 基于启发式规则模版的开放域候选三元组抽取

2.1.1 启发式规则模板的获取

依存句法分析是指通过分析语言单位内成分之间的依存关系揭示其句法结构的一种自然语言句子级分析方法。直观来讲，依存句法分析识别句子中的“主谓宾”、“定状补”等语法成

分，并分析各成分之间的关系。通过对大量的中文语句实例的句法依存分析树的观察后，发现实体关系的表述在句法依存树中具有一定的规律模式，并且这种模式与领域是无关的并独立于具体的关系类型，从而可以认为其具有较强的泛化能力，对这种规律进行抽象可以得到用于开放域实体关系三元组抽取的启发式模板。本文一共总结了四类启发式模板，用于从文本中抽取开放域候选三元组。



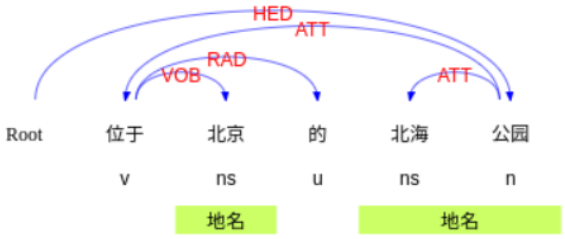


图 1: “美国总统特朗普……”的句法依存分析 图 2: “位于北京的北海公园”的句法依存分析

对于句子“美国总统特朗普出席会议”，利用语言技术平台 (LTP) [3] 提供的句法依存工具，分析结果如图1所示，可以发现此句中至少有一个明显的开放域实体关系三元组：（美国，总统，特朗普）。LTP 会将机构、人名、地名这三类命名实体标识出来，这对于抽取其中的三元组具有很大的帮助。通过观察可以发现，对于三元组（美国，总统，特朗普），三个元素在句子中依次通过 ATT（定中关系）链接，而且第一个和第三个元素被明确表明为命名实体（人名、地名、机构名），且第二个元素词性为名词，在本文中称这一类开放域实体关系三元组为“定中关系三元组”。以此为启发，构造一个基于启发式规则的“定中关系三元组”模板，模板满足以下规则：

- 头实体与尾实体皆为命名实体；
- 中间指示词必须为名词；
- 命名实体与中间指示词均为 ATT 依存关系，且依次指向前者。

对于句子“位于北京的北海公园”，其句法依存分析结果如图2所示。句中存在如下实体关系三元组：（北海公园，位于，北京），谓语元素（“位于”）是第三个元素（“北京”）VOB 关系（动宾关系）的父节点，而成为第一个元素（“北海公园”）ATT 关系（定中关系）的子节点，北海公园作为主语放到后面去了，这种结构有一个很明显的特征就是一定有一个“的”字与谓词构成 RAD 关系（右附加关系）。在本文中称这一类开放域实体关系三元组为“定语后置关系三元组”。以此为启发，构造一个基于启发式规则的“定语后置关系三元组”模板，模板满足以下规则：

- 指示词为动词；
- 指示词同时拥有 ATT 关系（定中关系）、VOB 关系（动宾关系）、RAD 关系（右附加关系）；
- 指示词的 ATT 关系父节点和 VOB 关系子节点为命名实体。

对于句子“青海地处青藏高原”，其句法依存分析结果如图3所示。句子中存在如下关系三元组：（青海，地处，青藏高原）通过观察可以发现该三元组以词性为动词的“地处”为中心，链接其他两个元素的依存关系都是以“地处”为父节点，关系依次为 SBV（主谓关系）、

