

# 知识图谱数据构建的“硬骨头”，阿里工程师如何拿下？

（原创） 2018-03-14 游维 阿里技术



阿里妹导读：搜索“西红柿”，你不但能知道它的营养功效、热量，还能顺带学会煲个牛腩、炒个鸡蛋！搜索引擎何时变成“暖男”了？原来背后有“知识图谱”这个强大的秘密武器。

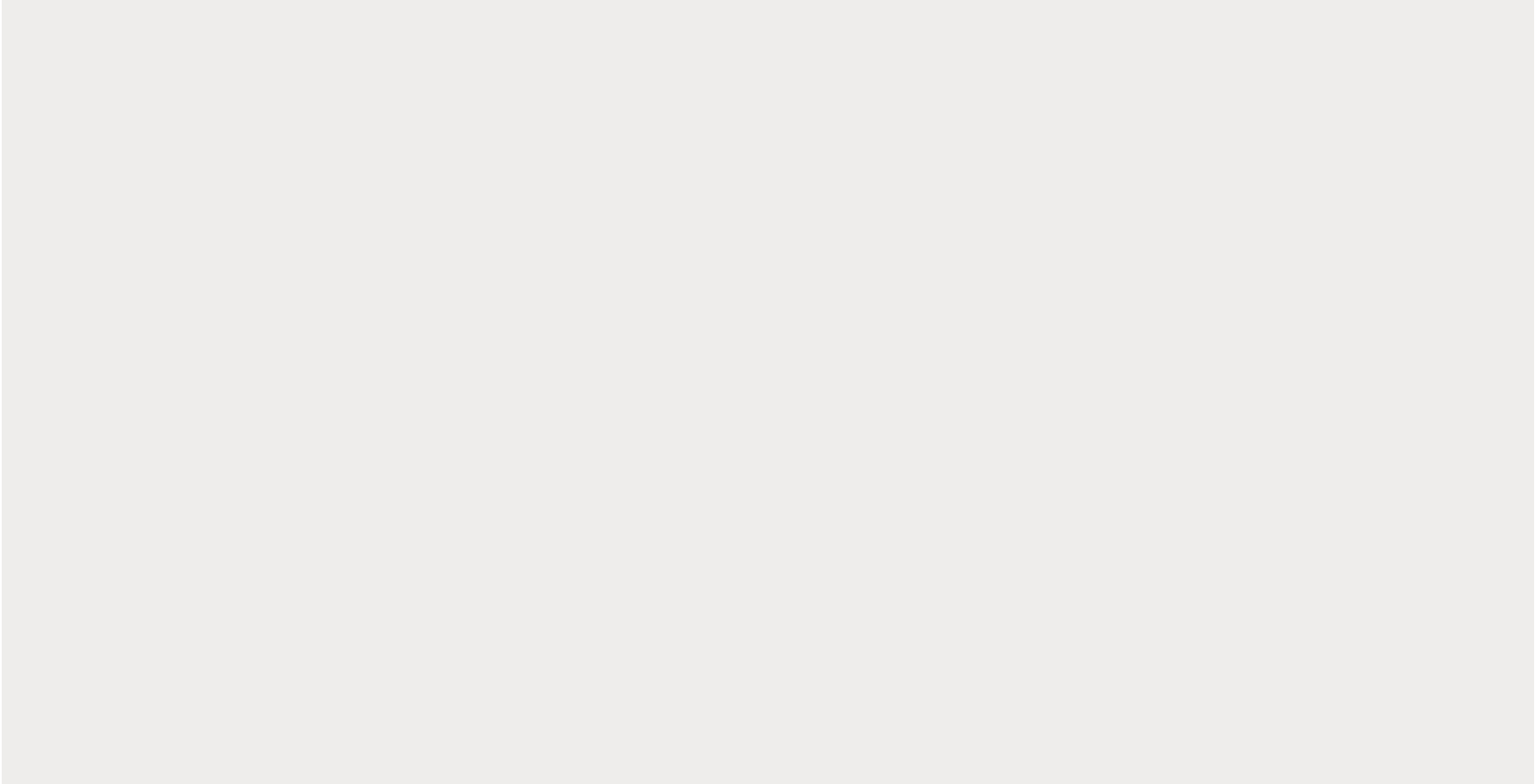
作为近年来搜索领域和自然语言处理领域的热点，知识图谱正引领着搜索引擎向知识引擎的转变。在阿里的“神马搜索”中，知识图谱及其相关技术的广泛应用不仅能帮助用户找到最想要的信息，更能让用户有意想不到的知识收获。

## 背景简介

为了不断提升搜索体验，神马搜索的知识图谱与应用团队，一直在不断探索和完善图谱的构建技

术。其中，开放信息抽取（Open Information Extraction），或称通用信息抽取，旨在从大规模无结构的自然语言文本中抽取结构化信息。它是知识图谱数据构建的核心技术之一，决定了知识图谱可持续扩增的能力。





## “神马搜索”界面

本文聚焦于开放信息抽取中的重要子任务——关系抽取，首先对关系抽取的各种主流技术进行概述，而后结合业务中的选择与应用，重点介绍了基于DeepDive的方法，并详述它在神马知识图谱数据构建工作中的应用进展。

## 关系抽取概述

### 关系抽取技术分类

现有的关系抽取技术主要可分为三种：

- **有监督的学习方法**：该方法将关系抽取任务当做分类问题，根据训练数据设计有效的特征，从而学习各种分类模型，然后使用训练好的分类器预测关系。该方法的问题在于需要大量的人工标注训练语料，而语料标注工作通常非常耗时耗力。
- **半监督的学习方法**：该方法主要采用Bootstrapping进行关系抽取。对于要抽取的关系，该方法首先手工设定若干种子实例，然后迭代地从数据中抽取关系对应的关系模板和更多的实例。
- **无监督的学习方法**：该方法假设拥有相同语义关系的实体对拥有相似的上下文信息。因此可以利用每个实体对对应上下文信息来代表该实体对的语义关系，并对所有实体对的语义关系进行聚类。

