Automatically Labeled Data Generation for Large Scale Event Extraction
大规模事件抽取的自动标记数据生成

2020 年

目录

摘要

大多数事件抽取任务都依赖于人工标注数据，但是人工标注数据十分昂贵，事件类型少，每个事件类型数量也少，因此有监督的事件抽取很难利用到大规模事件抽取中。本文提出利用知识库（Freebase 和 FrameNet,Wikipedia）来自动标记数据。本文自动标记的数据包含人工标注的数据，可以从人工标注的数据中提升性能。

1 简介

ACE 的缺点：

1. ACE 数据集中，33 个事件类型是在 599 个英文文档中人工标注的。
2. ACE 数据集中 60% 的事件类型都是少于一百个样例，甚至有三个事件类型是少于 10 个样本的。
3. 预定义的 33 个事件类型不能运用到大规模事件抽取中。

图一是标记句子的示例。

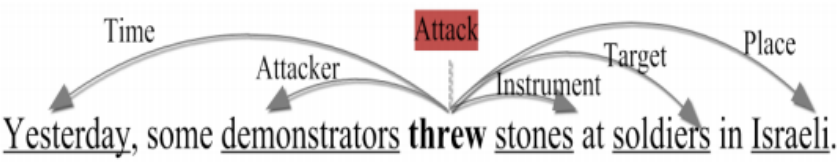

Figure 1: This sentence expresses an *Attack* event triggered by *threw* and containing five arguments.

图 1: 示例

将远程监督（DS）运用到事件抽取的挑战：

- 在知识库中，触发词并不是提前给出的。为了解决这个问题，文在运用远程监督之前，先发现触发词，再去自动标记事件元素。
- 根据远程监督在关系抽取内的应用，我们假设一个句子包含所有的事件元素，然而，对于特定事件，元素是分布在多个句子中。

为了解决上述问题，我们提出一个方法利用 Freebase 和 FrameNet 来为大规模事件抽取自动标记数据。首先，我们提出了一种利用 Freebase 对每种事件类型的参数进行优先级排序和选择关键参数或代表参数的方法（详见 3.1 节）；其次，我们只使用关键参数来标记事件并找出触发词；第三，外部语言知识资源 FrameNet，然后，我们提出了一种 Soft Distant Supervision（SDS）方法，它可以自动标注训练数据，假设任意一个包含 Freebase 中所有关键参数的句子和一个对应的触发词都有可能以某种方式表示该事件，在那句话中出现的论点很可能在那件事中起到相应的作用。最后，通过人工和自动两种方式对自动标注的训练数据进行质量评价。此外，我们采用基于 CNN 的 EE 方法，对自动标注的数据进行多实例学习，作为进一步研究该数据的基线。

总之，本文的贡献如下：

- 据本文作者所知，这是第一次利用 Freebase 和 FrameNet 来大规模标记数据。

- 我们提出了一种利用 Freebase 计算事件关键参数的方法，并利用它们自动检测事件和相应的触发词。此外，我们使用框架网来过滤有噪音的触发器，并扩展更多的触发器。
- 实验结果显示此方法是有效的。同时，我们的自动标注数据可以扩充传统的人工标注数据，从而显著提高提取性能。

2 背景

Freebase: 中间不仅有三元组原子知识表示，还创造了虚拟节点结构，被称为组合值类型（CVT）。

FrameNet: 包含一千多个框架以及 10000 多个词法单元，每个框架可以被看作是一种事件类型的语义框架。每个框架都有一组引理，部分词法标注可以唤起该框架，称为词法单元（LU）。

Wikipedia: 我们使用的维基百科于 2016 年 1 月发布。其中 630 万篇文章都用于我们的实验。我们使用 Wikipedia 是因为它是相对最新的，而且 Freebase 中的很多信息都来自 Wikipedia。