VOB（动宾关系），在本文中称这一类开放域实体关系三元组为“主谓宾关系三元组”。以此为启发，构造一个基于启发式规则的“主谓宾关系三元组”模板，模板满足以下规则：

- 指示词为动词；
- 指示词的 SBV 和 VOB 依存关系指向两个命名实体。

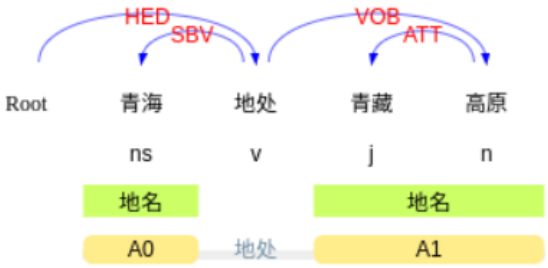


图 3: “青海地处青藏高原”的句法依存分析

图 4: “东京坐落于日本”的句法依存分析

对于句子“东京坐落于日本”，其句法依存分析结果如图4所示。句中存在如下实体关系三元组：（东京，坐落于，日本）。该三元组在句中的结构与“普通主谓宾关系三元组”很类似，但是这里的谓语不是简单的动词，而是动补短语，动词与补语之间通过 CMP 关系（动补关系）链接，动补短语与主语的关系依旧是 SBV（主谓关系），但与宾语（“日本”）为 POB（介宾关系）。在本文中我们称这一类开放域实体关系三元组为“主谓动补关系三元组”。以此为启发，构造一个基于启发式规则的“主谓动补关系三元组”模板，模板满足以下规则：

- 指示词为动词；
- 指示词拥有 CMP（动补关系），并且补语还拥有 POB（介宾关系）；
- 指示词的 SBV 关系指向命名实体，补语的 POB 关系子节点为命名实体。

2.1.2 候选开放域关系三元组的抽取

至此，基于不同的依存句法结构，构造了四类启发式规则模板，它们分别对应“定中关系三元组”、“主谓宾关系三元组”、“定语后置关系三元组”、“主谓动补关系三元组”。对于给定的中文自然语言文本语料，首先利用 LTP 进行句法依存分析，从而得到句法依存树，然后利用已经构造的启发式规则模板进行候选开放域关系三元组的抽取，可以依次获得“定中关系三元组”、“主谓宾关系三元组”、“定语后置关系三元组”、“主谓动补关系三元组”等四类不同依存结构的候选开放域关系三元组。

很显然，由于自然语言底层工具的具有一定的错误率，以及使用启发式规则模板也可能抽取到错误的三元组，因此对所抽取的候选三元组仍需进一步筛选，以获得较高准确率的三元组。

2.2 基于搜索引擎的开放域三元组置信度计算

根据 LTP 依存句法及其词性标注抽取出来的三元组包含一定的错误，例如，对于句子“中国驻美国记者发来报道。”（如图5），使用启发式规则模板“主谓宾关系三元组”，会抽取到（中国，驻，美国）这样没有价值的三元组，这类错误受限于构造的启发式规则模板，文本中

必然会有一些句子满足模板的要求但却并不是抽取正确的三元组。为了去除掉这样的三元组，可为每一个抽取出来的三元组计算一个分数，即置信度，作为这个三元组是否正确的评判，置信度低的三元组将被认为是错误的而被去除掉，而置信度高的三元组将被保留。

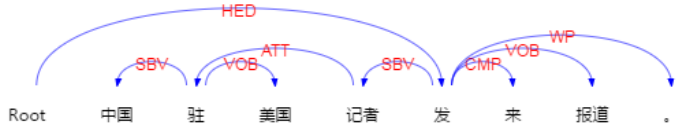


图 5: “中国驻美国……”的句法依存分析

2.2.1 候选三元组正负例对比分析

置信度对于正确抽取的三元组与未正确抽取的三元组应有较大的区分度，因此置信度的计算方法，应从正确抽取的三元组与未正确抽取的三元组的对比分析入手。例如，对于候选三元组（中国，首都，北京），这是一个正确抽取的三元组，不看中间的指示词“首都”，即对于头实体“中国”与尾实体“北京”，它们之间的可能的关系指示词有如“城市”、“一线城市”这样的词，这些词往往和“首都”具有非常相似的语义。但是对于如（中国，驻，美国）的负例候选三元组，通过一定背景知识可以得知，实体“中国”和“美国”的关系指示词还可能是“建交”、“合作伙伴”、“依赖”等这样的词，虽然其中有些词组成的三元组仍是没有太大价值的，但这样的词，往往和“驻”在语义上是不相近的。启发于此，可以把三元组关系指示词与其可能的关系指示词之间的语义相似度作为它的的置信度，而一个三元组可能的关系指示词，我们可借助搜索引擎获取。

