**文章编号**：1003-0077(2017)05-0114-06

# 基于弱监督和半自动方法的中文关系抽取数据集构建

马超义，徐蔚然

（北京邮电大学 信息与通信工程学院，北京 100876）

**摘　要**：关系抽取是信息抽取中的一项基础任务，对信息检索、问答系统、知识图谱等有非常重要的意义。现有的关系抽取数据集存在包含类别太少、句子标注困难、不易扩展等缺陷，且只有英文数据集，不能很好地解决中文关系抽取任务。该文采用弱监督和半自动的方法，构建了一份中文关系抽取数据集，弥补了上述不足。首先借助维基百科抽取出丰富的关系对，从百度搜索返回结果及搜狗新闻语料中抽取包含实体对的句子，完成弱监督句子抽取过程。将句子放入 RNN 关系抽取系统进行打分，选取标注价值高的句子提交人工标注，对标注结果进行处理，最终得到中文关系抽取数据集。

**关键词**：关系抽取；数据集；弱监督；半自动

**中图分类号**：TP391　　**文献标识码**：A

# Semi-automatic Construction of Chinese Relation Extraction Data Set Based on a Weakly Supervised Method

MA Chaoyi, XU Weiran

(School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China)

**Abstract**: The relation extraction is a fundamental task in information extraction, with practical significance in information retrieval, question answering system and knowledge mapping, etc. The existing relation extraction data set are for English, containing very limited categories and neglecting sentence level annotations. This paper constructs a Chinese relation extraction data set using a weakly supervised and semi-automatic method. It firstly extracts a large amount of relation pairs from Wikipedia, then extracts sentences that contains entity pairs from the corpus of Sougou News and Baidu. Thus the weakly supervised sentence extracting is completed. These sentences are then scored in an RNN-based relation extraction system, selecting sentences with higher score for manual annotation. Finally the Chinese relation extraction data set is completed after manual annotation.

**Key words**: relation extraction; data set; weakly supervised; semi-automatic

## 1　引言

随着互联网技术的发展，网络日益成为人们生活中不可缺少的一部分。信息抽取能够帮助人们在海量的信息中快速定位到自己真正需要的信息，它是一个以自由文本作为输入，产生固定格式的、无歧义的输出数据的过程。

关系抽取是信息抽取的一项重要子任务，是指利用包含一对命名实体的自然语言文本来确定两者之间的关系。对信息抽取技术的研究和应用有重要意义，对信息检索、问答系统、信息过滤、机器翻译等有非常积极的意义。比如，在搜索服务中，用户想要知道某明星的出生日期，而网络搜索通常只返回包含用户搜索词的页面，无法洞悉用户的需求进而直接返回答案。而关系抽取的目的正是希望通过对网络中各类自由文本的解析，返回最有可能的结果作为答案。

实体关系抽取的方法，主要有基于知识工程的方法和基于机器学习的方法。基于知识工程的方法

**收稿日期**：2016-08-16　**定稿日期**：2017-04-26

**基金项目**：教育部博士点学科专项科研基金(20130005110004)

依赖于专家构建的知识库，花费大量的人力和时间，并且系统移植困难，所以基于机器学习的方法成为目前的主流。机器学习方法效果的好坏很大程度上依赖于训练数据集的质量。目前被广泛采用的是2009 年构建的 SemEval-2010 Task 8 数据集，数据集包含九种关系定义，最新的关系抽取系统可以达到 85％以上的准确率。

考虑到现有数据集类别种类不够丰富，分类效果已很难提高，且不能很好地解决中文关系抽取的任务。本文在总结了现有关系数据集的基础上：①采用弱监督的方法获取待处理语料，丰富了关系类别及句子类型；②采用半自动的方式，处理获得的语料，在保证准确性的基础上，大大降低了人工标注工作量；③最终通过标注得到中文关系抽取数据集，供中文关系抽取任务使用。

