2020.5 月总结

月报主要分为两部分，第一部分为对 5 月工作的总结，第二部分为对 6 月工作的计划。

1.5 月工作总结

五月工作主要分为一下三点：

1. 对于 2020ACM multimedia 论文进行修改
2. 对于多模态网站的构建
3. 相关视觉实体链接论文学习

2.ACM MM 论文修改

根据 IJCAI 论文投稿的评审意见和 ACM MM 的会议要求，对之前的视觉实体链接论文进行修改，主要的评审意见和修改方案如下：

Meta-reviewer：The strong point of the paper are (i) the novel task of visual entity linking, (ii) a first set of experimental results on the task (iii) a new and large-scale dataset for this challenging task (iv) good performance results compared to all the baselines. However, the paper presents also a number of weak points that can be summarized in the fact that the presentation has several weaknesses that makes the paper difficult read, some technical details are not well explained, and what is more important the method of constructing the dataset and the characteristics of the dataset are not sufficiently explained. In particular there the paper does not provide enough information to evaluate the quality of the dataset. For instance the authors state that they use DBPedia 2016 as their knowledge graph. However, usually it is not trivial to find all listed images for the nodes for any public knowledge graph/ontology, and I assume the same should happen with DBPedia (e.g. there are often many broken links and one simply cannot download the image). There is no discussion on the number of images used per node or in total, the quality of these images, what they look like, etc.

Another weakness pointed out from a reviewer concerns the evaluation criteria. In "4.3 Main Results", authors first talk about different feature types (visual, textual, and structural) in a very confusing way. Then they move on to say that "we first use the Faster-RCNN visual object detection method to generate the corresponding bounding boxes for comparison, then, randomly connect the entity in the candidate list created by the above method to the entity bounding boxes". What does it mean to randomly assign entity boxes to objects? It is not clear here, and it looks like that this is how the authors run their entire evaluation (i.e. this is how they compute accuracy scores).

主要的意见分为两点：

1. 这篇论文没有提供足够的信息来评估数据集的质量。对于数据集描述的部分可以增加，详细描述数据集构建过程和一些数据集的参数，讨论每个节点使用的图像的数量或总数、这些图像的质量、它们的外观等。

2. 在“4.3 主要结果”中，作者首先以一种混乱的方式讨论了不同的特征类型(视觉的、文本的和结构的)。由于我们任务的开创性，所以我们只找到一个对比实验，但是这显然是不够，我们通过设计一些利用一些单模态附加额外技术框架的方法作为实验设定。需要在4.1 节进行更详细的描述，解释清楚实验的设定，进行润色。

Reviewer#1：

The manuscript proposes the novel task of visual entity linking, and provides initial experimental results on the task. Given an image, a bounding box for an object in this image, and an image caption, the task is to link the object bounding box to a KB entity. The authors collect a data set where they run their experiments, and compare to some baselines.

I found the paper very hard to read, unclear, and not organized. The tasks are not clearly explained, and the mathematical notation is unnecessarily convoluted. For example, \bm{X}, \bm{x}, \bm{x}^i, \bm{x}_t, and \bm{x}_{t,l} are used to denote many variables. Also, \bm{y}, \bm{y}', \bm{y}_v, \bm{y}_s, and \bm{y}^{\bm{n}_i} are used to denote other variables. This is extremely hard to read, and unnecessarily convoluted.

Moreover, the authors propose a new task, curate the data for this task and do not say anywhere that neither this data nor their model are going to be publicly available to the research community.

There are simply too many moving pieces in this manuscript. There is a NER trained using BERT + Bi-LSTM + CRF, scene graph parsing, KG entity linking, etc.

In 3.2, Visual features, it reads "we use the VGG-16 structural configuration". What does that mean?

I recommend the authors to write a considerably larger paper with revised maths, where the text is well organised, etc. I suggest that each small component of the task/model be thoroughly evaluated, discussed, and motivated.

1. 考虑添加一个模态注意力可视化的实验，使用可视化的图像来表示我们的模态注意力模块的有效性。
2. 简化参数设定，对参数进行对称性的声明，减少参数的复杂程度。
   对文章的描述进行修改，有些地方的表达可能还不太严谨。

Reviewer#2：

This paper focuses on recognising fine-grained KG entities in visual scenes, which can be of interested to many researchers. The topic in this paper has pointed out there are still many limitations in state-of-the-art image retrieval systems that are armed with cutting-edge and sophisticated “black-box” models. The proposed solutions to the problem in such a fine-grained setup have provided a path to address the challenges, which been largely neglected but is important in many real-world scenarios.