2.2.2 基于搜索引擎的候选关系指示词获取

互联网是人们获取信息的重要途径，一个三元组的候选关系指示词也可从互联网中获取。本文按照如下方法为每一个抽取出的三元组获取候选关系指示词：首先把三元组中间的关系指示词去掉，对应（中国，驻，美国），将头实体“中国”和尾实体“美国”放到搜索引擎中进行搜索。搜索引擎会返回大量的句子，其中大部分的句子并没有包含头实体与尾实体，这种句子显然没有太大的价值，而包含了头实体或尾实体的句子，其同时包含关系指示词的可能性很大，因此将这样的句子进行分词，得到的词语作为候选关系指示词保留，以用作之后的相似度计算。

同一个三元组的不同候选关系指示词中，出现频率较高的词很可能与这个三元组的关系指示词相似度非常高。但是一些常见词，例如“的”、“我”出现的频率也会很高，它们与关系指示词的相似度肯定是非常低的，因此这些词不应该作为候选关系词保留。要去除这样的词，通常可以设置一个停用词词表，出现在停用词词表中的词将不作为候选关系指示词。但是这种方法有一个最显著的问题：停用词词表需要人工地往里添加数据并且并不总能包括所有的停用词。一些停用词会因为它没有添加到词表而误作为候选关系指示词，而且这种情况出现的几率是非常大的。为解决此问题，本文在停用词的基础上，引入 TF-IDF 来进一步改善常用词的过滤。

TF-IDF 是一种统计方法，用以评估词语对于文本的重要程度。词语的重要性随着它在文本中出现的频率成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 实际上是：TF×IDF，TF 为词频 (Term Frequency)，IDF 为逆向文件频率 (Inverse Document Frequency)。

举例来说，在搜索“（中共中央，？，习近平）”时，从相关句子中一共分词得到了三个词：“总书记”、“同志”、“的”，那么三个词的 TF×IDF 计算方法及结果如表1所示：

词语	一次搜索中出现次数	计算	TF	语料库中含该词文件数	语料库中文件总数	计算	IDF	TF×IDF
总书记	6	6/18	0.33	220	18,000	$\log \frac{18000}{220}$	4.40	1.45
同志	2	2/18	0.11	262	18,000	$\log \frac{18000}{262}$	4.22	0.46
的	10	10/18	0.56	16,164	18,000	$\log \frac{18000}{16164}$	0.11	0.06

表 1: 词语的 TF×IDF 计算方法.

由表1可以看出，通过 TF×IDF 计算后，“的”这样的常见词已经排在了末尾。TF-IDF 这种方法好处在于不需要人工构造词表，只需将一份语料分好词，便可计算出每个词的 IDF 值，并且只要语料足够充分，未登录词的情况会很少。之后在选取候选关系指示词的时候，我们可以通过选取前 N 个词语作为候选关系指示词，以过滤“的”、“我”等常用词。

2.2.3 关系指示词与候选关系指示词间的相似度计算

如2.2.1节和2.2.2节中所述，为了获取给定的候选三元组的置信度，首先借助搜索引擎，依靠 TF-IDF 的计算结果获取相对准确的候选关系指示词集合，然后与候选三元组中关系指示词进行相似度计算，根据相似程度度量置信度。