这三种方法中，有监督学习法因为能够抽取并有效利用特征，在获得高准确率和召回率方面更有优势，是目前业界应用最广泛的一类方法。



## 远程监督算法

为了打破有监督学习中人工数据标注的局限性，Mintz等人提出了远程监督（Distant Supervision）算法，该算法的核心思想是将文本与大规模知识图谱进行实体对齐，利用知识图谱已有的实体间关系对文本进行标注。远程监督基于的基本假设是：如果从知识图谱中可获取三元组 $R(E1, E2)$ （注： $R$ 代表关系， $E1$ 、 $E2$ 代表两个实体），且 $E1$ 和 $E2$ 共现与句子 $S$ 中，则 $S$ 表达了 $E1$ 和 $E2$ 间的关系 $R$ ，标注为训练正例。

远程监督算法是目前主流的关系抽取系统广泛采用的方法，也是该领域的研究热点之一。该算法很好地解决了数据标注的规模问题，但它基于的基本假设过强，会引入大量噪音数据。例如，从知识图谱获取三元组：创始人（乔布斯，苹果公司），下表句1和句2正确表达了该关系，但句3和句4并没有表达这样的关系，因此对句3和句4应用基本假设时会得到错误的标注信息。这个问题通常称为 the wrong label problem。



出现 the wrong label problem 的根本原因，是远程监督假设一个实体对只对应一种关系，但实际上实体对间可以同时具有多种关系，如上例中还存在CEO（乔布斯，苹果公司）的关系，实体对间也可能不存在通常定义的某种关系，而仅因为共同涉及了某个话题才在句中共现。

为了减小 the wrong label problem 的影响，学术界陆续提出了多种改进算法，主要包括：

- 基于规则的方法：通过对wrong label cases的统计分析，添加规则，将原本获得正例标注的wrong label cases直接标为负例，或通过分值控制，抵消原有的正标注。
- 基于图模型的方法：构建因子图（factor graph）等能表征变量间关联的图模型，通过对特征的学习和对特征权重的推算减小wrong label cases对全局的影响。
- 基于多示例学习（multi-instance learning）的方法：将所有包含（ $E1$ ， $E2$ ）的句子组成一个bag，从每个bag对句子进行筛选来生成训练样本。此类方法最早提出时假设如果知识图谱中存在 $R(E1, E2)$ ，则语料中含（ $E1$ ， $E2$ ）的所有instance中至少有一个表达了关系 $R$ 。一般与

无向图模型结合，计算出每个包中置信度最高的样例，标为正向训练示例。该假设比远程监督的假设合理，但可能损失很多训练样本，造成有用信息的丢失和训练的不充分。为了能得到更丰富的训练样本，又提出了multi-instance multi-labels的方法。该方法的假设是，同一个包中，一个sentence只能表示（E1，E2）的一种关系，也就是只能给出一个label，但是不同的sentence可以表征（E1，E2）的不同关系，从而得到不同的label。多label标注的label值不是正或负，而是某一种关系。它为同时挖掘一个实体对的多种关系提供了可能的实现途径。另一种改进的方法是从一个包中选取多个valid sentences作为训练集，一般与深度学习方法相结合，这种方法更详细的讲解和实现会安排在后续介绍深度学习模型的章节中。

## 神马知识图谱构建中的关系抽取方法选择

知识图谱的数据构建，就数据源而言，分为结构化数据，半结构化数据和无结构数据三类。其中，无结构数据是最庞大、最易获取的资源，同时也是在处理和利用方面难度最大的资源。神马知识图谱构建至今，已经发展为一个拥有近5000万实体，近30亿关系的大规模知识图谱。在经历了前期以结构化和半结构化数据为主的领域图谱构建阶段，神马知识图谱的数据构建重点已经逐渐转移为如何准确高效地利用无结构数据进行实体与关系的自动识别与抽取。这一构建策略使得神马知识图谱在通用领域的建设和可持续扩增方面有很强的竞争力。

远程监督算法利用知识图谱的已有信息，使得有监督学习中所需的大规模文本标注成为可能。一方面，远程监督在很大程度上提升了有监督学习关系抽取的规模和准确率，为大规模的知识图谱数据构建和补充提供了可能；另一方面，远程监督对现有知识图谱的数据和规模有较强的依赖，丰富的标注数据对机器学习能力的提升有很大帮助。为了充分利用知识图谱规模和远程监督学习这种相辅相成的特性，在神马知识图谱的现阶段数据构建业务中，我们采用了以图谱现有的大规模实体与关系数据为依托，以远程监督算法为工具的关系抽取技术。

在上一章的综述中，我们介绍过多种基于远程监督思想的改进方法。在具体的业务实现中，我们选取了领域内与业务需求最为契合的两种代表性方法：基于DeepDive的抽取系统和基于深度学习抽取算法。两种方法相辅相成，各有优势：DeepDive系统较多依赖于自然语言处理工具和基于上下文的特征进行抽取，在语料规模的选择上更为灵活，能进行有针对性的关系抽取，且能方便地在抽取过程中进行人工检验和干预；而深度学习的方法主要应用了词向量和卷积神经网络，在大规模语料处理和多关系抽取的人物中有明显的优势。在下面的章节中，我们来更详细地了解这两种方法的实现与应用。

