无需人工！无需训练！构建知识图谱 BERT一下就行了！

原创 Sherry 夕小瑶的卖萌屋 2020-12-07 22:20

文：Sherry

今天给大家带来的是一篇号称可以自动建立知识图谱的文章《Language Models are Open Knowledge Graphs》，文中提出了一个叫Match and Map（MAMA）的模型，无需人工！无需训练！只需语料和预训练好模型，就可以从头建立出知识图谱，甚至可以挖掘出人类发现不了的新关系。当Wikipedia再次邂逅BERT，知识图谱就诞生啦！

通常来说知识图谱的建立需要人工定义好的关系或者是实体类别，然后基于这些我们称之为schema的骨架进行建立整个图谱。而传统的自动识别关系及实体的方法大都基于训练。而MAMA就不一样了，它就像妈妈一样可以帮我们实现全自动图谱建立：

- 不需要人工定义的schema，而是依靠开放实体抽取和开放关系抽取的方法去建立图谱。
- 不需要在开放关系抽取或者实体抽取的任务上训练，而仅仅依靠预训练模型就可以完成建立知识图谱的整个过程。
- 模型不针对一个单一的关系逐条分析，一次喂给MAMA整个段落，她就回报给你所有triple

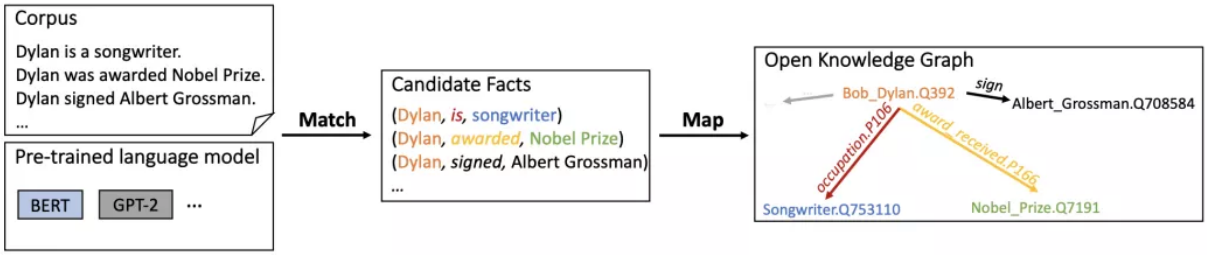
到底是怎么做到的呢?

开放知识图谱

想要建立MAMA，我们先回顾一下知识图谱中都有哪些基本元素：（熟悉知识图谱的同学们可以跳过这部分）知识图谱，我们想要把大量的非结构化的知识（一般是大量的网页及其中的文本）转化成结构化的图结构，那我们的基本结构中既要有知识也要有图。 目前，知识图谱中一共储存两类知识：一类是实体，一般是诸如人名地名这类的名词；另外一类是这些实体之间的关系，比如出生地，职

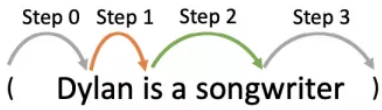
业。 有了知识，我们只需要把它建立成图结构，那么把实体看成图中的点，关系看成图中的边就可以了。 开放知识图谱一般用三元组（起始实体，关系，结束实体）来表示边，所有边都被以这个形式储存之后图谱就建立好啦。

MAMA怎样构建图谱呢?



要构建知识图谱第一步是获取基本原料：一个清洗好的语料库和一个预训练模型。 文中直接采用了维基百科作为语料，预训练模型则直接用发布的模型就可以了。

接下来关键的一步是自动抽取三元组，也是本文的主要贡献点。 实体抽取的技术已经相对成熟，给定一个语料中的段落，我们先用开源工具抽取出它的所有实体，来构成我们可能建立的关系候选。我们按照他们在句子中出现的顺序，分为头实体和尾实体。然后重点来了！我们利用BERT这类预训练模型的注意力权重来提取实体间的关系。



Step	Action	Intermediate candidates	Matching degrees
0	START	(Dylan,	0
1	YIELD	(Dylan, is	0.3
2	YIELD	(Dylan, is songwriter	0.7
3	STOP	(Dylan, is, songwriter)	0.7

(a) Matching example.

(b) Attention matrix for matching degree calculation.

对于一个（头实体，尾实体）对，我们用Beam search的方法从一个头实体出发生成一个到尾实体的序列。比如图中从Dylan出发，以songwriter结束。对于每一位置，我们看注意力权重矩阵里attend到这个实体的这一列，并且只关注在句子中当前位置之后的token的注意力权重，选择权重最大的下一个token加入当前序列。例子中从Dylan出发选择了is这个token，然后重复之前的操作，下一个我们选到了songwriter，那么搜索结束，我们就得到了一个（Dylan,is, songwriter）的序列。聪明的

小伙伴们已经发现了，这样提取出来的序列不就是我们想要的三元组吗？没错！我们再加上一些修修补补，MAMA就可以为我们完成构建图谱的工作啦！

按上面这样选出来的序列虽然可以简要表示我们所需要的信息，但它还不是严格意义上的关系三元组——我们有可能提取出多个token作为关系，文中针对这个问题对关系提取加入了一些限制：

- 首先，我们只保留注意力权重和大于阈值的序列。这是为了防止BERT这类模型单纯地提取出符合语言模型的序列，而不是那些对实体有特殊意义的关系。

> 一个反例:在阈值筛选之前，模型会从句子 Rolling Stone wrote: “No other pop song has so thoroughly challenged artistic conventions” 中抽取关系(Rolling Stone, wrote, pop song)

- 提取出来的关系必须在整个语料中出现足够多的次数。这样是为了防止出现一些过于细节偏门的关系。

> 例如 (Dylan, signed to Sam Peckinpah’s film, Pat Garrett and Billy the Kid)，这里的关系特指签约了Sam Peckinpah的电影，非常罕见且缺乏泛化性。

- 关系序列必须是句子中出现的连续token。这样可以防止提取出没有意义的关系。

> 例如(Rolling Stone, wrote challenged, conventions)，这里wrote 和chanllanged不表示合理的关系。

现在，我们就已经可以用MAMA从语料库中建立一个知识图谱啦！

MAMA效果如何?