对于符号化的关系指示词，无法使用有效方法计算其语义相似度，为此我们引入词向量表示将符号化的关系指示词向量化，进而通过计算两个向量的 cosine 相似度获取其语义相似度，词向量使用 word2vec 方法 [8] 进行训练（见3.1.2节）。这里我们把候选词中与关系指示词相似度最大的值作为三元组的置信度。如，对于三元组（中国，驻，美国），最后选出的前 N 个候选词中很可能没有“驻”，而是“建交”、“依赖”这样的词，显然，这样的词与“驻”语义上并不相近，因此其置信度可能会很低，而对于三元组（青海，位于，青藏高原），最后选出的前 N 个候选词中是有很大可能性包含“位于”的，即使没有“位于”，“坐落”、“地处”等词出现的概率也是非常大的，它们与“驻”语义上是非常相近的，因此其置信度可能会很高。通过选定适合的 N 值和置信度阈值（见3.3.1，3.3.2节），即可很大程度上过滤掉未正确抽取的三元组，保留正确抽取的三元组。

3 实验

3.1 实验设置与准备

3.1.1 数据集与评估方法

由于新闻报道用语规范，相比于其他文本更准确、简练，本文采用新闻语料作为实验所需的数据集。从各大门户网站共爬取 18000 条新闻报道，涉及军事、娱乐、体育等九个不同领域（如表2所示）。

新闻领域	文章数量	句子数量	新闻领域	文章数量	句子数量
财经	2,000	202,611	教育	2,000	143,505
军事	2,000	144,505	科技	2,000	139,186
汽车	2,000	119,084	时尚	2,000	58,278
体育	2,000	123,489	娱乐	2,000	78,255
政治	2,000	105,350	————	————	————

表 2: 实验数据集.

在评估阶段，本文使用准确率、召回率、以及 F1 值作为评估标准，以评估方法的性能。每个领域的新闻文本都为 2000 条，将每个领域的新闻语料输入 LTP 中获取候选实体关系三元组。之后，从得到的三元组中随机抽取 100 个作为测试，这 100 个样例代表了这个领域的特征。然后进行人工标注，计算其准确率、召回率以及 F1 值，这三样属性代表了本文方法对这个领域的适用程度。因为数据集中一共包含了九个领域的新闻文本，所以重复进行九次上述过程，将得到的九个领域的准确率、召回率，F1 值取平均后得出最终实验结果。

3.1.2 词向量训练

如2.2.3节所述，将三元组关系指示词词向量与候选关系指示词词向量进行 cosine 相似度计算，以得到两者的语义相似度。用于计算相似度的词向量是由 word2vec 提前训练好的。我们使用约十万条搜狗新闻语料进行训练，word2vec 使用 skip-gram 模型 [9]，获得了约 140 万中文词汇的词向量，部分词的词向量如表3和表4所示：

词语	相似度
总统	——
首相	0.853143
国防部长	0.834258
克里斯蒂娜 · 费尔南德斯	0.825957
司法部长	0.825429
普京	0.818455
代总统	0.817884
梅德韦杰夫	0.815461

表 3: 搜狗数据——50 维.

词语	相似度
总统	——
首相	0.781625
司法部长	0.777123
议长	0.771791
国防部长	0.767719
代总统	0.760391
财政部长	0.738668
外交部长	0.734386

表 4: 搜狗数据——100 维.

通过对比可以发现，相比与 50 维的词向量，100 维的词向量与目标词语（总统）的相似度的下降幅度更大，这是因为维度越高的词向量，如果其之间相似度越高就要求方向相近的维数越多，而随着维数的增多，两个词向量中方向相近的维数越多的概率会越来越小。除此之外，还可以发现 50 维的词向量中，与目标词语（总统）相似的词中出现了大量的人名，而 100 维的词向量中，与目标词语（总统）相似的词中则基本上都是和“总统”一样的职位词，这表明，100 维的词向量相对于 50 维的词向量，训练得更充分，词向量表示的信息更全面，因此实验部分，文本将使用 100 维的向量进行相似度计算。

3.2 基于启发式规则模板抽取候选关系三元组

对于3.1.1节中的 18000 条新闻文本，首先利用 LTP 进行句法依存分析，从而得到句法依存树，然后利用已经构造的启发式规则模板进行候选三元组的抽取，可以依次获得“定中关系三元组”、“主谓宾关系三元组”、“定语后置关系三元组”、“主谓动补关系三元组”等四类不同依存结构的候选三元组。各个新闻领域获取的候选三元组结果如表5所示：

