史上最可爱的关系抽取指南？从一条规则到十个开源项目

夕小瑶的卖萌屋 2019-12-29

以下文章来源于AINLP ，作者太子長琴

正文来自订阅号：AINLP

正文作者：太子長琴（NLP算法工程师）

小夕，我来给你出个题，看看你的反应如何？

好呀好呀，你说~

“梁启超有一个著名的儿子，叫梁思成;同时还有一个著名的学生，叫徐志摩。这两个人有一个共同的爱人，叫林徽因。林徽因的父亲是林长民，林长民不仅自己是名人，还有一个名人哥哥，叫林觉民。林徽因最后与梁思成结婚了，徐志摩娶的则是陆小曼。”
那么问题来了，梁启超和林长民是什么关系？

亲家！

哇，你怎么一口就答出来了！？
!?(・_・;?

嘻嘻，我悄悄拿我的关系抽取模型跑了一下~(￣▽￣)"

啥是关系抽取?

呐，跟我来~！

信息抽取是NLP领域的一个经典任务了，如何从不同来源的自然语言文本中提取真实可用的“知识”，并保证这些“知识”通常是**清楚的、事实性的信息**，是当今信息抽取领域的一个难点。

信息抽取三个最重要/最受关注的子任务：

- **实体抽取**
  也就是命名实体识别，包括实体的检测（find）和分类（classify）
- **关系抽取**
  通常我们说的**三元组（triple）**抽取，一个谓词(predicate)带2个形参(argument)，如 Founding-location(IBM,New York)
- **事件抽取**
  相当于一种多元关系的抽取

属性一般的形式是（实体，属性，属性值），**关系**的一般形式是（实体，关系，实体）。简单来区分的话，关系涉及到两个实体，而属性只有一个实体。
不过呢，属性和关系的提取是可以相互借鉴的。下面小夕以关系抽取为出发点，从**算法、数据集、评价指标、比赛**这几个方面给大家介绍关系抽取的相关知识。**文末还有一个蛋哦～～～**

关系提取方法

基于模板

这种方法比较简单，一般都是根据先验知识设计一些模式，然后在语料中匹配这些模式。举几个例子：

- 马云作为企业家，对应的模式是：`XX (?:作为|是) YY`

- 刘强东是京东的创始人，对应的模式是：`XX (?:作为|是) YY 的? ZZ`

这里的 XX YY 和 ZZ 自然就是前一步识别出来的实体了。

相关资源包括：

- 100 Best GitHub: Expert System | Meta-Guide.com

基于句法分析

主要是找到主谓宾，一般都是在句法分析的基础上进行的。举几个例子：

- 感冒是一种病，对应的句法结构为：感冒（SBV），是（Root），病（VOB）。
- 王思聪是王健林的儿子，对应的句法结构为：王思聪（SBV），是（Root），王健林（ATT），儿子（VOB）

其中，SBV 是主谓关系，VOB 是动宾关系，ATT 是定中关系。

相关资源包括：

- lemonhu/open-entity-relation-extraction: Knowledge triples extraction and knowledge base construction based on dependency syntax for open domain text.
- aoldoni/tetre: TETRE: a Toolkit for Exploring Text for Relation Extraction
- gabrielStanovsky/template-oie: Extract templated Open Information Extraction

基于机器学习

使用基本步骤如下：

- （通常在一个句子中）寻找实体对
- 判断实体对之间是否存在关系
- 送到分类器判断关系的类别（预先定义好的）是什么

标准流程：

- 预先定义好类别集合
- 选择相关实体集合
- 标注
- 设计特征
- 训练分类器
- 评估结果

特征：

- 词相关
- 词法相关

- 句法相关
- 实体相关

之前那篇笔记里涉及的比较全面，而且现在几乎都是结合深度学习模型做了，这块就不再赘述了。

相关资源：

- machinalis/iepy: Information Extraction in Python
- marcolagi/quantulum: Python library for information extraction of quantities from unstructured text

基于深度学习

一般包括两种做法：Pipeline 和 Joint model，前者就是把实体识别和关系分类分开；后者一起做。

特征一般是基于 Word embedding，Position embedding，POS，NER，WordNet；模型一般都是基于 CNN，RNN。

- 端到端目前最好的是基于 Bert 的，在此之前，最好的是 Wang et al. 2016 的 Multi-Attention CNN。
- 关系分类最好的是 (Cai et al., 2016) 的 BRCNN（Bidirectional Recurrent Convolutional Neural Network）。