3 生成数据的方法

这一节主要包含以下四个部分：

- 关键元素检测。
- 触发词检测。
- 触发词过滤和扩展。
- 自动标记数据生成。

如图 4 所示。

3.1 Key Argument Detection

有些元素在事件中十分的重要，通过这些元素可以轻易地将不同事件区别开来。本文使用 Key Rate (KR) 来评估一个元素在事件中的重要性。这个重要性取决于两个因素：Role Saliency （角色显著性）和 Event Relevance（事件相关性）。

Role Saliency (RS): 表示在给定一个特定事件类型下，通过这个元素可以识别出这个事件的程度。也就是，这个元素可以将不同事件区别开的重要程度。RS 的定义如下：

$$RS_{ij} = \frac{Count(A_i, ET_j)}{Count(ET_j)} \tag{1}$$

RS_{ij} 是第 j 个事件中，第 i 个元素的重要程度。$Count(A_i, ET_j)$ 是在 Freebase 中，事件 j 中，元素 i 出现的次数。$Count(ET_j)$ 是事件 j 在 Freebase 出现的次数。

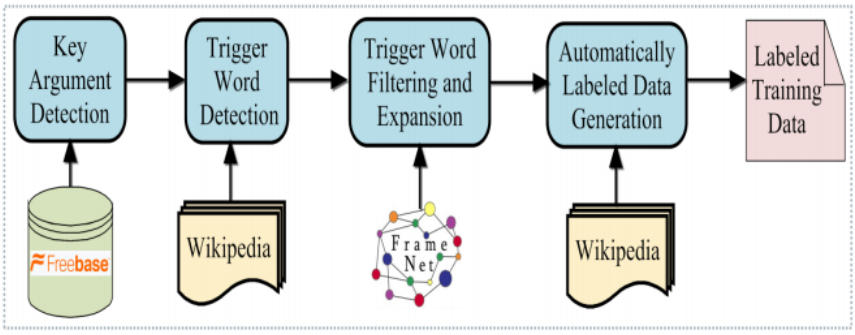

Figure 4: The architecture of automatically labeling training data for large scale event extraction.

图 2: 生成数据

Event Relevance (ER): 反映这个元素可以将事件抽取出来的能力。如果这个元素出现在每一个事件中，那么这个元素就会有很低的事件相关性。ER 的计算式如下：

$$ER_i = log \frac{Sum(ET)}{1 + Count(ETC_i)} \tag{2}$$

ER_i 是第 i 个元素的事件相关性，Sum(ET) 是在知识库中所有事件类型的数量。$Count(ETC_i)$ 是包含第 i 个元素的事件类型，最后 KR 的计算如下：

$$KR_{ij} = RS_{ij} * ER_i$$

本文计算了每个事件类型的所有元素，并根据 KR 进行了排序，然后本文选取了前 K 个元素作为关键元素。

3.2 Trigger Word Detection

本节使用上一节的关键元素来标记在维基百科中可能表示事件的句子。本文使用 Standford CoreNLP tool 来处理维基百科中的原始文本，进行词性标注，命名实体识别。最后，我们选择 Freebase 中包含事件实例所有关键参数的语句作为表示相应事件的语句。然后我们用这些标记的句子来检测触发词。

动词往往会成为表示一个事件的关键。在一个事件类型中，一个动词比其他动词出现的次数多，那么这个动词会触发这个事件。但是，这个动词出现在每一个事件中，就不会触发这个事件。因此我们提出 Trigger Candidate Frequency (TCF) 和 Trigger Event Type Frequency (TETF) 去评估两个方面。最后，使用 Trigger Rate (TR) 来作为最后这个动词成为触发词的可能性。

$$TR_{ij} = TCF_{ij} * TETF_i$$

$$TCF_{ij} = \frac{Count(V_i, ETS_j)}{Count(ETS_j)}$$

$$TETF_i = log\frac{Sum(ET)}{1 + Count(ETI_i)}$$

TR_{ij} 是在第 j 个事件类型中，第 i 个动词的触发概率。$Count(V_i, ETS_j)$ 是第 j 个事件类型下，包含第 i 个动词句子的数量。$Count(ETS_j)$ 是第 j 个事件类型的数量。$Count(ETI_i)$ 是包含动词 i 的事件数量。最后，为每个事件选择合适的触发词。