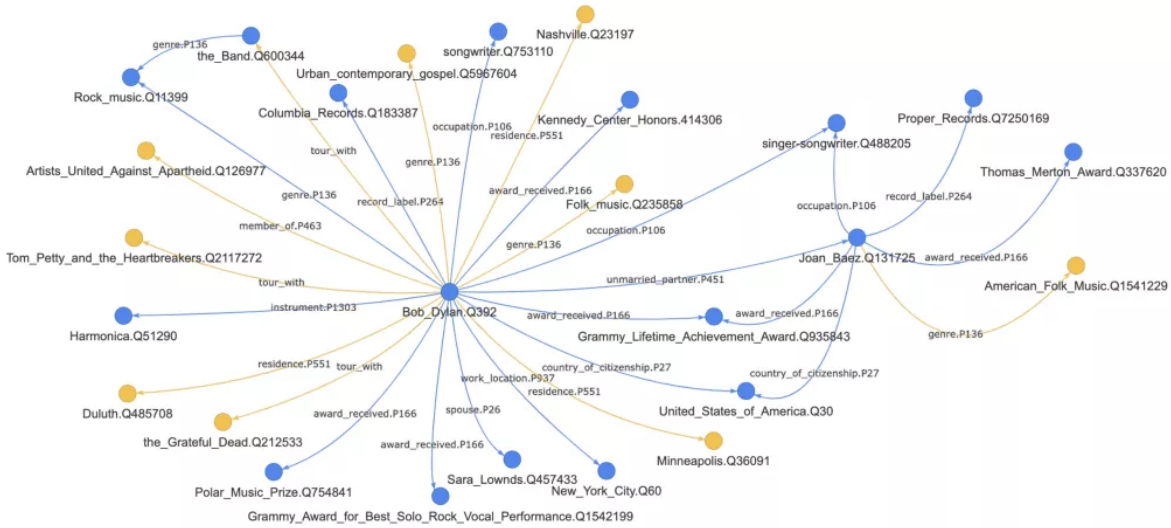
为了方便和其他方法比较，我们需要把这个开放图谱和已有的数据集对应上。使用已经比较成熟的实体链接，关系映射方法就可以了。

Method	Precision%	Recall%	F1%
OpenIE 5.1 [2]	56.98	14.54	23.16
Stanford OpenIE (Angeli et al., 2015)	61.55	17.35	27.07
MAMA-BERT$_{\text{BASE}}$ (ours)	61.57	18.79	28.79
MAMA-BERT$_{\text{LARGE}}$ (ours)	61.69	18.99	29.05
MAMA-GPT-2 (ours)	61.62	18.17	28.07
MAMA-GPT-2$_{\text{MEDIUM}}$ (ours)	62.10	18.65	28.69
MAMA-GPT-2$_{\text{LARGE}}$ (ours)	62.38	19.00	29.12
MAMA-GPT-2$_{\text{XL}}$ (ours)	62.69	19.47	**29.72**

Table 2: Compare the quality of mapped facts on TAC KBP.

这样造出来的MAMA无论在准确率还是召回率上都超过了之前的方法。

除了那些可以被对应到人造数据集中的关系之外，MAMA的一大亮点在于她可以发现其他没有被schema预先定义的关系：



图中蓝色的关系是在预定义schema中出现的部分，MAMA额外还生成了33%的新关系（黄色）。其中像Dylan和其他歌手曾经合作过，曾经是某个乐队的成员等，这样的信息是人工schema中所没有的，但对于歌手来说却是很重要。如果可以自动完善知识图谱和schema的构建，那就解决了KG中很难穷尽所有关系的难题了。

一些评价

个人认为，MAMA的整体思想还是很新颖且值得借鉴的。但是实验部分以及一些细节上的设置还需要更精细的设置。一大缺陷在于他没有和其他的SOTA进行比较，效果尚未可知。总体来说，为自动化的知识图谱构建提供了一个不错的思路。

论文链接：

https://arxiv.org/pdf/2010.11967.pdf

讲解视频：

https://www.youtube.com/watch?v=NAJOZTNkhII&t=276s

萌屋作者：Sherry。

本科毕业于复旦数院，转行NLP目前在加拿大滑铁卢大学读CS PhD。经历了从NOIer到学数学再重回CS的转变，却坚信AI的未来需要更多来数学和自认知科学的理论指导。主要关注问答，信息抽取，以及有关深度模型泛化及鲁棒性相关内容。

作品推荐：

1. Google Cloud TPUs支持Pytorch框架啦！

后台回复关键词【**入群**】

加入卖萌屋NLP/IR/Rec与求职讨论群

有顶会审稿人、大厂研究员、知乎大V和妹纸

等你来撩哦~

喜欢此内容的人还喜欢

我不看好data2vec这类多模态融合的研究

夕小瑶的卖萌屋