## 2　任务设定

### 2.1　相关工作总结

现有关系抽取数据集主要有两个，一是 SemEval-2010 Task 8 数据集，该数据集构建于 2009 年，共包含九种互不相容的关系，如因果关系、包含关系等。数据集包含 10 717 条数据，其中每条数据是一个包含实体对的句子，类别标签为实体对在该句中表现出的关系，例如：

My new ＜e1＞ apartment ＜/e1＞ has a ＜e2＞ large kitchen ＜/e2＞.　--Component-Whole

该数据集被广泛应用，已被引用 108 次。目前关系抽取系统在该数据集上的分类效果已达到 85％以上，在错分的句子中很多通过人工都很难准确识别，因而需要更丰富的类别和更多的句子供关系抽取系统使用。

另一评测数据集是 TAC-KBP 关系抽取任务给出的官方答案。该任务中共包含 41 种关系类别，约 33 000 句。句子类别相对丰富，但句子都来自官方给出的新闻语料，类型不够丰富且包含网页中的多余字符。由于允许利用共指信息完成推断，因而答案常常包含整段信息，不能很好地应用于关系抽取任务。

### 2.2　弱监督与半自动的抽取框架

本文在参考了现有英文数据集的基础上，构建了关系抽取的中文数据集，框架如图 1 所示。首先，通过弱监督的方式，从数据库中抽取实体对，进而从自由文本中取得更多更精确的句子，然后交由半自动的标注系统，进一步抽取标注价值更大的句子完成标注工作，在保证句子可靠性和多样性的基础上，降低了人工标注的难度。

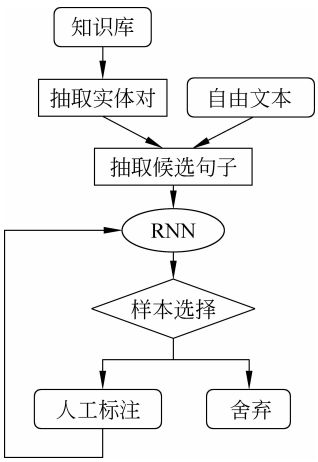


图 1　数据集构建流程

## 3　弱监督句子抽取过程

关系抽取的标注工作需借助句子中表达的内容确定句子中实体对的关系。若直接从网络数据中找出包含实体对，且描述恰为已定义关系类型的句子，将耗费大量的人力，效率低下且标注效果也将受到影响。

因此本文提出了一种弱监督的句子抽取方法。首先，借助外部知识库，找出属于特定关系的实体对，再从自由文本中选择包含这些实体对的句子。比如，在寻找属于“出生地”这个关系的句子时，借助外部知识库，发现 ＜奥巴马，夏威夷＞实体对属于该关系，然后抽取包含 “奥巴马”和“夏威夷”两个词的句子，相较于随机抽取的句子，如只包含“奥巴马”而不包含“夏威夷”的句子，这种方式得到的结果更有可能属于“出生地”这个类别，而且可以很方便地标注句子中待确定关系的实体对的位置。

### 3.1　关系定义

关系定义时，最重要的两点是关系的完备性和独立性。完备性是指我们定义的关系应尽可能地覆盖所有的句子，同时关系之间应相互独立，即不出现一个句子同时属于两个关系的情况。由于关系类型

的多样性，通常在已经能够包含大多数句子的情况下，会将剩余关系全部归于“其他”。比如，SemEval-2010 Task 8 数据集，定义了九种相互独立的关系，然后将其余不属于所列九种关系的句子全部归于第十个类别“其他”。