## DeepDive系统介绍

### DeepDive概述

DeepDive (<http://deepdive.stanford.edu/>) 是斯坦福大学开发的信息抽取系统，能处理文本、表格、

图表、图片等多种格式的无结构数据，从中抽取结构化的信息。系统集成了文件分析、信息提取、信息整合、概率预测等功能。Deepdive的主要应用是特定领域的信息抽取，系统构建至今，已在交通、考古、地理、医疗等多个领域的项目实践中取得了良好的效果；在开放领域的应用，如TAC-KBP竞赛、维基百科的infobox信息自动增补等项目中也有不错的表现。

DeepDive系统的基本输入包括：

- 无结构数据，如自然语言文本
- 现有知识库或知识图谱中的相关知识
- 若干启发式规则

DeepDive系统的基本输出包括：

- 规定形式的结构化知识，可以为关系（实体1，实体2）或者属性（实体，属性值）等形式
- 对每一条提取信息的概率预测

DeepDive系统运行过程中还包括一个重要的迭代环节，即每轮输出生成后，用户需要对运行结果进行错误分析，通过特征调整、更新知识库信息、修改规则等手段干预系统的学习，这样的交互与迭代计算能使得系统的输出不断得到改进。

## DeepDive系统架构和 workflows

DeepDive的系统架构如下图所示，大致分为数据处理、数据标注、学习推理和交互迭代四个流程：



## 数据处理

### 1、输入与切分

在数据处理流程中，DeepDive首先接收用户的输入数据，通常是自然语言文本，以句子为单位进行切分。同时自动生成文本id和每个句子在文本中的index。doc\_id + sentence\_index 构成了每个句子的全局唯一标识。

### 2、NLP标注

对于每个切分好的句子，DeepDive会使用内嵌的Stanford CoreNLP工具进行自然语言处理和标注，包括token切分，词根还原、POS标注、NER标注、token在文本中的起始位置标注、依存语法分析等。

### 3、候选实体对提取

根据需要抽取的实体类型和NER结果，首先对实体mentions进行定位和提取，而后根据一定的配对规则生成候选实体对。需要特别注意，在DeepDive中，每一个实体mention的标定都是全局唯一的，由doc\_id、sentence\_index以及该mention在句子中的起始和结束位置共同标识。因此，不同位置出现的同名的实体对（E1，E2）将拥有不同的（E1\_id，E2\_id），最终的预测结果也将不同。

### 4、特征提取

该步骤的目的是将每一个候选实体对用一组特征表示出来，以便后续的机器学习模块能够学习到每个特征与所要预测关系的相关性。Deepdive内含自动特征生成模块DDlib，主要提取基于上下文的语义特征，例如两个实体mention间的token sequence、NER tag sequence、实体前后的n-gram等。Deepdive也支持用户自定义的特征提取算法。

## 数据标注

在数据标注阶段，我们得到了候选实体对以及它们对应的特征集合。在数据标注阶段，我们将运用远程监督算法和启发式规则，对每个候选实体对进行label标注，得到机器学习所需的正例和负例样本。

### 1、 远程监督

实现远程监督标注，首先需要从已知的知识库或知识图谱中获取相关的三元组。以婚姻关系为例，DeepDive从DBpedia中获取已有的夫妻实体对。若候选实体对能在已知的夫妻实体对中找到匹配映射时，该候选对标记为正例。负例的标注针对需要抽取的不同关系有不同的可选方法。例如可以将没有在知识库中出现的实体对标注为负例，但在知识库收入不完整的情况下该方法会引入噪音负例；也可以用知识库中互斥关系下的实例来做负例标注，例如父母-子女关系，兄弟姐妹关系，都与婚姻关系互斥，用于标注负例基本不会引入噪音。

### 2、 启发式规则

正负样本的标注还可以通过用户编写启发式规则来实现。以抽取婚姻关系为例，可以定义如下规则：

- Candidates with person mentions that are too far apart in the sentence are marked as false.
- Candidates with person mentions that have another person in between are marked as false.
- Candidates with person mentions that have words like "wife" or "husband" in between are marked as true.

用户可以通过预留的user defined function接口，对启发式规则进行编写和修改。

### 3、 Label冲突的解决

当远程监督生成和启发式规则生成的label冲突，或不同规则生成的label产生冲突时，DeepDive采用majority vote算法进行解决。例如，一个候选对在DBpedia中找到了映射，label为1，同时又满足2中第2条规则，得到label为-1，majority vote对所有label求和： $\text{sum} = 1 - 1 = 0$ ，最终得到的label为doubt。

## 学习与推理

通过数据标注得到训练集后，在学习与推理阶段，Deepdive主要通过基于因子图模型的推理，学习特征的权重，并最终得到对候选三元组为真的概率预测值。

因子图是一种概率图模型，用于表征变量和变量间的函数关系，借助因子图可以进行权重的学习和



边缘概率的推算。DeepDive系统中，因子图的顶点有两种，一种是随机变量，即提取的候选实体对，另一种是随机变量的函数，即所有的特征和根据规则得到的函数，比方两个实体间的距离是否大于一定阈值等。因子图的边表示了实体对和特征及规则的关联关系。

当训练文本的规模很大，涉及的实体众多时，生成的因子图可能非常复杂庞大，DeepDive采用吉布斯采样（Gibbs sampling）进行来简化基于图的概率推算。在特征权重的学习中，采用标准的SGD过程，并根据吉布斯采样的结果预测梯度值。为了使特征权重的获得更灵活合理，除了系统默认的推理过程，用户还可以通过直接赋值来调整某个特征的权重。篇幅关系，更详细的学习与推理过程本文不做展开介绍，更多的信息可参考DeepDive的官网。

