知识库丰富的神经关系提取

Bayu Distiawan Trisedya1，Gerhard Weikum2，齐建中 1，张瑞 1 *

1 1 澳大利亚墨尔本大学

2 德国萨尔信息大学马克斯·普朗克信息学研究所

{btrisedya @ student，jianzhong.qi @，rui .zhang @} unimelb.edu.au weikum@mpi-inf.mpg.de

摘 要

我们研究了用于知识库（KB）丰富化的关系提取。 具体来说，我们旨在从三元组形式的句子中提取实体及其关系，并以端到端的方式将提取的三元组的元素映射到现有的 KB。以前的研究集中在提取本身上，并依赖于命名实体歧义消除（NED）将三元组映射到 KB 空间中。 这样，NED 错误可能会导致提取错误，从而影响整体精度和召回率。 为了解决这个问题，我们提出了一种基于神经编码器-解码器模型的知识库丰富的端到端关系提取模型。 我们通过远距离监督，共参考解析和副词短语检测来收集高质量的训练数据。 我们提出了一种基于 n 元语法的注意力模型，该模型捕获句子中的多词实体名称。 我们的模型采用共同学习的单词和实体嵌入来支持命名实体消歧。 最后，我们的模型使用改进的波束搜索和三元分类器来帮助生成高质量的三元组。 在两个真实数据集上，我们的模型在 F1 得分方面分别比最新基准高出 15.51％和 8.38％。

1 简介

知识库（KBs）通常以知识图（KGs）的形式出现，已成为许多任务中必不可少的资源，包括问答系统，推荐系统和自然语言生成。大型 KB，例如 DBpedia（Auer 等，2007），Wikidata（Vrandecic 和 Krotzsch，2014）和 Yago（Suchanek 等，2007）包含数以百万计的有关实体的事实，这些事实以主题的形式表示。 谓语-宾语三元组。 但是，这些知识库还远远不够完整，无法连续地进行丰富和管理。

Input sentence:
"New York University is a private university in Manhattan."
Unsupervised approach output:
⟨NYU,is,private university⟩ ⟨NYU,is private university in,Manhattan⟩
Supervised approach output:
⟨NYU, instance of, Private University⟩ ⟨NYU, located in, Manhattan⟩
Canonicalized output:
⟨Q49210, P31, Q902104⟩ ⟨Q49210, P131, Q11299⟩

表 1：关系提取示例。

以前的研究工作是基于嵌入的模型（Nguyen 等人，2018; Wang 等人，2015）和实体对齐模型（Chen 等人，2017; Sun 等人，2017; Trisedya 等人，2019） 丰富的知识库。 继序列到序列体系结构(Bahdanau 等人,2015)成功用于从结构化数据生成句子(Marcheggiani 和 Perez-Beltrachini，2018; Trisedya 等人，2018）之后，我们采用了这种架构 相反，从句子中提取三元组。

在本文中，我们研究如何通过从文本源中提取关系来丰富知识库。 具体来说，我们旨在提取 hh，r，ti 形式的三元组，其中 h 是头部实体，t 是尾部实体，r 是实体之间的关系。重要的是，由于 KB 在实体上的覆盖范围通常比关系上的覆盖范围要好得多，因此我们假设 h 和 t 是 KB 中的现有实体，r 是一个谓词，它属于我们感兴趣的预定义谓词集合，但是关系 hh， r，ti 在 KB 中尚不存在。 我们旨在找到 h 和 t 之间的更多关系，并将它们添加到 KB 中。 例如，从表 1 中第一个提取的三元组中，我们可以识别出 KB 中已经存在的两个实体“ NYU”（纽约大学的缩写）和“私人大学”。 谓词“ instance of”也位于我们感兴趣的预定义谓词集中，但是 KB 中不存在 hNYU，Private Universityi 实例的关系。 我们旨在将此关系添加到我们的知识库中。这是 KB 富集的典型情况（与从头开始构建 KB 或出于其他目的（例如 Q&A 或摘要）执行关系提取相反。

KB 富集要求通过将提取的三元组的实体和关系映射到它们的适当实体和 KB 中的谓词 ID 来规范化实体和关系。 表 1 举例说明了从句子中提取的三元组。 提取的第一个三元组的实体和谓词（包括 NYU，私立大学的实例和私立大学）分别映射到其唯一的 ID Q49210，P31 和 Q902104，以符合 KB 的语义空间。