从论文的趋势看，端到端相对主流一些，不过对于我们的需求来说，关系分类更适合些。更多相关论文和模型可以进一步阅读 NLP-progress/relationship_extraction，这里就不贴那些论文的东西了。

基于半监督

半监督是利用少量高精度的 pattern 种子或种子 tuple 来 bootstrap 一个分类器。具体而言，在大规模语料中查找包含已有 pattern 实体对的句子，然后提取和归纳实体的上下文来学习新的 pattern。

还是举个栗子，比如我们有一个种子 tuple：（Queen，创作了，波西米亚狂想曲），然后可能找到了这些句子：

- 波西米亚狂想曲是由 Queen 演唱的歌曲。
- 波西米亚狂想曲是 Queen 最伟大的作品之一。
- Queen 这首将近 6 分钟的波西米亚狂想曲包括四个部分。

进而可以提取出类似这样的一些 pattern：

- （艺人，演唱，歌曲）
- （歌曲，是，艺人，作品）
- （艺人，作品，包括）

这些 pattern 又可以进一步寻找新的 pattern（把艺人和歌曲替换掉）。最终算法如下：

```
function BOOTSTRAP(Relation R) returns new relation tuples
 tuples←Gather a set of seed tuples that have relation R
 iterate
  sentences←find sentences that contain entities in tuples
  patterns←generalize the context between and around entities in sentences
  newpairs←use patterns to grep for more tuples
  newpairs←newpairs with high confidence
  tuples←tuples + newpairs
 return tuples
```

Bootstrapping 系统会给新的 pattern 一个置信度以避免语义飘移。比如 “在演唱会现场粉丝的要求下，周杰伦不得不演唱了一首网络歌曲《学猫叫》”，（周杰伦，演唱，学猫叫）显然不是我们想要的。关于置信度的计算可以参考上面提到的笔记，对一个 pattern 主要考量两方面因素：pattern 在现有 tuple 上的 hits 和在整个 Documents 上的 finds。

基于远程监督

远程监督从大规模数据库中获取的大量种子中产生出许多带噪声的 pattern features，然后用一个分类器组合这些 pattern。

举个栗子，比如要学习 “出生地-出生日期” 关系，半监督的方法中，我们可能只有少量几个启动的种子，但是现在我们可以在 Wikipedia-based 数据库（比如 DBPedia 或 Freebase）中获取大量包含 “出生地-出生日期” 的 tuple，比如（`<Albert Einstein, Ulm>, <Hubble, Marshfield>` 等等）。然后用命名实体工具提取包含两个实体并 match 给定 tuple 的句子：

- Hubble 出生于 Marshfield
- Einstein，生于 1879，Ulm
- Hubble 的出生地是 Marshfield

可以从中提取训练集，一个训练实例对应一个（关系，实体1，实体2）。

- `<出生地, Edwin Hubble, Marshfield>`
- `<出生地, Albert Einstein, Ulm>`
- `<出生日期, Albert Einstein, 1879>`

接下来可以用基于特征的分类器或直接使用神经网络分类器（不需要构建特征）。对于前者，可以从多个方面构建特征，比如实体 label，实体间单词、路径，相邻词等。每个 tuple 包括多个训练实例的特征，每个实例又可以从多个句子中获取词法和句法特征。最终的算法如下：

```
function DISTANT SUPERVISION(Database D, Text T) returns relation classifier C
 foreach relation R
  foreach tuple (e1,e2) of entities with relation R in D
   sentences←Sentences in T that contain e1 and e2
   f←Frequent features in sentences
   observations←observations + new training tuple (e1, e2, f, R)
 C←Train supervised classifier on observations
 return C
```

最终的分类器将会根据特征发现关系。

基于无监督

无监督关系提取的目的就是在没有标注数据，甚至没有任何关系列表的情况下从 Web 或大规模语料中提取关系。这个任务一般叫 open information extraction 或 Open IE，关系通常都是几个单词（常以动词开头）。