但由于 SemEval-2010 Task 8 中类型不够丰富，现有方法已能达到很高的分类准确性。本文采用了 TAC-KBP 2015 年 slot-filling 任务中对句子关系的定义，共 41 种，更加细致也更加丰富。这 41 种关系可以按照候选实体的类型、数量分别分类。按候选实体类型可以分为三种：名称、数值、字符串。其中名称类实体包括人名、地名、组织机构名。数值类实体为数字或者日期。字符串型实体是除以上两种类型外的其他实体，如宗教信仰、死亡原因等关系所对应的实体。每种关系含义的具体描述在 slot-filling 任务的任务说明[①]中有详细介绍。关系具体名称及类型如表 1 所示。

**表 1 slot-filling 关系列表**

| Type | Slot_name | Type | Slot_name |
|---|---|---|---|
| PER | per:alternate_names | PER | per:date_of_death |
| PER | per:children | PER | per:cause_of_death |
| PER | per:country_of_birth | PER | per:charges |
| PER | per:city_of_birth | PER | per:religion |
| PER | per:city_of_death | PER | per:title |
| PER | per:schools_attende | ORG | org:alternate_names |
| PER | per:cities_of_residence | ORG | org:city_of_headquarters |
| PER | per:country_of_death | ORG | org:country_of_headquarters |
| PER | per:date_of_birth | ORG | org:founded_by |
| PER | per:origin | ORG | org:member_of |
| PER | per:other_family | ORG | org:members |
| PER | per:parents | ORG | org:parents |
| PER | per:countries_of_residence | ORG | org:political_religious_affiliation |
| PER | per:siblings | ORG | org:shareholders |
| PER | per:spouse | ORG | org:date_dissolved |
| PER | per:stateorprovince_of_birth | ORG | org:top_members_employees |
| PER | per:stateorprovinc | ORG | org:subsidiaries |
| PER | per:statesorprovinces_of_residence | ORG | org:stateorprovince_of_headquarters |
| PER | per:age | ORG | org:date_founded |
| PER | per:employee_or_member_of | ORG | org:number_of_employees_members |
| | | ORG | org:website |

### 3.2 实体对获取

本文采用弱监督的方式，借助结构化数据库完成关系对获取，目的是从结构化数据库中获取属于特定关系的实体对。在此之前，需要将已经定义的关系与知识库中描述的关系类型进行对应。

为保证所抽取实体对的可靠性及多样性，本文采用维基百科作为辅助的外部数据库。我们下载了离线的中文维基百科数据库[②]，共包含 11GB 语料，包含约 400 万词条。每个词条对应一个维基百科页面，维基百科中人名、组织机构名等页面都包含实体关系描述部分，如图 2 所示是词条奥巴马的关系描述部分。

这部分信息记录在离线数据库的 info-box 部分，含有 info-box 的词条共 20 万个。通过人工筛选，我们得到了与这 41 个英文关系所对应的维基关系描述共 331 条，如与 alternname 对应的维基关系有别名、alias、nickname 等。通过抽取这些关系包含的实体，并经过一些简单的字符处理，最终得到候

① http://www.nist.gov/tac/2015/KBP/ColdStart/guidelines/TAC_KBP_2015_Slot_Descriptions_V1.0.pdf

② https://dumps.wikimedia.org/zhwiki/


图 2　维基百科页面示例

选关系对共 24 多万条，如表 2 所示。

**表 2　关系对实例**

| KBP-relation | Wiki-relation | Entity1 | Entity2 |
|---|---|---|---|
| alternate_names | 别名 | 游泳中心 | 水立方 |
| City_of_residences | 居住城市 | 奥巴马 | 芝加哥 |
| schools_attended | 毕业院校 | 陈希同 | 北京大学 |
| Title | 头衔 | 二月河 | 作家 |

### 3.3　包含实体对的句子获取

在对 TAC-KBP 任务数据集进行分析的过程中，我们发现通过新闻语料获取的句子形式不够丰富，不能充分地包含各种类型的句子。因此，我们在句子抽取过程中加入了百度搜索的结果，既可以很方便地得到包含实体对的句子，又能得到各种类型的句子形式，且能方便地拓展句子数量，解决类别间样本不平衡的问题。