以前有关关系提取的研究都采用了无监督和有监督的方法。 无监督方法通常从一小组手动定义的提取模式开始，以检测实体名称和与输入文本中的关系有关的短语。 这种范例被称为开放信息提取（Open IE）（Banko 等，2007; Corro and Gemulla，2013; Gashteovski 等，2017）。 在这种方式中，实体和谓词都以其表面形式捕获而没有规范化。监督方法训练统计和神经模型来推断句子中两个已知实体之间的关系（Mintz 等人，2009; Riedel 等人，2010,2013; Zeng 等人，2015; Lin 等人，2016）。 这些研究大多数采用预处理步骤来识别实体。 只有很少的研究通过对现有 KB 进行逻辑推理来消除提取的实体的歧义，将提取的三元组映射完全整合到唯一标识的 KB 实体上（例如（Suchanek 等人，2009; Sa 等人，2017））。

因此，大多数现有方法都需要将命名实体歧义消除（NED）（请参见 Shen 等人（2015）的调查）作为一个单独的处理步骤。 另外，将关系短语映射到 KB 谓词上需要另一个映射步骤，通常需要借助复述词典来进行。 这种两阶段体系结构固有地倾向于在其两个阶段之间传播错误：NED 错误可能导致提取错误（反之亦然），从而导致将错误的关系添加到 KB 中。

我们的目的是通过提出一种端到端的神经学习模型来联合提取和规范化任务，以从句子中联合提取三元组并将其映射到现有的知识库中。我们的方法基于编码器-解码器框架（Cho 等人，2014），将任务视为将句子翻译成三元组元素序列的任务。 对于表 1 中的示例，我们的模型旨在将“纽约大学是曼哈顿的一所私立大学”转换为 ID 为“ Q49210 P31 Q902104 Q49210 P131 Q11299”的序列，从中我们可以得出要添加的两个三元组 KB。

但是，带有注意力的标准编解码器模型（Bahdanau 等人，2015）无法捕获表示谓词的多字实体名称和口头或名词短语。 为了解决这个问题，我们提出了一种新的基于 n-gram 的注意力形式，该形式计算注意力集中的 n gram 组合来捕获补充了注意力模型中标准单词级别注意力的言语或名词短语语境。 因此，我们的模型可以更好地捕获实体和关系的多词上下文。 我们的模型利用了预训练的单词和实体嵌入，它们是通过 skip gram（Mikolov 等，2013）和 TransE（Bor des 等，2013）共同学习的。 我们共同学习的嵌入的优点是双重的。 首先，嵌入物捕获单词和实体之间的关系，这对于命名实体消除歧义至关重要。 其次，实体嵌入保留了实体之间的关系，这有助于建立一个高度准确的分类器来过滤无效的提取三元组。 为了解决缺少完全标记的训练数据的问题，我们采用遥远的监督来生成对齐的句子对和三重作为训练数据。 我们通过共同引用分辨率（Clark 和 Manning，2016）和基于字典的释义检测（Ganitkevitch 等，2013； Grycner 和 Weikum，2016）来增强该过程。 共参考解析有助于提取具有隐式实体名称的句子，这会使候选句子的集合扩大到与 KB 中现有的三元组对齐。 复述检测有助于过滤不表达实体之间任何关系的句子。

本文的主要贡献是：

•我们提出了一种端对端模型，用于提取和规范化三元组以丰富 KB。该模型减少了关系提取和 NED 之间的错误传播，而现有方法则容易发生这种错误传播。

•我们提出了一个基于 n-gram 的注意力模型，以有效地将实体及其关系的多词提及映射到唯一标识的实体和谓词中。 我们建议联合学习单词和实体嵌入，以捕获单词和实体之间的关系，以消除命名实体的歧义。 我们还提出了一种改进的波束搜索和三元分类器，以生成高质量的三元组。

•我们在两个现实世界的数据集上评估提出的模型。 我们将遥远的超视力与辅助参考分辨率和副词短语检测相结合，以获取高质量的训练数据。 实验结果表明，我们的模型始终优于强大的神经关系提取基准（Lin 等人，2016）以及最新的 NED 模型（Hoffart 等人，2011; Kolitsas 等人） 。，2018）。

2 相关工作

2.1 开放信息提取 Banko 等。

（2007 年）介绍了开放信息提取（Open IE）的范例，并提出了一个包括三个阶段的管道：学习者，提取者和评估者。 学习者使用依赖项解析信息以无监督的方式学习提取模式。 前拖拉机通过将名词短语标识为自变量并将连接短语标识为谓词来生成候选三元组。评估者根据统计证据为每个候选三元组分配一个概率。这种方法易于提取不正确的，冗长的和无意义的三元组。 各种后续研究（Fader 等人，2011； Mausam 等人，2012； Angeli 等人，2015； Mausam，2016）通过添加手工制作的图案或使用遥距模式证明了 Open IE 的准确性。 监督。Corro 和 Gemulla（2013）开发了 ClausIE，这是一种分析句子中的从句并从该结构中得出三元组的方法。 Gashteovski 等。 （2017）开发了 MinIE，通过使生成的三元组更加简洁来推进 ClausIE。