## 交互迭代

迭代阶段保证通过一定的人工干预对系统的错误进行纠正，从而使得系统的准召率不断提升。交互迭代一般包括以下几个步骤：

### 1、准召率的快速估算

- 准确率：在P集中随机挑选100个，看为TP的比例。
- 召回率：在输入集中随机挑选100个positive case，看有多少个落在计算出的P集中。

### 2、错误分类与归纳

将得到的每个extraction failure（包括FP和FN）按错误原因进行分类和归纳，并按错误发生的频率进行排序，一般而言，最主要错误原因包括：

- 在候选集生成阶段没有捕获应捕获的实体，一般是token切分、token拼接或NER问题
- 特征获取问题，没能获取到区分度高的特征
- 特征计算问题，区分度高的特征在训练中没有获得相应的高分（包括正负高分）

### 3、错误修正

根据错误原因，通过添加或修改规则、对特征进行添加或删除、对特征的权重进行调整等行为，调整系统，重新运行修改后的相应流程，得到新的计算结果。

## 神马知识图谱构建中的DeepDive应用与改进

在了解了DeepDive的工作流程之后，本章将介绍我们如何在神马知识图谱的数据构建业务中使用DeepDive。为了充分利用语料信息、提高系统运行效率，我们在语料处理和标注、输入规模的控制、输入质量的提升等环节，对DeepDive做了一些改进，并将这些改进成功运用到业务落地的过程中。

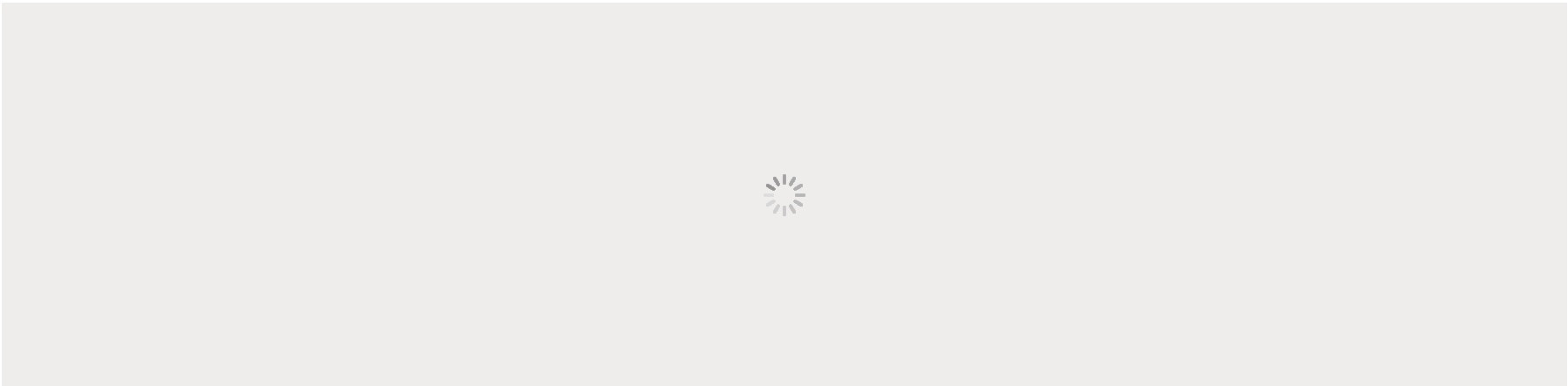
## 中文NLP标注

NLP标注是数据处理的一个重要环节。DeepDive自带的Stanford CoreNLP工具主要是针对英文的处理，而在知识图谱的应用中，主要的处理需求是针对中文的。因此，我们开发了中文NLP标注的外部流程来取代CoreNLP，主要变动如下：

- 使用Ali分词代替CoreNLP的token切分，删除词根还原、POS标注和依存语法分析，保留NER标注和token在文本中的起始位置标注。
- token切分由以词为单位，改为以实体为单位。在NER环节，将Ali分词切碎的token以实体为粒度重新组合。例如分词结果“华盛顿”、“州立”、“大学”将被组合为“华盛顿州立大学”，并作为一个完整的实体获得“University”的NER标签。
- 长句的切分：文本中的某些段落可能因为缺少正确的标点或包含众多并列项等原因，出现切分后的句子长度超过一定阈值（如200个中文字符）的情况，使NER步骤耗时过长。这种情况将按预定义的一系列规则进行重新切分。

### 主语自动增补

数据处理环节的另一个改进是添加了主语自动补充的流程。以中文百科文本为例，统计发现，有将近40%的句子缺少主语。如下图刘德华的百科介绍，第二段中所有句子均缺少主语。



主语的缺失很多时候直接意味着候选实体对中其中一个实体的缺失，这将导致系统对大量含有有用信息的句子无法进行学习，严重影响系统的准确率和召回率。主语的自动补充涉及两方面的判断：

- 主语缺失的判断
- 缺失主语的添加

由于目前业务应用中涉及的绝大多数是百科文本，缺失主语的添加采用了比较简单的策略，即从当前句的上一句提取主语，如果上一句也缺失主语，则将百科标题的NER结果作为要添加的主语。主语缺失的判断相对复杂，目前主要采用基于规则的方法。假设需要提取的候选对（E1， E2）对应的实体类型为（T1， T2），则判定流程如下图所示：