新闻领域	候选三元组	定中关系	主谓宾关系	定语后置关系	主谓动补关系
财经	1,745	1,409	270	34	32
教育	888	767	90	19	12
军事	2,223	1,373	741	40	69
科技	918	779	113	13	13
汽车	532	468	45	12	7
时尚	376	264	94	11	7
体育	2,676	1,524	1,027	62	63
娱乐	748	519	201	13	15
政治	2,537	2,197	281	24	35

表 5: 候选关系三元组的抽取.

从表5中可以看出，在 4 种启发式规则模板中，满足定中关系的三元组最多，其次是满足主谓宾关系的三元组，满足定语后置与主谓动补的三元组则较少。这样的结果是由新闻文本的语言特征决定的，这表明在新闻语料中，大部分的句子满足定中关系与主谓关系，而较少的句子会用到定语后置与主谓动补关系。

3.3 基于搜索引擎结果的候选关系三元组置信度计算

3.3.1 利用搜索引擎获取候选关系指示词

针对不同启发式规则模板得到的候选关系三元组,抽取候选关系指示词时也有不同。在2.1.1节中, 以利用词语间定中关系启发式规则模板抽取出来的三元组，其关系指示词都为名词，对于这类三元组，应抽取相关句子中的名词作为相关词。利用其他启发式规则模板抽取出来的三元组，其关系指示词都为动词，所以抽取时将相关词句子中的动词作为相关词。

得到了相关词后，计算其 TF×IDF 值，将排在前 N 的词保存作为相似度计算的候选词。上文提到，对于 N 的取值有一定要求，首先应能正确将相关词与一般常用词“的”、“我”等区分，其次不能取得太小导致相关词没有作为候选关系指示词。为确定 N 的取值，本文从用 LTP 抽取出来的三元组中随机取 250 个进行实验，去除关系指示词后将它们放入搜索引擎，将每个三元组相关词中 TF×IDF 值最高的前 N 个词作为候选词保存，得到了图6显示的结果。图6记录的是当 N 取不同值时，候选词中至少有一个词和关系指示词的相似度 >0.85 的三元组所占的比例：

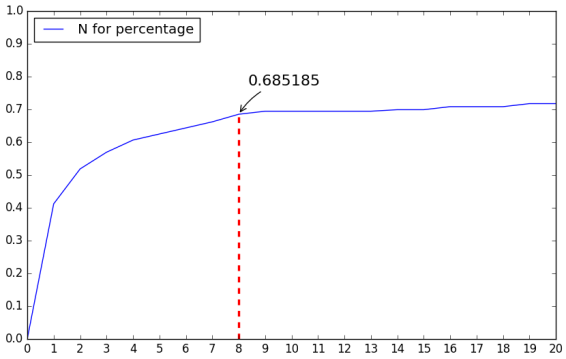


图 6: N 与满足条件三元组所占比例关系图

由图中可以发现，在 N 取值小于 8 时，曲线上升非常迅速，而当 N 到达 8 之后，曲线渐渐趋于平缓。也就是说，有大约 70% 的三元组在前 8 个相关词中能找到一个与关系指示词非常相近的词。这表明，与关系指示词相似的词集中在 TF-IDF 值排序前 8，而从第 8 个词之后，基本上都是一些“的”，“我”这样的常用词，因此曲线在 N 取 8 之前，上升非常迅速，而当 N 取值大于 8 后，曲线却渐渐趋于稳定。所以在从搜索引擎中抽取相关词的时候，将候选词数量的阈值设置成 8，即只取 TF*IDF 值前 8 的相关词作为候选关系指示词。