Stanovsky 等。 （2018）通过将关系提取引入序列标记中，为 Open IE 提出了一个受监督的学习者。 对 bi-LSTM 模型进行了训练，以预测输入的每个标记的标签（实体，谓词或其他）。 与我们最相关的工作是 Neural Open IE（Cui 等人，2018），它提出了一种具有注意力模型的编码器/解码器来提取三元组。 但是，这项工作并不适合于提取规范化实体的关系。 另一类研究使用神经学习进行语义角色标记（He 等，2018），但此处的目标是识别单个输入句子的谓词-论元结构，而不是从语料库中提取关系。

所有这些方法都会生成三元组，其中头尾实体和谓词保持其表面形式。 因此，相同实体的不同名称和短语会导致多个三元组，如果以这种方式添加，将会污染 KG。 将三元组映

射到 KG 中唯一标识的实体的唯一方法是通过实体链接（NED）方法进行后处理（Shen 等人，2015）或与后续映射进行聚类（Gal′arraga 等人，2014）。

2.2 实体感知关系提取

受 Brin（1998）的启发，最先进的方法通过利用现有 KG 的老化种子事实利用远程监督（Mintz 等，2009； Suchanek 等， 2009； Carlson 等，2010）。这些方法从种子事实中学习提取模式，将这些模式应用于提取新的事实候选者，重复此原理，最后使用统计推断（例如分类器）来减少误报率。 其中一些方法基于这样一个假设，即在同一句子中种子事实的实体的同时出现是表达实体之间语义关系的指示。 这是潜在的错误提示的来源。 后续研究（Hoffmann 等人，2010； Riedel 等人，2010，2013； Surdeanu 等人，2012）通过各种方式克服了这一局限性，包括使用特定于关系的词典和潜在因子模型。 尽管如此，这些方法仍通过实体的表面形式来处理实体，而忽略了它们到 KG 中现有实体的映射。

Suchanek 等。 （2009）和 Sa 等。 （2017）使用概率逻辑推理分别基于约束解决方案或概率图形模型的蒙特卡洛抽样来消除误报。 这些方法将实体链接（即 NED）集成到其模型中。 但是，两者都具有很高的计算复杂度，并且都依赖于建模约束和适当的先验条件。