具体的主语补充实例和处理过程举例如下：



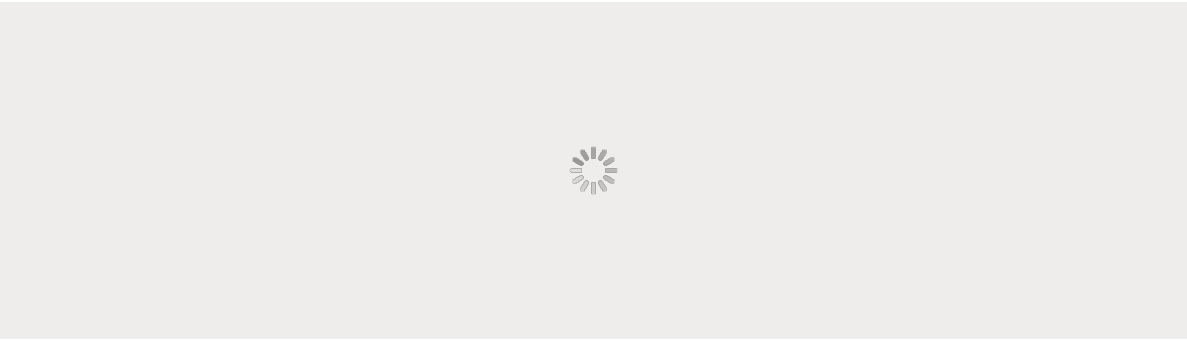
以百科文本为例，经实验统计，上述主语自动补充算法的准确率大约在92%。从关系抽取的结果来看，在所有的错误抽取case中，由主语增补导致的错误比例不超过2%。

基于关系相关关键词的输入过滤

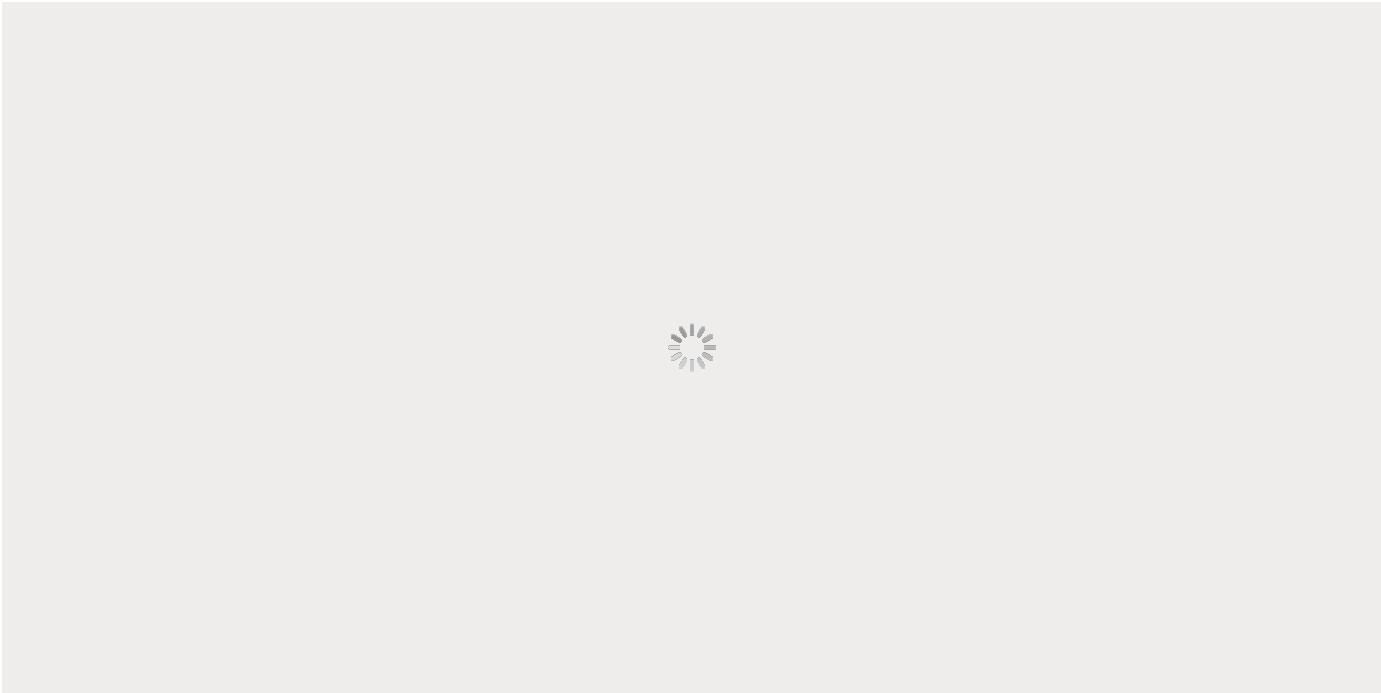


DeepDive是一个机器学习系统，输入集的大小直接影响系统的运行时间，尤其在耗时较长的特征计算和学习推理步骤。在保证系统召回率的前提下，合理减小输入集规模能有效提升系统的运行效率。

假设需要提取的三元组为R（E1， E2）且（E1， E2）对应的实体类型为（T1， T2）。DeepDive的默认运行机制是：在数据处理阶段，提取所有满足类型为（T1， T2）的实体对作为候选，不考虑上下文是否有表达关系R的可能性。例如，抽取婚姻关系时，只要一个句子中出现大于等于两个的人物实体，该句子就会作为输入参与系统整个数据处理、标注和学习的过程。以下五个例句中，除了句1，其它4句完全不涉及婚姻关系：



尤其当句中的两个人物实体无法通过远程监督获取正例或负例标签时，此类输入无法在学习环节为系统的准确率带来增益。为减小此类输入带来的系统运行时间损耗，我们提出了以下改进算法：



实验证明，利用改进算法得到的输入集规模有显著的减小，以百科文本的抽取为例，婚姻关系的输入集可缩小至原输入集的13%，人物和毕业院校关系的输入集可缩小至原输入集的36%。输入集的缩小能显著减少系统运行时间，且实验证明，排除了大量doubt标注实体候选对的干扰，系统的准确率也有较大幅度的提升。