3.3.2 置信度阈值的确定与三元组准确率

获取了三元组的候选关系指示词后，将与关系指示词最相近的候选词的相似度值作为三元组的置信度。最后对于置信度选择一个阈值 T 作为三元组正确与错误的分割线，即置信度高于 T 的三元组将被判定为正确抽取的三元组，而置信度低于 T 的三元组将被判定为未正确抽取的三元组。

本文通过以下实验来确定阈值 T 的取值，以获得最好的实验效果：按照3.1.1节中所述，从每个领域中获得的实体关系三元组集合随机抽样出 100 个进行人工标注，计算其准确率、召回率以及 F1 值，重复进行九次，取平均后得出最终实验结果如图7、8所示：

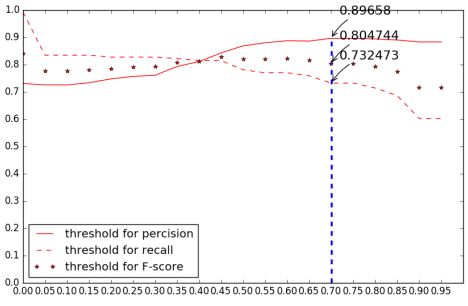


图 7: 阈值与准确率、召回率、F1 值关系图

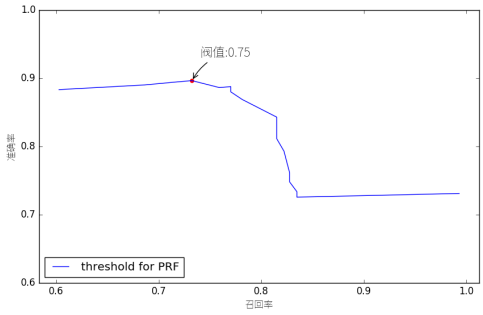


图 8: 准确率、召回率随阈值变化图

可以看到随着置信度的阈值从 0 提升到 0.95, 准确率从 70% 逐渐提升到 88% 左右，在阈值为 0.7 的时候，准确率都达到了 89.65%，而随着阈值继续提升，准确率反而出现了下降，这是因为即使经过置信度计算，仍有一部分未正确抽取的三元组置信度接近 1，随着阈值的提高，一部分正确抽取的三元组（如置信度为 0.9 的三元组）被误判，导致了准确率的下降，但是从曲线下降的幅度可以发现，这样的错误并不多。

而对于召回率，当阈值取 0 的时候，召回率达到了 100%，本文有必要对召回率的计算进行说明。众所周知，召回率 = 抽取结果正例数/文本中正例总数，在本文实验中，利用 LTP 对从网站上爬取的新闻文本进行抽取，得到抽取后的三元组。对于抽取后得到的三元组，可以使用人工标注的方法，标注出正确的三元组，但是对于互联网海量新闻文本中的正例总数，却很难统计。事实上，从互联网中提取数据，可能都面临这样的问题。如李维刚等人的工作 [15]，其采用模式扩展的方法，通过定义少量种子集合，在互联网中搜索匹配的关系实例，不断迭代进行自学习，最终获得了 98.42% 的准确率，但是对于召回率，却没有进行很好的评价。

召回率虽然在本文2.1节三元组抽取时难以计算，但是在本文3.3节置信度计算时却可以评价。以2.1节中抽取出来的候选三元组为基础，标注出其中正确抽取的三元组，再经过3.3节方法得到过滤后的三元组，统计其中正确抽取的三元组，算出召回率。本文主要针对的是知识图谱补全任务，更关心置信度计算方法的召回率，而不是从互联网新闻语料中抽取候选三元组的召回率，因此本文将召回率作为评价3.3节置信度计算方法的一项指标。

当阈值取 0 时，表明不管三元组的置信度多低，只要其大于 0，便判定为正确的三元组，等同于没有使用文中3.3节中的置信度计算方法，因为此时并没有过滤任何三元组，因此召回率在阈值取 0 时为 100%。从图7中可以看出随着阈值的提高，实验的召回率开始下降，这是因为把阈值取得越高，越少的三元组会被判定成正确的三元组，这难免会使一些原本是正确的三元组被误判，因此实验的召回率随着置信度阈值的提高一直下降。由于知识图谱构建及其补全要求较高的准确率，本文选取 0.7 作为置信度的阈值。