我们首先使用了搜狗实验室提供的 sogouCA 新闻数据集，来自搜狐新闻 2012 年 6—7 月期间国内、国际、体育、社会、娱乐等 18 个频道共 3GB 的新闻数据。从中匹配包含已获得实体对的句子。同时，为保证候选句子类型的多样性，我们利用百度 API，抓取了搜索目标实体对后的返回页面内容，通过字符匹配得到包含目标实体对的句子，这样大大地丰富了原有的结果。最终我们从 sougouCA 得到了 5 万多条候选句子，从百度返回结果中获取了 15 万多条句子。由于百度结果的丰富性，结果可随时进行扩充。

为方便接下来的句子分类任务，我们在句子中加入标记符对包含的实体进行定位。最终句子形式如下：

＜e1＞刘墉＜/e1＞，台湾著名＜e2＞作家＜/e2＞，由造成轰动的《萤窗小语》开始，到近年《爱就注定了一生的漂泊》，总共出版了 30 多本书。

## 4　基于 RNN 的半自动关系标注过程

使用半监督的方法，可以确保抽取的句子中包含特定关系的实体对，但由于句子源自自由文本，仍存在很大的冗余，直接提交人工标注仍需较大工作量。因而本文采用了一种半自动的标注方式，将得到的句子首先交由训练好的 RNN(recurrent neural network)关系抽取系统打分，选择置信度较高的提交人工标注，进一步降低了人工标注的数量，且每个句子有预设的类别，降低了标注的复杂度。

### 4.1　RNN 框架介绍

传统关系抽取方法如基于模式匹配的关系抽取、基于字典驱动的关系抽取等，都需要根据句子的语法特性，设定具体的模式，结果依赖于模式的优劣及多样性。与传统关系抽取方法相比，基于机器学习的方法有更好的拓展性。该方法的实质是将关系抽取看作一个分类问题，通过具体的机器学习算法，借助标注语料构造分类器，然后将其应用于特定关系的判别。

随着深度学习理论的不断发展，RNN 在自然语言处理，尤其是句子级的分类任务中取得了很大进展。相较于传统方法依赖自然语言处理工具，进行实体识别、词性标注、句法解析等预处理工作，RNN 可只利用词向量来表示每一个词，作为网络的输入。并且利用神经网络来做句子分类的效果已逐渐超过传统方法。

如图 3 所示，分类器共包含三个部分，分别是：①词向量层，将输入句子中的每个词转为词向量表示；②双向的循环层，将词序列正向反向分别输入，得到词级别的特征；③最大池化层，将前一层得到的词级别特征合并成句子级别的特征。最终将句子级别的特征用于分类。

借助词向量的表达能力及 RNN 本身对词的记忆能力，该模型在关系抽取任务中取得了很好的效果，已被广泛应用于各种句子分类任务。以上模型经过训练，可以在英文训练集 SemEval-2010 Task 8 上取得 80.0%的准确率。

### 4.2　半监督标注过程

在半监督标注过程中，我们采用了基于置信度的标注策略，使用一个句子 $S$ 被分为某一类别的最大概率值 $p_0 = \max p_i(S)$ 作为置信度，置信度越高