需要指出的是，虽然在输入环节通过关系相关关键词进行过滤减小输入规模，能最有效地提高系统运行效率（因为跳过了包含特征提取在内的所有后续计算步骤），但该环节的过滤是以句子为单

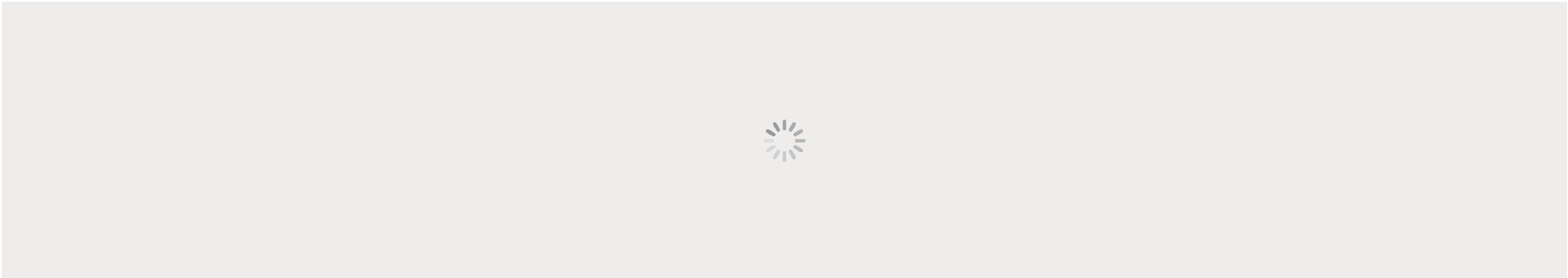
位，而非作用于抽取的候选实体对。来看一个婚姻关系提取的多人物示例：

- 除了孙楠、那英等表演嘉宾盛装出席外，担任本场音乐会监制的华谊兄弟总裁王中磊先生、冯小刚导演和夫人徐帆，以及葛优、宋丹丹、李冰冰等演艺明星也一一现身红毯，到场支持此次音乐会。

因为含有婚姻关系相关的关键词“夫人”，该句子将被保留为系统输入。从该句提取的多个人物候选实体对需要依靠更完善的启发式规则来完成进一步的标注和过滤。

### 实体对到多实体的扩展

关系抽取的绝大部分任务仅涉及三元组的抽取。三元组一般有两种形式，一种是两个实体具有某种关系，形如R（E1， E2），例如：婚姻关系（刘德华，朱丽倩）；另一种是实体的属性值，形如P（E， V），例如：身高（刘德华， 1.74米）。DeepDive默认的关系抽取模式都是基于三元组的。但在实际应用中，有很多复杂的关系用三元组难以完整表达，例如，人物的教育经历，包括人物、人物的毕业院校、所学专业、取得学位、毕业时间等。这些复杂的多实体关系在神马知识图谱中用复合类型来表示。因此，为使抽取任务能兼容复合类型的构建时，我们对DeepDive的代码做了一些修改，将候选实体对的提取，扩展为候选实体组的提取。代码修改涉及主抽取模块中的app.ddlog、底层用于特征自动生成的DDLlib和udf中的map\_entity\_mention.py、extract\_relation\_features.py等文件。下图展示了一个扩展后的实体组抽取实例，抽取关系为（人物、所在机构、职位）：