3.4 实验结果与分析

本文将开放域三元组抽取任务分为两步。首先利用启发式规则模板将三元组从新闻语料中抽出，然后再利用搜索引擎为每个三元组获取置信度，通过设置置信度阈值进一步过滤掉未正确抽取的三元组，得到最终的实验结果。

表6是只用启发式规则与启发式规则和搜索引擎相结合两种方法的实验效果对比

方式	准确率	召回率	F1 值	阈值
启发式规则	73.11%	100%	84.17%	0
启发式规则 + 搜索引擎	89.65%	73.24%	80.47%	0.7

表 6: 最终实验结果

候选三元组 （基于启发式规则抽取）	候选关系指示词 (基于搜索引擎抽取)	置信度 （基于搜索引擎计算）	系统判断	人工标注
(伊拉克, 首都, 巴格达)	首都	100%	True	True
(美国, 前财政部部长, 萨默斯)	部长	99.99%	True	True
(青海, 地处, 青藏高原)	位于	78.01%	True	True
(利比亚政府军, 攻打, 班加西)	击退	45.67%	False	False
(蔡演威, 如愿到, 济南)	特批	29.92%	False	False
(美国, 依赖, 俄国)	痛恨	36.44%	False	False
(江苏, 油田, 高邮)	农产品	30.30%	False	False

表 7: 部分实验结果

实验证明，本文的方法是拥有比较高的准确率与召回率的，表7是从正确判定的三元组中

选出的一些样例。从表中不难看出，本文的方法对于错误三元组的过滤效果较为明显。首先对于正确抽取的三元组，置信度普遍偏高，而未正确抽取的三元组则普遍偏低，具有一定的区分性，这表明3.3节中置信度的计算方法是有效的。其次，对于未正确抽取的三元组，无论是因为其没有价值（如：蔡演威, 如愿到, 济南），还是因为 LTP 抽取错误（江苏, 油田, 高邮）, 本文方法都能进行很好地过滤，体现出了一定的适应性。另外对于一些不太为人所知的信息如（新华社，记者，XXX），也能进行过滤，因此本文方法具备一定的应用价值。

4 结论

本文从中文句子中词语之间的依存特点入手，通过分析实体关系在语句中的表述方式，提出了四种启发式规则模板以辅助句法依存分析工具，从新闻语料中抽取候选三元组，再利用搜索引擎计算三元组的置信度，进一步过滤三元组，获得最终的实体关系三元组。实验结果显示，我们的方法简单而有效，获得了 89% 的准确率及高达到 80% 以上的 F1 值，满足实际要求，可以应用于知识图谱构建以及知识补全的任务上。

参考文献

[1] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

[2] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3, 2010.

[3] Wanxiang Che, Zhenghua Li, and Ting Liu. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics, 2010.

[4] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[5] J Fukumoto, F Masui, M Shimohata, and M Sasaki. Oki electric industry: Description of the oki system as used for muc-7. In *Proceedings of the 7th Message Understanding Conference*, 1998.

[6] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics, 2004.

[7] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444, 2002.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[10] Eric Miller. An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*, 25(1):15–19, 1998.

[11] Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. Patty: a taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1135–1145. Association for Computational Linguistics, 2012.

[12] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics, 2012.

[13] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, 2014.

[14] 刘克彬，李芳，刘磊，韩颖. 基于核函数中文关系自动抽取系统的实现. 计算机研究与发展, 44(8):1406–1411, 2007.

[15] 李维刚, 刘挺, and 李生. 基于网络挖掘的实体关系元组自动获取. 电子学报, 35(11):2111–2116, 2007.

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

[17] 陈宇, 郑德权, 赵铁军, et al. 基于 deep belief nets 的中文名实体关系抽取. 软件学报, 23(10):2572–2585, 2012.