最近的研究使用神经网络来学习三元组的提取。

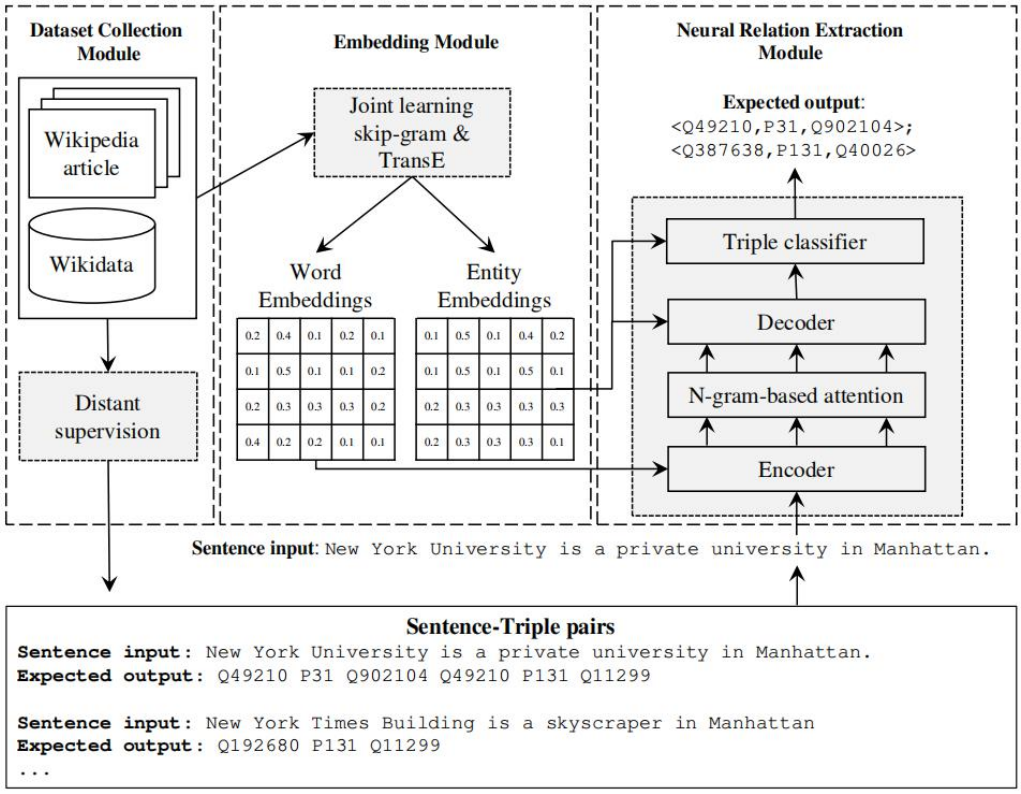

图 1：我们提出的解决方案概述。

Nguyen and Grishman（2015）提出了具有多尺寸窗口内核的卷积网络。 Zeng 等。（2015 年）提出了分段卷积神经网络（PCNN）。 Lin 等。 （2016，2017）通过提出 PCNN 并在句子层面给予关注来改进此方法。 该方法在实验研究中效果最好； 因此，我们选择它作为我们比较方法的主要基准。 后续研究考虑了进一步的变化：Zhou 等。 （2018）提出了等级注意，Ji 等。（2017）合并了实体描述，Miwa 和 Bansal（2016）合并了句法功能，而 Sorokin 和 Gurevych（2017）使用了背景知识进行语境化。 这些神经模型都不适合 KG 富集，因为实体的规范化不在其范围之内。

这些神经模型都不适合 KG 富集，因为实体的规范化不在其范围之内。

3 提议的模型

我们从问题定义开始。 令 G =（E，R）是现有的 KG，其中 E 和 R 分别是 G 中的实体和关系（谓词）的集合。 我们将句子 S = hw1，w2，...，wii 作为输入，其中 wi 是句子中位置 i 处的标记。 我们旨在从句子中提取一组三元组 O = {o1，o2，...，oj}，其中 oj = hhj，rj，tj i，hj，tj∈E 和 rj∈R。表 1 说明了我们问题的输入和目标输出。

3.1 解决方案框架

图 1 说明了总体解决方案框架。 我们的框架由三个组件组成：数据收集模块，嵌入模块和神经关系提取模块。

在数据收集模块中（在第 3.2 节中有详细介绍），我们将现有知识库中的已知三元组与包含文本语料库中此类三元组的句子对齐。 对齐的句子和三元组对将在以后的神经关系提取模块中用作训练数据。这种协调是通过远程监督来完成的。为了获得大量高质量的比对，我们使用共参考分辨率扩展了处理过程，以提取具有隐式实体名称的句子，从而扩大了要比对的候选句子的集合。 我们进一步使用基于字典的复述检测来过滤不表达实体之间任何关系的句子。

在嵌入模块（在第 3.3 节中有详细介绍）中，我们提出了一种结合单词-实体嵌入的联合学习方法，方法是结合跳跃语法（Mikolov 等人，2013 年）以计算单词嵌入和 TransE（Bordes 等人，2013 年）。 （2013 年）。 联合学习的目的是捕获单词和实体的相似性，以帮助将实体名称映射到相关的实体 ID 中。 此外，结果实体嵌入用于训练三元分类器，该分类器有助于过滤由我们的神经关系提取模型生成的无效三元组。

在神经关系提取模块中（在第 3.4 节中有详细介绍），我们通过将注意力机制扩展到句子的 n-gram 标记，提出了一个基于 n-gram 的注意力模型。 n gram 注意会计算注意权重的 n-gram 组合，以捕获补充标准注意模型的单词级别注意的言语或名词短语上下文。 这种

扩展有助于我们的模型更好地捕获实体和关系的多词上下文。

编码器-解码器模型的输出是实体和谓词 ID 的序列，其中每个三个 ID 表示一个三元组。为了生成高质量的三元组，我们提出了两种策略。 第一种策略使用修改后的波束搜索，该算法将提取的实体的词汇相似性与输入句子中实体名称的表面形式相结合，以确保正确的实体谓词。 第二种策略使用三元分类器，该分类器使用来自联合学习的实体嵌入进行训练，以过滤无效的三元组。 在 3.5 节中详细介绍了三元组生成过程。

3.2 数据集收集

我们的目的是通过提出一种有监督的关系提取模型，从句子中提取三元组以丰富知识库。 为了训练这样的模型，我们需要大量的带有完整标签的训练数据，这些数据以句子三重对的形式出现。 根据 Sorokin 和 Gurevych（2017）的研究，我们使用遥远的监督（Mintz 等，2009）将 Wikipedia1 中的句子与 Wikidata2 中的三元组对齐（Vrandecic 和 Króotzsch，2014）。

我们通过与实体提及相关联的超链接将句子中的实体提及映射到 Wikidata 中的相应实体条目（即 Wikidata ID），该超链接被记录在 Wikidata 中作为实体条目的 url 属性。 每对可能包含一个句子和多个三元组。 我们根据谓词复述的顺序对三元组的顺序进行排序，这些谓语复述表示句子中实体之间的关系。我们通过提取包含 Wikidata 三元组的头和尾实体的句子来收集句子三重对。 为了生成高质量的句子三元组对，我们提出了两个附加步骤：（1）使用共引用解析来提取包含隐式实体名称的句子，以及（2）使用释义过滤不表达任何关系的句子 检测。 我们在下面详细介绍这些步骤。

