基于标签校正的端到端实体关系联合抽取

陈鹏之 张瑾 刘悦 程学旗
中国科学院计算技术研究所，北京，100190

论文摘要

实体关系抽取是信息抽取领域中的重要研究课题之一。近几年，随着深度学习的快速发展与应用，联合式抽取被广泛应用于实体抽取和实体间的关系预测。虽然端到端的联合抽取方法在该领域得到了较大关注，但这类方法目前未考虑multi-token实体；同时，抽取过程中忽略了关系预测与实体抽取之间的相互影响。针对以上问题，该文结合Encoder-Decoder框架的特点，引入标签校正机制，提出了一种基于标签校正的端到端实体关系联合抽取方法CopyLC。实验结果证明，在更严格的评价方式下，该文所提出的方法与当前主流方法相比，在NYT和WebNLG数据集上均能获得更好的抽取效果。

论文简介

基于端到端的联合抽取方法主要存在两个问题：（1）实体抽取过程中，该类方法默认每个实体仅由一个单词构成，模型没有抽取出完整的实体；（2）该类方法抽取出的实体对准确率低，在抽取时，关系预测和实体抽取会相互影响，使得关系预测不准确，进而导致抽取的三元组准确率低。

针对上述问题，本文提出一种基于标签校正的联合抽取方法，称为CopyLC。在解码阶段使用四个权重矩阵，配合Attention、Copy、Mask机制，分别抽取主、客实体的首尾，以此抽取完整的实体。在解码完成时，通过标签序列对抽取的实体进行校正。最后，将以上两部分结合，构建一个多任务学习模型。

实验分析

本文的实验主要是为了验证本文提出的实体关系抽取方法在NYT和WebNLG数据集上的效果。

模型	NYT 数据集		
	P	R	F1
CopyRE-One(ours)	0.611	0.532	0.568
CopyRE-Mul(ours)	0.610	0.567	0.587
CopyLC-One	0.625	0.558	0.590
CopyLC-Mul	**0.637**	**0.574**	**0.603**

模型	WebNLG 数据集		
	P	R	F1
CopyRE-One(ours)	0.313	0.274	0.292
CopyRE-Mul(ours)	0.329	0.276	0.300
CopyLC-One	0.334	0.273	0.300
CopyLC-Mul	**0.358**	**0.295**	**0.323**

从实验结果可以看出，本文的方法在P、R及F1上均有提升。为了分析标签校正对抽取效果的影响程度，本文设计了消融实验。

模型	NYT 数据集		
	P	R	F1
CopyLC-One	**0.625**	**0.558**	**0.590**
CopyLC-One -LC	0.607	0.548	0.576
CopyLC-Mul	**0.637**	**0.574**	**0.603**
CopyLC-Mul -LC	0.614	0.549	0.580

模型	WebNLG 数据集		
	P	R	F1
CopyLC-One	**0.334**	**0.273**	**0.300**
CopyLC-One -LC	0.310	0.254	0.280
CopyLC-Mul	**0.358**	**0.295**	**0.323**
CopyLC-Mul -LC	0.326	0.277	0.299

通过消融实验，本文发现标签校正模块对抽取效果有提升，三元组的P、R及F1值均得到提升。为了进一步探究标签校正的具体影响，本文在CopyLC-One模型上进行实验，统计模型在关系预测和实体对（主、客实体共同构成）抽取上的效果，如下所示。

模型	属性	NYT 数据集		
		P	R	F1
CopyLC	关系	**0.897**	**0.802**	**0.847**
	实体对	**0.724**	**0.647**	**0.683**
CopyLC -LC	关系	0.885	0.799	0.839
	实体对	0.698	0.630	0.662

模型	属性	WebNLG 数据集		
		P	R	F1
CopyLC	关系	**0.836**	**0.714**	**0.770**
	实体对	**0.637**	**0.506**	**0.564**
CopyLC -LC	关系	0.820	0.672	0.739
	实体对	0.618	0.506	0.556

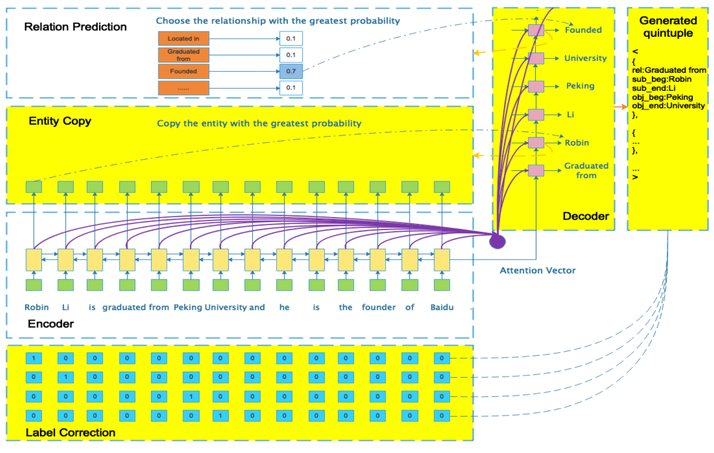
实验结果表明标签校正对关系预测和实体抽取的效果均有提升。本文认为原因是模型采用了Encoder-Decoder架构，标签校正使实体抽取更准确，抽取出的实体会影响下一个三元组中的关系预测，使关系预测更准确，反之，关系预测也会影响实体的抽取。

论文结论

针对当前基于端到端的联合抽取方法存在抽取中未考虑实体multi-token以及抽取过程中关系预测与实体抽取相互影响等问题，本文结合Encoder-Decoder框架的特点，引入标签校正机制，提出了一种基于标签校正的实体关系联合抽取方法。从实验结果来看，本文提出的方法，在更严格的评价方法下，相比较当前主流方法在NYT和WebNLG数据集上均能获得更好的抽取效果。

系统模型

模型的框架如下图所示，包含三个主要模块。
1)**Encoder**：通过双向LSTM对输入语句进行编码。
2)**Decoder**：负责关系预测和实体抽取。实体抽取阶段使用Copy、Attention、Mask机制抽取主、客实体的首尾。
3)**Label Correction**：使用四个序列分别校正抽取出的主、客实体首尾。



算法原理

实体抽取：

使用Copy机制替代固定词表生成，从输入文本中分别将两个实体抽取出来。本文用两步来抽取一个实体，第一步，找到实体首部，第二步，找到实体尾部，每两步可以抽取出任意长度实体。利用Encoder的输出与Decoder的输出得到输入句子中每个单词的表达向量 ，进而得到主、客实体首尾的表达向量。例如计算主实体首位置，使用softmax得到主实体首部的概率分布，选择概率值最大的位置上对应的单词作为主实体的开头，下一步需要抽取出主实体的结尾位置，抽取方法与抽取主实体开始位置类似。主实体的结尾位置必须位于主实体开始位置之后，为了保证抽取的实体有意义，本文使用Mask机制：

$$M_i^{sub} = \begin{cases} 1, i \geq j \\ 0, i < j \end{cases}$$

通过Mask机制可以计算主实体结尾位置的表达向量。类似的，本文分别设计了另外三种Mask机制辅助抽取客实体首尾。

标签校正：

当模型抽取出多对实体关系三元组后，使用标签校正模块校正抽取出的实体。用正确的实体对模型抽取出的实体进行校正。本文用四个序列分别记录主实体首尾、客实体首尾位置的标签。在输入句子中，将主实体开始对应的位置标记为1，其余位置标记为0。同理，用三个序列分别标注并表示主实体结尾、客实体开始、结尾的标签序列，最后用四个权重矩阵拟合。

西安电子科技大学 XIDIAN UNIVERSITY