基于关系优先的三元组知识抽取

庄传志[1,2],张宁豫[3,4],邓淑 敏[3,4],余海 阳[3,4],叶宏斌[3,4],陈华钧[3,4], 靳小龙[1;2]

1中国科学院网络数据科学与技术重点实验室, 中国科学院计算技术研究所

2 中国科学院大学计算机与控制学院

3浙江 大学

4 AZFT知识引擎实验室

摘要:

三元组知识抽取是信息抽取的重要组成部分，在知识库构建等领域具有重要作用。大多数抽取方法都是首先抽取实体或实体对，而这往往导致实体对组合时的冗余问题，且很难处理重叠难题。针对这些问题，本文提出一种关系优先的抽取框架，并基于该框架提出了**基于关系优先的知识抽取模型RFTE**。该模型首先对文本中包含的关系进行分类，然后将预测的关系类型和文本拼接进行基于序列标注的实体识别以抽取该关系对应的实体对，最后将抽取的实体和关系类别进行组合得到三元组知识。我们的模型解耦了不同关系类型的实体识别，显著降低了搜索空间带来的影响。实验结果显示，该模型在三元组知识抽取数据集NYT、WebNLG、SKE上都取得较好的效果。

关键词：实体识别, 关系分类,三元组抽取,知识图谱

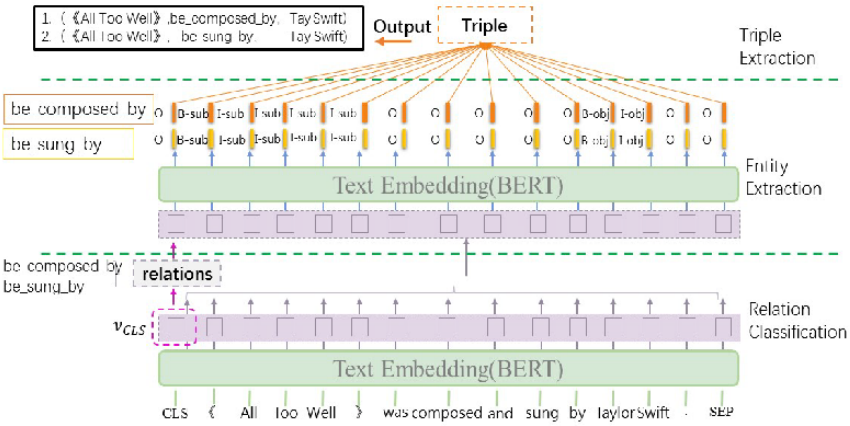

图2. 基于关系优先的RFTE模型

简介:

三元组知识抽取可以从海量无结构文本中识别出实体以及实体对之间的关系，也称为三元组抽取或关系抽取。例如：在句子“芝加哥位于美国的中西部”中，可以抽取出（芝加哥, 坐落于,美国）这样的知识，其中“芝加哥”和“美国”分别称为三元组的头实体（subject, 记为sub）和尾实体（object, 记为obj）。

三元组抽取框架:

总的来说，先前的三元组抽取框架可分为如下3类:

1.先实体对后关系（见图1(a)）：先实体对后关系是指先通过序列标注抽取实体，然后通过实体对之间的组合进行关系判别。

2.先头实体，后关系、尾实体（见图1(b)）：先头实体，后关系、尾实体是指先通过序列标注抽取头实体，然后通过拼接头实体的特征同时抽取关系和尾实体。

3.同时抽取实体对和关系（见图1(c)）：同时抽取实体对和关系指先通过序列标注获取每个单词的关系及实体标签信息，然后通过特定的规则对各类标签信息进行组合，以同时抽取关系和尾实体。

4.在真实场景中，我们发现关系的数量通常远小于头、尾实体的数量，此外文本中存在大量的不存在关系的句子，这启发我们提出第4类三元组知识抽取框架，即解耦最大似然函数的方式。该方法先抽取关系，然后再同时抽取头、尾实体，以实现嵌套三元组知识的抽取(见图1(d))。先抽取关系事实上可以极大的缩小模型的搜索空间，且解耦了不同关系之间的实体抽取。

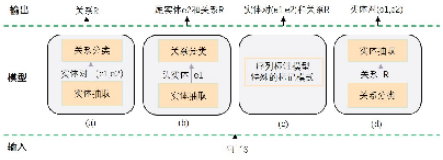

图1. 三元组抽取模型类型(a)先抽取实体对，再进行关系分类；(b)先抽取头实体e1,再抽取关系R对应的尾实体e2；(c)同时抽取关系和实体对；(d)我们提出的框架：先抽取关系R，再抽取关系R对应的实体对

RFTE模型:

在基于关系优先的框架下，我们提出的抽取模型如图2所示，模型由关系分类层、实体识别层和三元组抽取层组成。关系分类层对句子可能存在的关系进行分类抽取，实体识别层根据关系分类层的抽取结果对句子中的单词（或字）进行标注，在三元组抽取层对序列标注层得到的结果进行三元组知识组合并输出抽取得到的结果。

关系分类:

在关系分类层中，[CLS]标签对应的输出向量h_CLS，用来对整个句子中的关系类型信息进行编码,并将其放入一个全连接网络和sigmoid函数中。如下式所示：输出R的维度为关系类别的个数。

$$R = sigmoid\left(W_0\left(\tanh\left(h_{CLS}\right)\right) + b_0\right)$$

实体识别:

实体识别层的输入是关系分类层的输出和原始的文本。我们关系分类层得到的标签与原始文本拼接，得到{[CLS], relation, [SEP], w1, w2, ..., w_L}，并输入实体识别层。对输入序列的对应的任一向量h_i，我们将其放入另一个全连接网络和softmax函数中。

$$E = softmax\left(W_1\left(\tanh\left(h_i\right)\right) + b_1\right)$$

三元组抽取:

在关系分类模块和实体识别模块，我们可以获取候选的关系及实体，在本层我们设计的规则对其进行组合。具体的说，对于任意文本和给定的关系，我们从第一个头实体开始向后遍历，依次将尾实体和头实体、关系组合成三元组，当遇到下一个头实体时候，重新开始遍历并继续进行三元组组合直到句子末尾。

表1. 不同模型的主要结果对比

Model	NYT			WebNLG			SKE		
	P	R	F	P	R	F	P	R	F
Tagging	52.6	33.6	41.0	47.6	18.6	26.7	34.7	15.8	22.0
CopyR	60.2	55.6	57.8	38.1	36.9	37.5	40.2	36.2	38.0
HRL	74.1	65.1	69.3	69.5	62.9	66.0	58.2	42.2	48.9
MrMep	77.9	76.6	77.1	69.4	77.0	73.0	61.1	56.7	58.8
CasRel	**89.7**	89.5	89.6	**93.4**	90.1	91.8	-	-	-
RFTE	88.9	**90.5**	**89.7**	91.3	**92.5**	**91.9**	**81.1**	**80.2**	**80.6**

表2. 加入优化策略后的模型结果比对

Model	NYT			WebNLG			SKE		
	P	R	F	P	R	F	P	R	F
RFTE	88.9	90.5	89.7	91.3	92.5	91.9	81.1	80.2	80.6
RFTE+DA	89.0	90.8	89.9	91.5	92.8	92.1	**82.1**	**81.2**	**81.6**
RFTE+R	89.0	90.9	89.9	91.5	92.8	92.1	81.2	80.5	80.8
RFTE+DA+R	**89.8**	**91.0**	**90.4**	**91.8**	**92.9**	**92.3**	81.3	80.9	81.2

两种优化策略:

1.数据增强(Data Augmentation, DA):首先，我们对训练集中文本的实体替换为同类型的其他实体进行数据增广，并基于增广后的训练集训练模型。

2.规则约束（Rule, R）：我们统计不同类型实体为某关系的概率，并计算其共现统计值来构造后验约束规则，从而优化三元组解码。

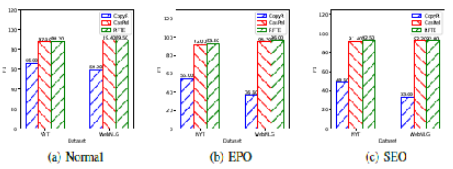

图4.模型在不同嵌套类型句子上的表现

实验结果:

从表1和表2中得出，我们的实验效果甚至略微好于目前最优的模型CasRel，并且显著优于HRL等基线模型。我们发现这两种优化策略进一步提升了模型的效果，这证明了实体替换数据增强和后验约束解码的有效性。

我们进一步对不同嵌套类型的句子进行了扩展实验。如图4所示，可以看出，CopyR模型在Normal，EPO和SEO三种类型上的表现均呈下降趋势，反映出从具有不同重叠模式的句子中提取相对三元组的难度越来越大。由于数据集中Normal样本较多，EPO和SEO的样本较少，且具有显著特征，CasRel模型和RFTE模型很容易在预训练模型的作用下取得比Normal更好的结果。但相比较于CasRel模型，RFTE模型在所有三个情况都取得更高的结果。

表3. 错误实例

致谢:

感谢国家自然科学基金项目（91846204, 61772501和91646120），SQ2018YFC000004, 2018YFB1402800，GF创新研究项目和阿里巴巴藏经阁（知识引擎）搜索计划对本研究的支持与资助。