在将句子与三元组对齐之前，在步骤（1）中，我们找到隐式实体名称以增加要对齐的候选句子的数量。 我们将共同引用解决方案（Clark 和 Manning，2016）应用于维基百科文章中的每个段落，并将提取的共同引用替换为适当的实体名称。 我们观察到，维基百科文章中段落的第一句可能包含代名词，指代主要实体。 例如，有奥巴马的文章在准图形与句子开始“他再次当选伊利诺伊州州参议员，1998 年”。 这可能会导致标准的共同引用解析丢失该段落其余部分的隐式实体名称。 为了解决这个问题，如果未提及 Wikipedia 页面的主要实体名称，我们会用粗心大意地替换段落第一句中的代词。 对于先前示例中的句子，我们将“ He”替换为“ Barack Obama”。 直觉是，Wikipedia 文章包含单个感兴趣实体的内容，并且段落第一句中提到的代词大多与主要实体有关。

在步骤（2）中，我们使用基于字典的复述检测来捕获句子中实体之间的关系。首先，我们通过从三个来源(包括PATTY (Nakashole 等人)、2012)，保利(Grycner 和 Weikum, 2016)，

和 PPDB (Ganitkevitch 等人，2013)产生 540 个谓词和 24013 个独特释义。

	#pairs	#triples	#entities	#predicates
All (WIKI)	255,654	330,005	279,888	158
Train+val	225,869	291,352	249,272	157
Test (WIKI)	29,785	38,653	38,690	109
Test (GEO)	1,000	1,095	124	11

表 2：数据集的统计信息。

例如，关系“出生地”的谓词释义是{出生于，出生于，...}。然后，我们使用该词典过滤不表达实体之间任何关系的句子。 我们使用精确的字符串匹配来查找句子中的语言或名词短语，这是三元组谓词的复述。 例如，对于三胞胎 h 巴拉克·奥巴马（Barack Obama）的出生地檀香山，将保留句子“巴拉克·奥巴马（Barack Obama）1961 年出生于夏威夷的檀香山”，而句子“巴拉克·奥巴马（Barack Obama）于 2010 年访问檀香山”将被删除（该句子可能 如果还有另一个有效的三胞胎巴拉克·奥巴马（Barack Obama），请访问檀香山（Honolului）。 这有助于过滤噪声以使句子三重对齐。

收集的数据集包含 255,654 个句子三重对。 对于每对，三元组的最大数量为四个（即，一个句子最多可以产生四个三元组）。 我们将数据集分为训练集（80％），开发集（10％）和测试集（10％）（我们将其称为 WIKI 测试数据集）。 为了进行压力测试（以与培训数据不同的文本样式测试建议的模型），我们还在 Wikipedia 之外收集了另一个测试数据集。我们对旅行网站的用户评论采用相同的程序。 首先，我们收集有关澳大利亚 100 个热门地标的用户评论。 然后，我们将经过调整的远程监督应用于评论，并收集了 1,000 对三元组对（我们将其称为 GEO 测试数据集）。 表 2 总结了我们的数据集的统计信息。

3.3 单词和实体嵌入的联合学习

我们的关系提取模型基于编码器-解码器框架，该框架已在神经机器翻译中广泛用于将文本从一种语言翻译为另一种语言。 在我们的设置中，我们的目标是将句子翻译成三元组，因此源输入的词汇是一组英语单词，而目标输出的词汇是现有 KG 中的一组实体和谓词 ID。为了计算源词汇和目标词汇的嵌入量，我们提出了单词和实体嵌入的联合学习方法，该方法可有效捕获单词和实体之间的相似性，以实现命名实体歧义消除（Yamada et al。，2016）。请注意，我们的方法与 Yamada 等人的方法不同。 （2016）。 我们使用联合学习的方法是结合 skip-gram（Mikolov 等人，2013）来计算单词嵌入，使用 TransE（Bordes 等人，2013）来计算实体嵌入（包括关系嵌入），而 Yamada 等人 。 （2016 年）使用基于 Wikipedia

的基于链接的度量（WLM）（Milne 和 Witten，2008 年），该度量未考虑关系嵌入。

我们的模型通过最小化基于边距的目标函数 J_E 来学习实体嵌入：

$$J_E = \sum_{t_r \in T_r} \sum_{t'_r \in T'_r} \max\left(0, \left[\gamma + f(t_r) - f(t'_r)\right]\right) \quad (1)$$

$$T_r = \{\langle h, r, t\rangle | \langle h, r, t\rangle \in G\} \quad (2)$$

$$T_r{}' = \left\{\langle h', r, t\rangle \mid h' \in E\right\} \cup \left\{\langle h, r, t'\rangle \mid t' \in E\right\} \quad (3)$$

$$f(t_r) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\| \quad (4)$$

此处，||x||是向量 x 的 L1-范数，γ 是边距超参数，Tr 是来自 KG G 的有效关系三元组的集合，Tr′是腐败关系三元组的集合 是 G 中的实体集）。 损坏的三元组用作负样本，它是通过用随机实体替换 Tr 中有效三元组的头或尾实体而创建的。 我们使用 Wiki 数据中的所有三元组（属于测试数据的三元组）来计算实体嵌入。