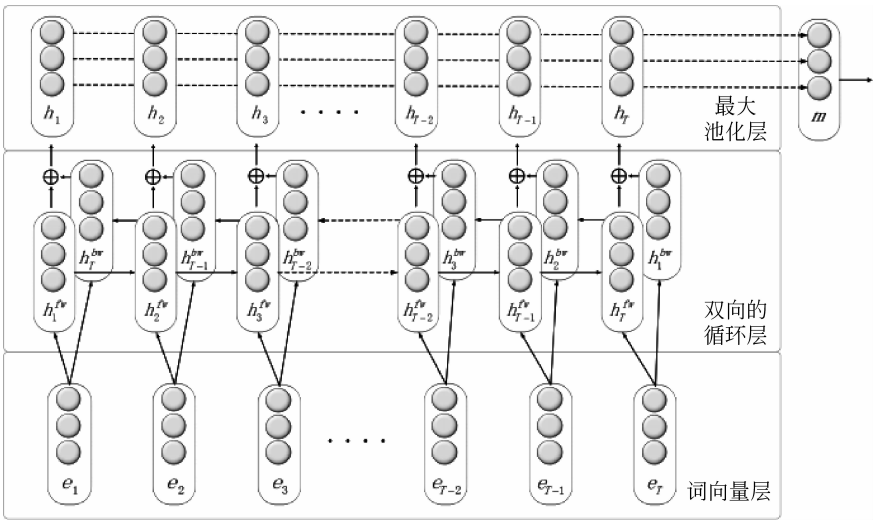


图 3 循环神经网络结构

的句子越容易被标注为“是”从而加入训练数据集，但太高的阈值会导致待标注句子数过少。本文最终选定 0.75 作为阈值，保证有足够的结果加入最终数据集。如果置信度大于某一阈值，则认为该句子具有更大的标注价值，需要提交人工标注。针对标注任务，我们设计了简易的标注工具，如图 4 所示。

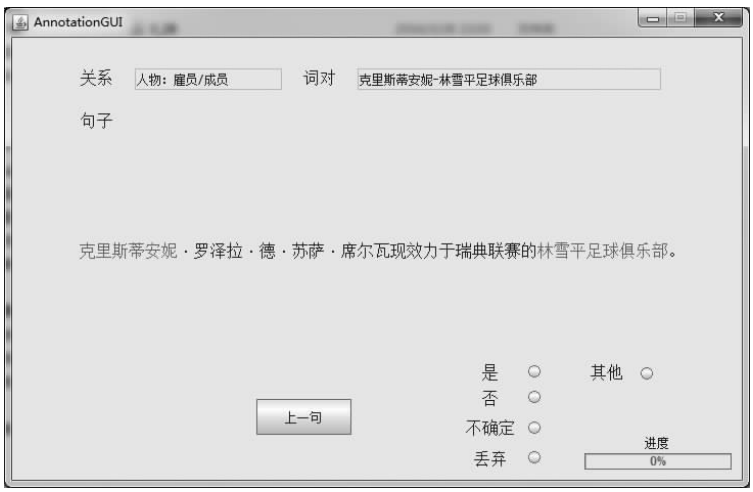


图 4 简易标注工具

从图 4 可以看出：标注选项共有五个：“是”“否”“不确定”“丢弃”“其他”，分别代表：句子属于该类别、句子不属于该类别、不能确定句子是否属于该类别、句子格式或表述有问题，直接丢弃、句子不属于 41 个关系中的任一类别。标注中，若标注人员一致标注为“是”或“其他”，则直接加入数据集。对于标注不一致或标为“不确定”的部分，可以帮助我们找出标注过程中定义不清楚的情况。

初始阶段，我们随机地从每个类别中选取一些句子进行标注，以完成对 RNN 的初始训练，接着从余下的数据集中选择句子放入分类器打分，根据打分结果决定是否提交人工标注，标注完成后的句子将继续应用于 RNN 的训练。重复以上步骤，完成对句子的处理，得到最终的数据集。

## 5 结果分析

最终，我们选择了 41 个关系类别，通过人工标注的方式，从维基百科中选择了 331 种关系定义作为这 41 种关系的拓展；通过半监督的方式，从维基百科中获取中文实体对 24 万多个，从百度结果页面及 sougouCA 新闻语料中抽取句子 20 万余句。经由 RNN 抽取系统，选择了约一万个句子提交人工标注，经过人工标注，将 5 031 个句子加入数据集，