3.3 Trigger Word Filtering and Expansion

通过以上的触发词检测，我们可以得到一个初始的触发词。然而，这个最初的触发词汇是不准确的，并且只包含了动词的触发词，名词触发词被忽略了。本文使用 FrameNet 来过滤不准确的动词以及扩展名词的触发词。使用词嵌入，将 Freebase 里面的事件映射到 FrameNet 的框架里，计算 Freebase 事件类型与 FrameNet 中事件框架的语义相似度。公式如下：

$$frame(i) = \underset{j}{argmax}\ (similarity(e_i, e_{j,k}\ )) \tag{3}$$

然后，我们过滤动词，它是在最初的动词触发词词典，而不是在映射框架。我们在映射框架中使用高置信度的名词来扩展触发词汇。

3.4 Automatically labeled data generation

最后，本文提出 Soft Distant Supervision，并用它进行自动生成训练数据。也就是假设任何包含 Freebase 中所有关键参数和相应触发词的语句都可能以某种方式表示该事件，并且该语句中出现的参数可能在该事件中扮演相应的角色。

4 Method of Event Extraction

在本文中，事件抽取有两个阶段，是一个多分类任务。第一个阶段是**事件分类**，如果在 Freebase 中，关键元素出现了，那么就进入第二个阶段**元素识别**，进行识别事件中的元素。使用 Multi-pooling Convolutional Neural Networks with Multi-instance Learning(DMCNNs-MIL) 来进行两个阶段的工作。为了防止错误标签带来的问题，本文在两个 DMCNN 上使用 Multi-instance Learning (MIL) 多示例学习。两个 DMCNN 分别在两个阶段使用多示例学习。在每个包上，目标函数使用交叉熵。

$$J(\theta) = \sum_{i=1}^{T} logp\left(y_i \left| m_j^i, \theta\right.\right)$$

事件 j 使用最大似然函数，拥有这些元素的情况下，最有可能是哪些事件。

$$j^* = arg\max_j p\left(r \left| m^i_j, \theta\right.\right)$$

采用随机梯度下降法（Adadelta),mini-batches.

5 Experiments

在本节，先人工评估自动标记的数据，然后评估在 ACE 数据上标记的数据，最后评估 DMCNN-MIL 在自动标记数据上面的性能。

5.1 Our Automatically Labeled Data

首先设置超参数，关键元素的个数为 2（后面做实验得到的结果），触发词重要程度为 0.8。先使用两个关键元素去标记数据，接着使用这些标记数据以及 FrameNet 来找触发词以及用 SDS 来生成标记数据。最后生成 72611 个标记数据。

5.2 Manual Evaluations of Labeled Data

从自动标记的数据中随机选择 500 个样本，每个样本由三个人进行标记，自动标记数据的准确率，如表所示。

阶段	平均准确率
触发词标记	88.9
元素标记	85.4

5.3 Automatic Evaluations of Labeled Data

本文使用两个方式将自动标记的数据加入到 ACE 的数据库。

1. 我们删除 ACE 数据中有这些重复事件类型的人工注释 ACE 数据，并将自动标记的数据添加到剩余的 ACE 训练数据中。我们叫扩展数据（ED）。
2. 我们直接把自动标注的数据加入到 ACE 相同事件类型中，我们把这个训练数据叫 ACE+ED