为了建立实体与单词嵌入之间的交互作用，我们遵循 Yamada 等人提出的 Anchor Context Model。 （2016）。

首先，我们通过结合原始文本和经修改的 Wikipedia 锚文本来生成文本语料库。 这是通过将句子中的实体名称替换为相关的实体或谓词 ID 来完成的。 例如，句子“纽约大学是曼哈顿的一所私立大学”被修改为“ Q49210 是 Q11299 中的 Q902104”。 然后，我们使用 skip-gram 方法从生成的语料库中计算单词嵌入（修改后的锚文本中的实体 ID 被视为 skip-gram 模型中的单词）。给定 n 个单词的序列[w1，w2，...，wn]，该模型通过最小化以下目标函数 JW 来学习单词嵌入：

$$J_W = \frac{1}{T} \sum_{t=1}^{n} \sum_{-c \le j \le c, j \ne 0} \log P\left(w_{t+j} | w_t\right) \quad (5)$$

$$P\left(w_{t+j} | w_t\right) = \frac{\exp\left(\mathbf{v}'_{w_{t+j}}{}^{\top} \mathbf{v}_{w_t}\right)}{\sum_{i=1}^{W} \left(\mathbf{v}'_i{}^{\top} \mathbf{v}_{w_t}\right)} \quad (6)$$

其中 c 是上下文窗口的大小，wt 表示目标词，wt + j 是上下文词； vw 和 v0w 是单词 w 的输入和输出矢量表示形式，W 是词性大小。 单词和实体嵌入的联合学习的总体目标功能是：

$$J = J_E + J_W \quad (7)$$

3.4 基于 N 元语法的注意力模型

我们提出的关系提取模型以端到端的方式集成了 KB 富集的提取和规范化任务。 为了建立这样的模型，我们采用编码器-解码器模型（Cho 等人，2014）将句子翻译成三元组序列。 编码器将句子编码为向量，解码器将其用作上下文以生成三元组序列。因为我们将输入和输出视为顺序，所以我们在编码器和解码器中使用 LSTM 网络（Hochreiter 和 Schmidhuber，1997）。

具有注意力模型的编码器/解码器（Bahdanau 等人，2015）已 用于机器翻译中。 但是，在关系提取任务中，注意力模型无法捕获多字实体名称。 在我们的初步调查中，我们发现注意模型在单词和实体之间产生了不对齐的情况。

上面的问题是由于不同实体名称中的单词相同（例如，纽约大学，华盛顿大学等不同大学名称中的大学一词）所致。 在训练期间，模型会更加注意“大学”一词，以区分名称相似的不同类型的实体，例如，纽约大学，纽约时报大楼或纽约人寿大厦，但不会区分名称不同的相同类型的实体 （例如，纽约大学和华盛顿大学）。 这可能会导致实体对齐方式出错，尤其是在预测不在训练数据中的实体的 ID 时。 即使我们将 hEntity-name 和 Entity-IDi 对添加为训练数据（请参阅“训练”部分），仍然会发生对齐错误。

我们通过提出一个基于 n-gram 的注意力模型来解决上述问题。 该模型使所有可能的 n-gram 句子输入引起注意。 注意权重被计算在词嵌入的 n-gram 组合上，因此用于解码器的上下文向量如下计算。

$$\mathbf{c}_t^d = \left[\mathbf{h}^e; \sum_{n=1}^{|N|} \mathbf{W}^n \left(\sum_{i=1}^{|X^n|} \alpha_i^n \mathbf{x}_i^n \right) \right] \qquad (8)$$

$$\alpha_i^n = \frac{\exp(\mathbf{h}^{e\top} \mathbf{V}^n \mathbf{x}_i^n)}{\sum_{j=1}^{|X^n|} \exp(\mathbf{h}^{e\top} \mathbf{V}^n \mathbf{x}_j^n)} \qquad (9)$$

此处，c^d_t 是解码器在时间步 t 的上下文向量，h^e 是编码器的最后一个隐藏状态，上标 n 表示 n 元语法组合，x 是输入句子的词嵌入，$|\ X^n\ |$ 是 n-gram 令牌组合的总数，N 表示 n-gram 组合中使用的 n 的最大值（在我们的实验中为 N = 3），W 和 V 是学习的参数矩阵，α 是关注权重。

训练

在训练阶段，除了使用远距离监督收集的句子三对（请参阅第 3.2 节）外，我们还将训

练集中的 KB 中所有实体的(Entity-name，Entity-ID)对添加到训练数据中，例如 ，(纽约大学，Q49210)。 这允许模型学习实体名称和实体 ID 之间的映射，尤其是对于看不见的实体。