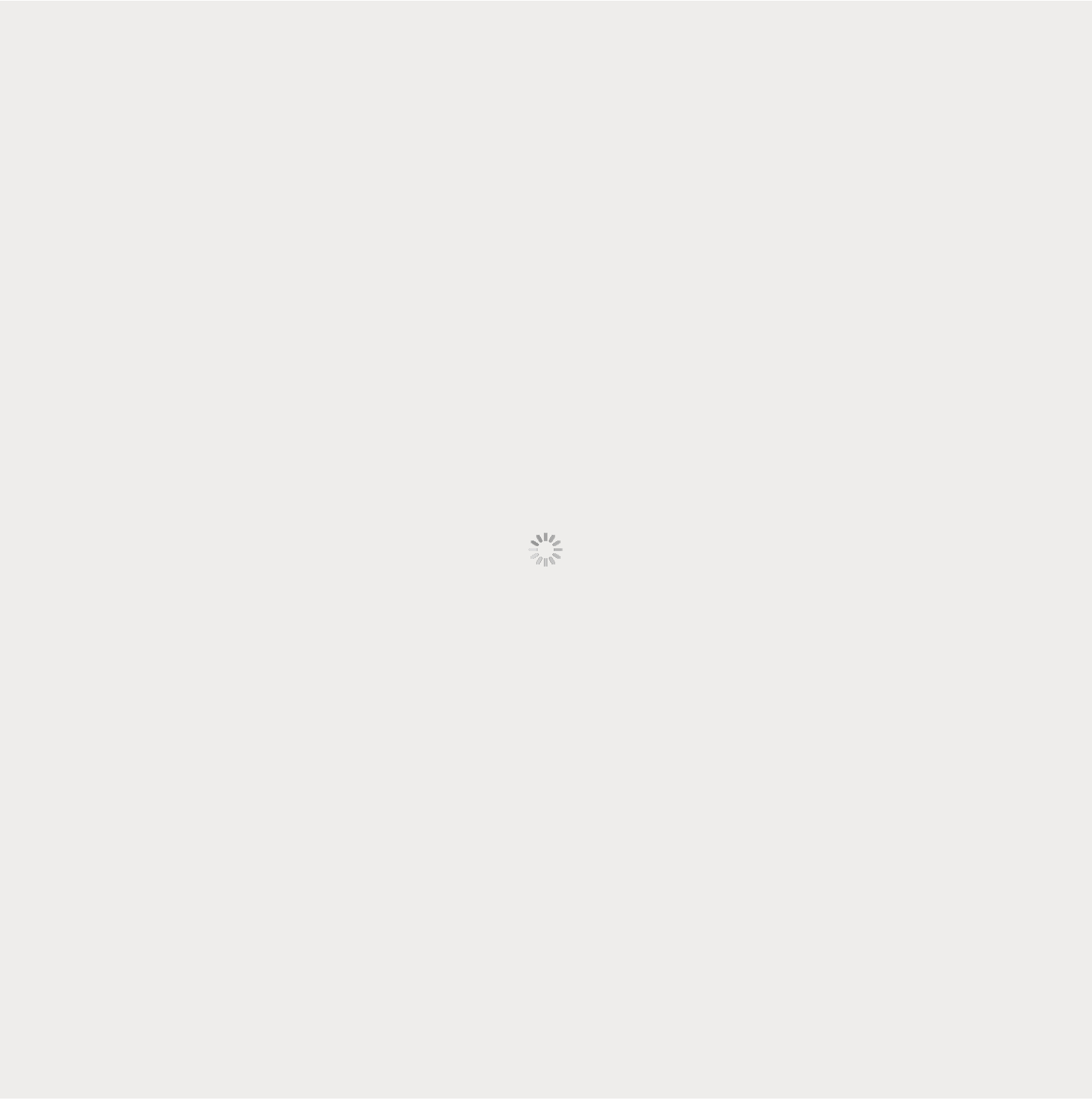
### 应用DeepDive的数据构建工作

本节首先给出一个输入示例以及该示例在DeepDive运行过程中每一步的输出结果，如下图所示。通过这个示例，我们可以对DeepDive各模块的功能和输出有更直观的认识。



为了更详细地了解DeepDive的应用和改进算法的效果， 以下我们给出一个具体的婚姻关系抽取任务的相关运行数据。

下表显示了该抽取任务在数据处理阶段各步骤的的耗时和产出数量：



在数据标注的远程监督阶段，我们除了使用知识图谱中已有的夫妻关系做正例标注，还使用了已有的父母-子女关系和兄弟姐妹关系做负例标注，得到正例数千个，正负标注候选实体的比例约为1:2。





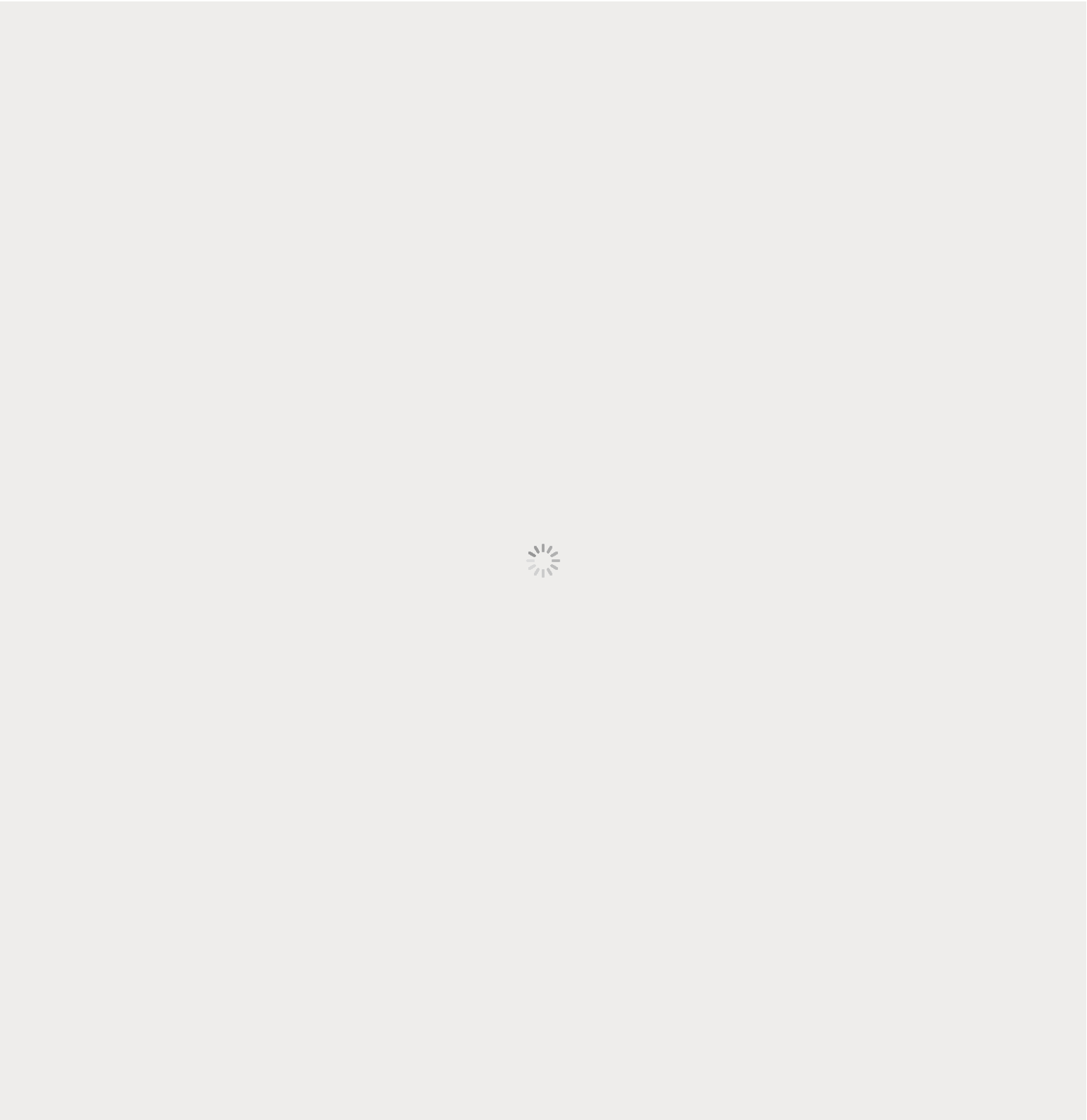
在DeepDive系统中，远程监督的wrong label problem可以依靠合理编写的启发式规则得到一定程度的纠正。观察婚姻关系的wrong label样例，我们发现较大比例的wrong label是夫妻实体以某种合作形式（如合作演出、合作演唱、合作著书等）共现在一个句子中，夫妻实体有一个出现在书名号中时，也容易发生误判。例如：