用 DMCNN 来训练这些数据，并在 ACE 数据中进行测试。测试结果如下：这个结果证明

Methods	Trigger Identification(%)			Trigger Identification + Classification(%)			Argument Identification(%)			Argument Role(%)		
	P	R	F	P	R	F	P	R	F	P	R	F
Li's structure trained with ACE	76.9	65.0	70.4	73.7	62.3	67.5	69.8	47.9	56.8	64.7	44.4	52.7
Chen's DMCNN trained with ACE	80.4	67.7	73.5	75.6	63.6	69.1	68.8	51.9	59.1	62.2	46.9	53.5
Nguyen's JRNN trained with ACE	68.5	75.7	71.9	66.0	73.0	69.3	61.4	64.2	62.8	54.2	56.7	55.4
DMCNN trained with ED Only	77.6	67.7	72.3	72.9	63.7	68.0	64.9	51.7	57.6	58.7	46.7	52.0
DMCNN trained with ACE+ED	79.7	69.6	**74.3**	75.7	66.0	**70.5**	71.4	56.9	**63.3**	62.8	50.1	**55.7**

Table 4: Overall performance on ACE blind test data

图 3: 测试结果

自动生成数据是有效的。

5.4 Discussion

Impact of Key Rate

特征	触发词（F1 ）	元素（F1)
ACE	69.1	53.5
AEC+RS	70.1	55.3
AEC+ER	69.5	54.2
AEC+KR	70.5	55.7

Impact of Trigger Rate and FrameNet
在这个部分证明 TR 和 FrameNet 找到触发词的有效性。用 Grid search 来进行寻找最优参数 (0.5, 0.6, 0.7, 0.8, 0.9, 1.0)。最后是 0.8。

Feature	Trigger	Argument
	F_1	F_1
ACE	69.1	53.5
ACE + TCF	69.3	53.8
ACE + TETF	69.2	53.7
ACE + TR	69.5	54.0
ACE + TR + FrameNet	**70.5**	**55.7**

Table 6: Effects of TCF, TETF,TR and FrameNet

图 4: 测试结果

5.5 Performance of DMCNN-MIL

有两种评估本文方法的方式：held-out and manual evaluation

Held-out Evaluation　用两个标准去评判自动预测事件的准确性：

1. 关键元素和事件类型是否匹配
2. 事件类型和元素角色是否匹配

我们可以看到，多实例学习可以有效地缓解远程监督事件抽取中的噪声问题。

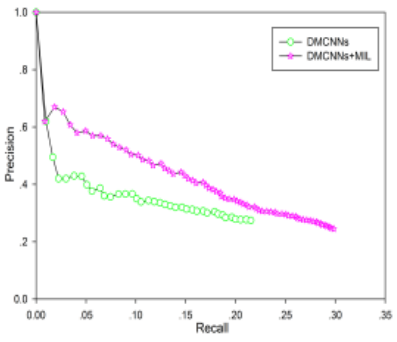

Figure 7: P-R curves for event classification.

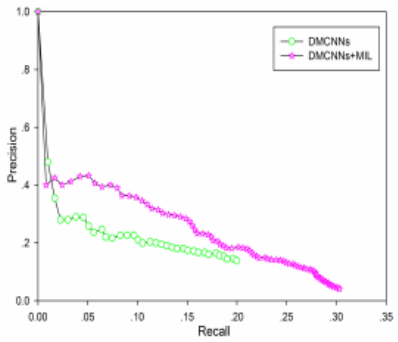

Figure 8: P-R curves for argument classification.

图 5: 测试结果

Methods	Event Classificaiton			
	Top 100	Top 300	Top 500	Average
DMCNNs	58.7	54.3	52.9	55.3
DMCNNs+MIL	**70.6**	**67.2**	**64.3**	**67.4**
Methods	Argument Classificaiton			
	Top 100	Top 300	Top 500	Average
DMCNNs	43.5	40.6	36.7	40.3
DMCNNs+MIL	**50.8**	**45.6**	**43.5**	**46.6**

Table 7: Precision for top 100, 300, and 500 events

图 6: 测试结果

Human Evaluation　这个部分仅仅评估没有在 Freebase 中出现的事件示例。

可以看出 DMCNNs-MIL 获得了最好的结果。

6 Conclusion and Future Work

在未来，我们将使用所提出的自动数据标记方法来处理更多的事件类型，并探索使用自动标记的数据来提取事件的更多模型。