3.5 三重生成

编码器-解码器模型的输出是实体和谓词 ID 的序列，其中每三个标记表示一个三重。 因此，要提取一个三元组，我们只需将生成的输出的每三个标记分组即可。 但是，由于实体嵌入（例如，纽约市的嵌入）之间的相似性，贪婪方法（即，选择解码器的最后一个 softmax 层中概率最高的实体）可能导致模型提取错误的实体 和芝加哥可能相似，因为它们都是美国的城市。 为了解决这个问题，我们提出了两种策略：使用改进的波束搜索对预测实体进行重新排序，以及使用三元分类器过滤无效的三元组。

Model		WIKI			GEO		
		Precision	Recall	F1	Precision	Recall	F1
Existing Models	MinIE (+AIDA)	0.3672	0.4856	0.4182	0.3574	0.3901	0.3730
	MinIE (+NeuralEL)	0.3511	0.3967	0.3725	0.3644	0.3811	0.3726
	ClausIE (+AIDA)	0.3617	0.4728	0.4099	0.3531	0.3951	0.3729
	ClausIE (+NeuralEL)	0.3445	0.3786	0.3607	0.3563	0.3791	0.3673
	CNN (+AIDA)	0.4035	0.3503	0.3750	0.3715	0.3165	0.3418
	CNN (+NeuralEL)	0.3689	0.3521	0.3603	0.3781	0.3005	0.3349
Encoder-Decoder Models	Single Attention	0.4591	0.3836	0.4180	0.4010	0.3912	0.3960
	Single Attention (+pre-trained)	0.4725	0.4053	0.4363	0.4314	0.4311	0.4312
	Single Attention (+beam)	0.6056	0.5231	0.5613	0.5869	0.4851	0.5312
	Single Attention (+triple classifier)	0.7378	0.5013	0.5970	0.6704	0.5301	0.5921
	Transformer	0.4628	0.3897	0.4231	0.4575	0.4620	0.4597
	Transformer (+pre-trained)	0.4748	0.4091	0.4395	0.4841	0.4831	0.4836
	Transformer (+beam)	0.5829	0.5025	0.5397	0.6181	0.6161	0.6171
	Transformer (+triple classifier)	0.7307	0.4866	0.5842	0.7124	0.5761	0.6370
Proposed	N-gram Attention	0.7014	0.6432	0.6710	0.6029	0.6033	0.6031
	N-gram Attention (+pre-trained)	0.7157	0.6634	0.6886	0.6581	0.6631	0.6606
	N-gram Attention (+beam)	0.7424	**0.6845**	0.7123	0.6816	**0.6861**	0.6838
	N-gram Attention (+triple classifier)	**0.8471**	0.6762	**0.7521**	**0.7705**	0.6771	**0.7208**

表 3：实验结果。

修改后的波束搜索重新排列由解码器通过计算实体名称（从 KB 获得）和的每个 n-gram 标记之间的编辑距离来预测的 top-k（在我们的实验中，k = 10）实体 ID。 输入的句子。 直觉是应在句子中提及实体名称，以便将相似度最高的实体选作输出。

我们的三分类器接受了来自联合学习的实体嵌入训练（请参见第 3.3 节）。 三重分类是评估实体嵌入质量的指标之一（Socher 等人，2013）。 我们建立一个分类器来确定三元(h，r，t)的有效性。 我们根据似然性分数（h + r - t）（计算实体嵌入的分数）训练二元分类器。 我们通过破坏有效的三元组来创建负样本（即用随机实体替换首尾实体）。 三元组分类器可有效过滤有效的三元组，例如(曼哈顿首都曼哈顿的纽约大学)。

4 实验

我们在包括 WIKI 和 GEO 测试数据集在内的两个真实数据集上评估我们的模型（请参阅第 3.2 节）。 我们使用精度，召回率和 F1 分数作为评估指标。

4. 1 超参数

我们使用网格搜索来找到网络的最佳超参数。 我们对编码器和解码器都使用 512 个隐藏单元。 我们使用 64 个维度的预训练单词和实体嵌入（请参见第 3.3 节）。 我们在编码器和解码器上都使用 0.5 的丢失率进行正则化。 我们使用亚当（Kingma and Ba，2015）的学习率为 0.0002。