类似的观察和总结可以编写成启发式规则，依靠从规则得到的负标注抵偿远程监督得到的正标注，

减小系统在学习和推理时的偏差。

虽然启发式规则的编写大多依靠专家知识或人工经验完成，但规则的完善和扩充可以依靠某些自动机制来辅助实现。例如，规则定义：句中出现“P\_1和P\_2结婚”，则（P\_1，P\_2）得到正标注。根据对“和”和“结婚”等token的扩展，我们可以得到“P\_1与P\_2结婚”、“P\_1和P2婚后”、“P\_1和P\_2的婚礼”等类似应该标注为正的语境。这里，token的扩展可以通过word2vec算法加人工过滤实现。下表给出了该抽取任务中用到的规则和相应的统计数据。整个数据标注过程耗为14m21s。



学习与推理过程耗时约38m50s。我们随机截取了部分知识图谱未收录的预测实体对的输出结果展示如下：



对于系统的准确率，我们取expectation为  $[0.95,1]$  区间内的输出结果进行分段统计，统计结果如下列图表所示：



对系统预测的错误样例进行分析，我们总结了几种错误类型，下表按照出现频率从高到低，给出了错误描述和错误示例：



系统召回率的计算相比准确率的计算更为复杂，在语料规模较大的情况下，准确估算召回率将耗费大量的人力。我们采用了抽样检测的方式来估算召回率，具体实践了以下三种方法（统计中 expectation 均取  $\geq 0.95$ ）：

- 抽样含有某个指定实体的所有 sentences，计算召回：含实体“杨幂”的 sentences 共 78 例，含（杨幂，刘恺威）实体对的 sentences 共 13 例，人工判断其中 9 例描述了该实体对的婚姻关系，其中 5 例被召回，召回率为 0.556。
- 用于远程监督正例标注的知识图谱实体对超过 4000 对，统计表明，其中 42.7% 的实体对出现在



了语料中，26.5%的实体对被召回，召回率为0.621。

- 输入集随机挑选100例positive cases，其中49例的expectation值 $\geq 0.95$ ，召回率为0.49。

基于DeepDive的关系抽取研究目前已较为完整，并已经在神马知识图谱的构建业务中落地。目前在数据构建中的应用涉及人物、历史、组织机构、图书、影视等多个核心领域，已抽取关系包括人物的父母、子女、兄弟姐妹、婚姻、历史事件及人物的合称、图书的作者、影视作品的导演和演员、人物的毕业院校和就业单位等。以百科全量语料为例，每个关系抽取任务候选sentence集合的规模在80w至1000w，经改进算法过滤，输入规模在15w至200w之间，生成的候选实体对规模在30w至500w之间。系统每轮迭代运行的时间在1小时至8小时之间，约经过3-4轮迭代可产出准确率和召回率都较高的数据给运营审核环节。系统运行至今，已累计产出候选三元组近3千万。

---

除此之外，基于深度学习模型的关系抽取技术及其在神马知识图谱数据构建中的应用，我们也在不断探索和实践。明天，阿里妹将继续为大家介绍相关的技术进展和业务落地过程中遇到的一些挑战，敬请关注哦。

## 参考文献

- [1]. 林衍凯、刘知远，基于深度学习的关系抽取
- [2]. Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In EMNLP. 1753–1762.
- [3]. Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao. 2017. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence
- [4]. Siliang Tang, Jinjian Zhang, Ning Zhang, Fei Wu, Jun Xiao, Yueting Zhuang. 2017. ENCORE: External Neural Constraints Regularized Distant Supervision for Relation Extraction. SIGIR'17
- [5]. Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. EMNLP.
- [6]. Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In Machine Learning and Knowledge Discovery in Databases. Springer. 148–163.
- [7]. Ce Zhang. 2015. DeepDive: A Data Management System for Automatic Knowledge Base Construction. PhD thesis.
- [8]. Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; and Weld, D. S. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 541–550. Association for Computational Linguistics.
- [9]. Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and

Computational Natural Language Learning, 455–465. Association for Computational Linguistics.

[10]. Shingo Takamatsu, Issei Sato and Hiroshi Nakagawa. 2012. Reducing Wrong Labels in Distant Supervision for Relation Extraction. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 721–729

[11]. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J.; et al. 2014. Relation classification via convolutional deep neural network. In COLING, 2335–2344.

[12]. Ce zhang, Cheistopher Re; et al. 2017. Communications of the ACM CACM Homepage archive  
Volume 60 Issue 5, Pages 93-102

[13]. Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2, 1003–1011. Association for Computational Linguistics.

[14]. <http://deepdive.stanford.edu/>

## 你可能还喜欢

点击下方图片即可阅读



如何用架构师思维解读区块链技术？



十年前，他如何自学技术进阿里？





首次公开！菜鸟弹性调度系统的架构设计



关注「**阿里技术**」  
把握前沿技术脉搏