ReVerb 系统从一个句子中提取关系一般包括四步：

- 在句子上进行 POS 和实体识别。
- 对句中每个动词，找到以动词开头并满足句法和词汇约束（合并相邻匹配项）的最长单词序列 w。
- 对每个短语 w，找到最左边的名词短语 x（不是相对代词，wh-单词或 “there”），在右边找到最近的名词短语 y。
- 使用置信度分类器（一个逻辑回归分类器）给关系 r=x, w, y) 一个置信度。

分类器是在 1000 个随机选择的句子上训练所得，首先提取关系，然后人工标注是否正确，最后训练分类器。使用到的一些特征如下（将提取到的关系及周围的词作为特征）：

```
(x,r,y) covers all words in s
the last preposition in r is for
the last preposition in r is on
len(s) ≤ 10
there is a coordinating conjunction to the left of r in
s
r matches a lone V in the syntactic constraints
there is preposition to the left of x in s
there is an NP to the right of y in s
```

小结

方法	优点	缺点
模板	精准高，领域内可定制	召回低，耗时耗力
句法分析	构建简单	召回低，与句法结果相关
机器学习	数据相关时精准较高	特征工程较复杂，数据标注成本较高，训练数据敏感
深度学习	数据相关时精准高，泛化能力较好	数据标注成本很高，训练数据敏感
半监督 Bootstrapping	成本低，可以发现新关系	对初始种子敏感，语义飘移，准确率低
远程监督	精准高，训练数据不敏感，无语义飘移	依赖已有数据库
无监督	成本很低，容易实现	需转为权威格式存储，动词为中心的局限性

比赛

比赛最有名的大概就是 SemEval 2018 Task 10 和 SemEval-2010 Task 8 了。前者是一个二分类任务，目的是识别给定属性能否区分两个给定的概念。

ATTRIBUTE	CONCEPT1	CONCEPT2	LABEL
bookcase	fridge	wood	1

ATTRIBUTE	CONCEPT1	CONCEPT2	LABEL
bucket	mug	round	0
angle	curve	sharp	1
pelican	turtle	water	0
wire	coil	metal	0

后者是关系分类任务，给定两个标记的 nominals，预测它们的关系和关系的方向。

> There were apples, **pears** and oranges in the **bowl**.

(content-container, pears, bowl)

数据集

除了上面的两个比赛的数据集，还有以下一些数据集：

- FewRel: A Large-Scale Supervised Few-shot Relation Classification Dataset with State-of-the-Art Evaluation
  - 70K sentences
  - 100 relations
  - Wikipedia corpus
- The New York Times Annotated Corpus - Linguistic Data Consortium
  - Stanford NER 提取实体
  - 自动与 Freebase knowledge base 中的实体关联
  - 关系也是基于 Freebase knowledge base 中的事实
- Stanford TACRED Homepage
  - 106,264 examples
  - newswire and web text from TAC KBP Comprehensive English Source Corpora 2009-2014 - Linguistic Data Consortium
  - 41 relation types as used in the TAC KBP challenges

评价方法

评价指标还是以 F1 为主：

- 属性判别是二分类任务，直接用 F1 评估。
- 关系分类使用 Macro-averaged F1（9 个关系，不包括 OTHER，考虑关系方向）。
- FewRel 主要是在 5-way-1-shot，5-way-5-shot，10-way-1-shot 和 10-way-5-shot 几个任务上的准确率。
- NYT 主要使用 N 个结果的精准率或 PR 图。
- TACRED 使用 Micro-averaged F1（不包括 no_relation type）。

更多资源

CrowdTruth Corpus 相关：

- CrowdTruth/CrowdTruth-core: CrowdTruth framework for crowdsourcing ground truth for training & evaluation of AI systems
- CrowdTruth/Open-Domain-Relation-Extraction: Crowdsourced data for open domain relation classification from sentences

- CrowdTruth/Medical-Relation-Extraction: Crowdsourced ground truth for medical relation extraction.

其他资源：

- roomylee/awesome-relation-extraction: A curated list of awesome resources dedicated to Relatio of the most important tasks in Natural Language Processing (NLP).
- NLP-progress/relationship_extraction.md at master · sebastianruder/NLP-progress

参考资料：

- **知识抽取-实体及关系抽取 - 徐阿衡 知乎**
- **知识图谱入门 (三) 知识抽取 - pelhans 的博客**

明天有「超级超级重磅」的文章放出噢！小伙伴们记得提前做好准备鸭(￣▽￣)