For the technical parts, authors propose a novel multi-modal learning method by fusing the features from three different modalities (visual, textual and heterogeneous graphs), followed by linking the entities with a deep learning-to-rank model, which is technically sound. To

evaluate the proposed framework, authors have specifically constructed a new and large-scale dataset and conducted extensive experiments over it. The performance comparisons have been made against recent state-of-the-art works and the variants of the proposed method. From the experimental results, authors have shown their method have outperformed all the compared works in terms of accuracy. In general, this paper is well written and easy to follow. I would like to highlight the strengths and weaknesses of this paper as follows:
Strengths:
- important information gains in visual scenes
- novel multi-modal features learning and exploiting
- a new and large-scale dataset for this challenging task (publicly accessible now?)
- outstanding performance results compared to all the baselines
Weaknesses:
- the presentation has several weaknesses (see below)
- still need more detailed experiment analysis for performance improvements over the counterparts.
- technical details are not well explained.
e.g. why can't the loss of visual features and textual features added directly in equation 7? How to determine the extra parameters?
- Several minor typos and errors:
page 2 "an images" (need carefully check spelling and grammar)
Table 1 font is too small, so the data is not very clear for the readers.

There is also an interesting issue here. If the entities in the collected news data do not exist in the corresponding KG, how to evaluate the linking process, or how to expand the visual entity linking?
模型方面根据评审意见来说没有太大的问题，可以解释一下我们选择损失函数的想法，主要对文章的描述进行修改。

最终对论文进行了修改，适当增加了相关工作的内容，增加了关于模态注意力机制的测评，通过模态可视化对不同 sample 的模态注意力进行分析，验证了模态注意力机制的有效性。

通过对模态注意力的可视化，我们对一些例子进行了分析：

In the example of the first row in Fig. 4, "Jobs, Apple's founder, attended the launch of the new iPhone". We first generate the candidate entity list of "Jobs", "Apple", "iPhone". For the first example, in the process of entity linking for "Jobs" entity, we can learn that the influence of visual modality is higher than that of textual modality according to the color depth of the modalities. For the other two entities "Apple" and "iPhone", the influence of visual modality is much lower than that of textual modality. Because there are few candidate entities of "Apple" and "iPhone", just rely on the textual modality we can easily find the knowledge graph entity corresponding to the contextual semantics, but there are many related entities for "Jobs" entity, so we need to use the feature vector of visual modality for the entity linking task, which is why different entity categories have different modality weights.

我们通过分析在不同实体中视觉和文本所含比不同的情况，深色表示注意力更多的模态，浅色表示表示注意力小的模态，验证了模态注意力模型机制可以使得模型更加注意相关信息，从而使得链接结果表现更佳。

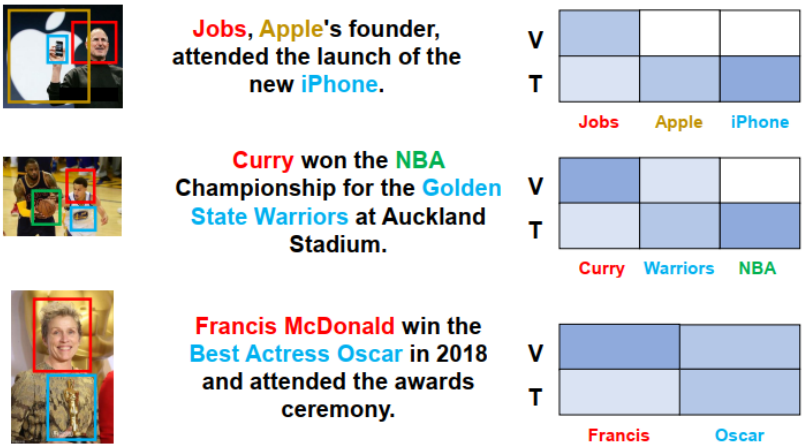


Figure 4: Visualization of modality attention from VELD test data. The model makes final predictions based on the weighted signal from all the modalities. Modalities- V:visual, T:textual.

3. 多模态网站的构建

对我们的多模态网站进行了一些更新，将 SPARQL 功能部署在实验室服务器上面，并设计了一个可视化界面进行访问，近期实验室网络和工作状态不稳定，等待稳定后会继续更新。