作为最终的结果。该方法可供中文关系抽取及多类别句子分类任务使用，填补了中文数据集的空白。

经过半监督的抽取过程及分类器的打分，我们对候选句子进行了有效的删减，保留的句子包含候选类别，大大降低了人工标注的难度。标注过程中，我们通过分析标注不一致及标注结果为“不确定”的句子，逐步明确了各类别的定义。最后，针对标注结果中某些类别句子数很少的情况，我们重新拓展了这些类别的句子数，保证各类别句子数不会太少。

我们利用文中提到的 RNN 结构在最终的数据集上进行了实验。随机抽取每个关系中 80％的句子作为训练集，20％的句子作为测试集。最终在 41 个类别上的分类准确率为 76％。相同分类器在 TAC-KBP 的数据集上，三万个句子、41 个类别上的分类准确率为 61.6％。说明最终的数据集更加规范有效，适用于关系抽取任务。

## 6　总结及展望

本文采用弱监督和半自动的方法，构建了一份中文关系抽取数据集，填补了中文关系抽取数据集的空白。在数据集构建过程中，参考现有英文数据集的构建方式，并针对其关系类别少、句子形式不够丰富、标注复杂等缺陷，采用弱监督的方式抽取句子，采用半自动的方法对结果进行进一步处理，大大降低了人工标注的难度。最终对数据集的评测效果证明了数据集的实用性。

由于中文关系抽取任务并没有广泛开展，本文中的关系定义参考了英文数据集的定义方式，后续可逐步拓展和完善。分类器设计部分也可根据中文语法句法等的特点加入更多信息，或修改神经网络，以达到很好的效果。

## 参考文献


[1] 陈立玮，冯岩松，赵东岩. 基于弱监督学习的海量网络数据关系抽取[J]. 计算机研究与发展，2013，50(9)：1825-1835.

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

[3] 牟晋娟，包宏. 中文实体关系抽取研究[J]. 计算机工程与设计，2009 (15)：3587-3590.

[4] 余东，李诺，申德荣，等. ERE：基于半结构化 Web 页面的实体关系抽取系统[J]. 计算机与数字工程，2014，42(9)：1581-1586.

[5] 杨静，徐蔚然，谭松波. COAE2014 情感关键句评测任务和评测数据设计[C]. 第六届中文倾向性分析评测委员会，2015：51.

[6] 杨博，蔡东风，杨华. 开放式信息抽取研究进展[J]. 中文信息学报，2014，28(4)：1-11.

[7] HENDRICKX I，KIM S N，KOZAREVA Z，et al. Semeval-2010 task 8：multi-way classification of semantic relations between pairs of nominals[C]//Proceedings of the Workshop on Semantic Evaluations：Recent Achievements and Future Directions. Association for Computational Linguistics，2009：94-99.

[8] 贾真，何大可，杨燕，等. 基于弱监督学习的中文网络百科关系抽取[J]. 智能系统学报，2015，10(1)：113-119.

[9] 杨宇飞，戴齐，贾真，等. 基于弱监督的属性关系抽取方法[J]. 计算机应用，2014，34(1)：64-68.

[10] 涂新辉，张红春，周琨峰，等. 中文维基百科的结构化信息抽取及词语相关度计算方法[J]. 中文信息学报，2012，26(3)：109-115.

[11] Mikolov T，Karafiá T M，Burget L，et al. Recurrent neural network based language model[C]//Proceedings of the Interspeech，2010(2)：3.

[12] 戴敏，朱珠，李寿山，等. 面向中文文本的情感信息抽取语料库构建[J]. 中文信息学报，2015，29(4)：67-73.

[13] Zhang Z. Weakly-supervised relation classification for information extraction[C]//Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM，2004：581-588.



马超义(1991—)，通信作者，硕士研究生，主要研究领域为自然语言处理和信息抽取。
E-mail：machaoyi@bupt.edu.cn



徐蔚然(1975—)，副教授，博士，研究生导师，主要研究领域为信息抽取、知识图谱等。
E-mail：xuweiran@bupt.edu.cn