4. 2 模型

我们将我们提出的模型 [3] 与三个现有模型进行了比较，包括 CNN（Lin 等人（2016）的最新监督方法），MiniE（Gashteovski 的最新无监督方法） 等人（2017），以及 Corro 和 Gemulla（2013）的 ClausIE。 为了通过这些模型映射提取的实体，我们使用了两个最先进的 NED 系统，包括 AIDA（Hoffart 等，2011）和 NeuralEL（Kolitsas 等，2018）。 AIDA 和 NeuralEL 的精度（在我们的测试数据集上测试）分别为 70％和 61％。 为了映射无监督方法输出的提取谓词（关系），我们使用基于字典的释义检测。 我们使用用于收集数据集的同一词典（即，三个释义词典的组合，包括 PATTY（Nakashole 等人，2012），POLY（Grycner 和 Weikum，2016）和 PPDB（Gan itkevitch 等人。 ，2013））。 如果正确谓词的释义之一（即黄金标准）出现在抽取谓词中，则我们用正确谓词 ID 替换抽取谓词。否则，我们将提取的谓词替换为“ NA”以表示无法识别的谓词。 我们还将 N-gram 注意模型与两个基于编码器/解码器的模型进行了比较，包括单注意模型（Bahdanau 等，2015）和 Transformer 模型（Vaswani 等，2017）。

4. 3 结果

表 3 显示端到端模型优于现有模型。 特别是，我们提出的 n-gram 注意模型在准确性，召回率和 F1 得分方面均取得了最佳结果。我们提出的模型在 WIKI 和 GEO 测试数据集上的 F1 评分分别优于最佳现有模型（MinIE）33.39％和 34.78％。 由于现有模型受 NED 误差传播的影响，因此可以得到这些结果。 不出所料，现有模型与 AIDA 的组合比与 NeuralEL 的组合具有更高的 F1 分数，因为 AIDA 的精度高于与 NeuralEL 的组合。

为了进一步显示错误传播的影响，我们建立了一个没有规范化任务的实验（即目标是预测已知实体之间的关系）。 通过允许 CNN 模型访问正确的实体，我们删除了 NED 预处理步骤。 同时，我们向我们提出的模型的解码器提供正确的实体。 在这种设置下，我们提出的模型在 WIKI 和 GEO 测试数据集上的精度分别达到 86.34％和 79.11％，而 CNN 的精度分别达到 81.92％和 75.82％。

我们提出的 n-gram 注意模型在 WIKI 和 GEO 测试数据集上的 F1 得分分别比端到端模型

高出 15.51％和 8.38％。 变形金刚模型也只产生与单注意力模型类似的性能，这比我们的模型差。 这些结果表明，我们的模型在输入句子中捕获的多词实体名称（在两个数据集中，有 82.9％的实体具有多词实体名称）比其他模型更好。

表 3 还显示了预训练的嵌入在所有方面均改善了模型的性能。 此外，预训练的嵌入有助于模型更快收敛。在我们的实验中，使用预训练嵌入的模型的平均收敛时间为 20 个纪元，而未使用预训练嵌入的模型的平均收敛时间为 30 - 40 个纪元。 我们的三分类器与改进的波束搜索相结合，提高了模型的性能。 改进的波束搜索通过基于输入句子中的表面形式提取正确的实体来提供较高的召回率，而三元分类器则通过过滤无效的三元组来提供高精度。

讨论

我们将进一步执行手动错误分析。 我们发现模型的错误输出是由两个不同实体的实体名称相同引起的（例如，迈克尔·乔丹（Michael Jordan）的名称指的是美国篮球运动员或英国足球运动员）。 修改后的波束搜索无法消除这些实体的歧义，因为它仅考虑词法相似性。我们考虑将基于上下文的相似性用作未来的工作。

5 结论

我们提出了一个 KB 富集的端到端关系提取模型，该模型集成了提取和规范化任务。 因此，我们的模型减少了现有方法容易产生的关系提取和 NED 之间的错误传播。 为了获得高质量的训练数据，我们采用了远程监控，并通过共参考分辨率和复述检测对其进行了增强。我们提出了一种基于 n 元语法的注意力模型，该模型可以更好地捕获句子中的多词实体名称。 此外，我们提出了一种改进的波束搜索和三重分类，可帮助模型生成高质量的三重。

实验结果表明，我们提出的模型在 WIKI 和 GEO 测试数据集上的 F1 评分分别优于现有模型 33.39％和 34.78％。 这些结果证实了我们的模型减少了 NED 和关系提取之间的误差传播。在两个真实数据集上，我们提出的 n-gram 注意模型在 F1 分数方面分别优于其他编码器-解码器模型 15.51％和 8.38％。 这些结果证实我们的模型更好地捕获了一个句子中的多单词实体名称。 将来，我们计划探索基于上下文的相似性，以补充词汇相似性，从而提高整体性能。

致谢

Bayu Distiawan Trisedya 得到了 Indosian 教育捐赠基金（LPDP）的支持。Bayu Distiawan Trisedya 即将访问马克斯·普朗克数学研究所。 这项工作得到了澳大利亚研究委员会（ARC）发现项目 DP180102050，谷歌院系研究奖和中国国家科学基金会（项目编号 61872070 和编号 61402155）的支持。

参考文献

略。