Edit query

show my namespaces, add a namespace

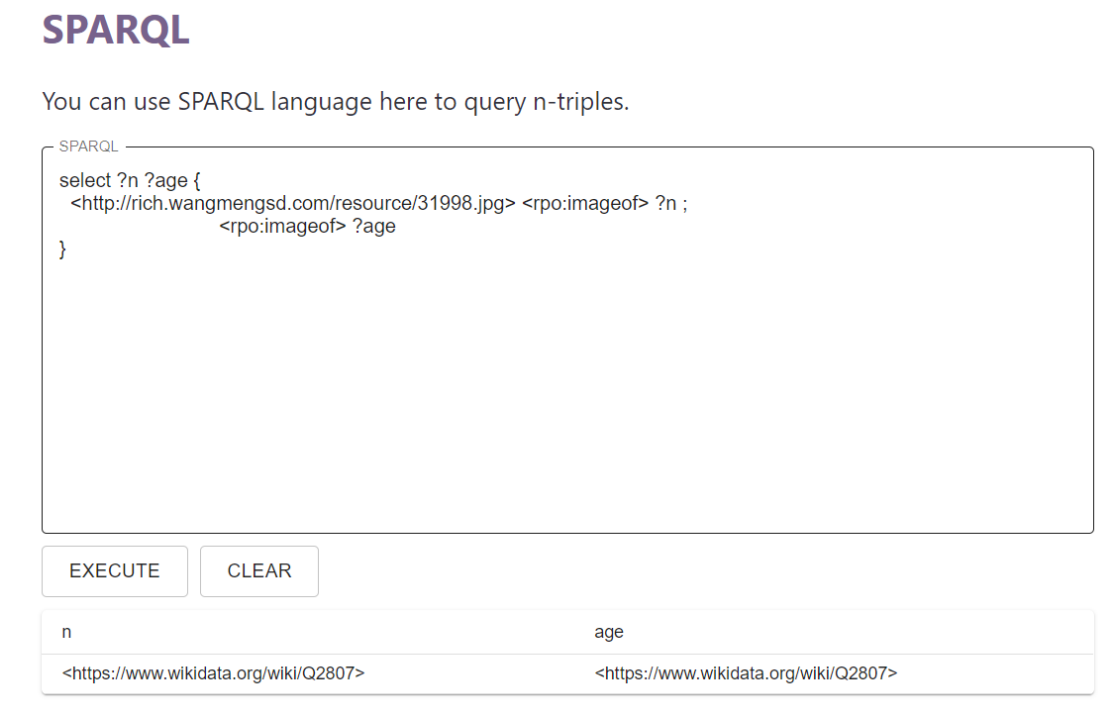
```
1 select ?n ?age {
2   <http://rich.wangmengsd.com/resource/31998.jpg> <rpo:imageof> ?n ;
3                                 <rpo:imageof> ?age
4 }
```

Execute

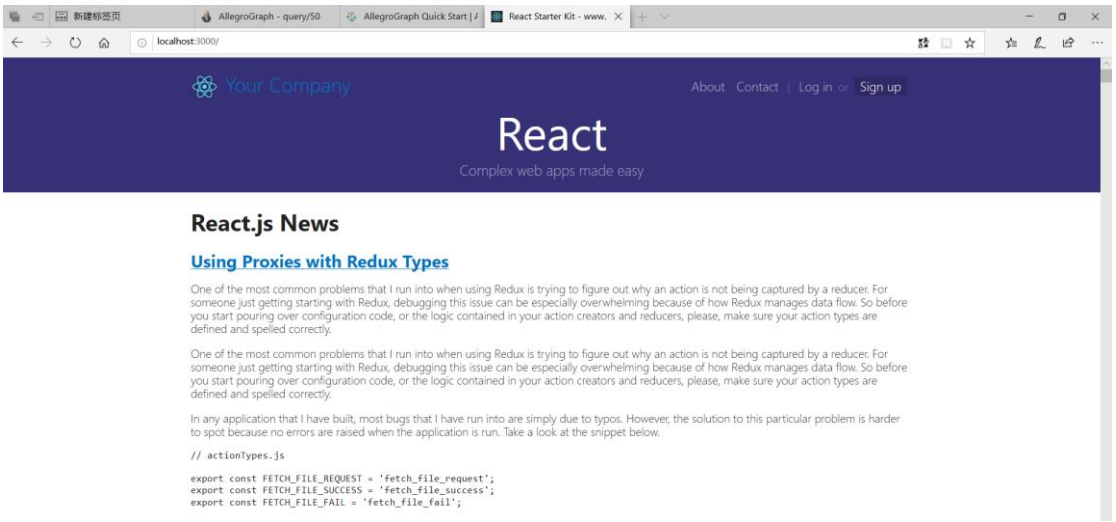
Result

Download as SPARQL JSON

n	age
<https://www.wikidata.org/wiki/Q2807>	<https://www.wikidata.org/wiki/Q2807>



如图所示，现有的网站已经可以在线进行 SPARQL 查询，可以直接用前端输入语句然后从 agraph 后端调用结果，同时也对响应时间进行了测试，千条数据级别的查询速度大概在 100ms 左右。



4. 相关论文阅读

《Neural Collective Entity Linking Based on Recurrent Random Walk Network Learning》

此文是 2019 年 IJCAI 中科大的一篇关于全局实体链接的文章，与上述论文的主要区别是使用一个堆叠随机游走层以加强证据使得相关的 EL 成为高概率决策的方法来进行实体链接，其中候选实体之间的语义相互依赖性主要是从一个外部知识库引导。在传统的目标函数

中引入一个语义规则器，保持集体 EL 决策的一致性，从而使外部 EL 决策具有一致性。

论文模型如下：

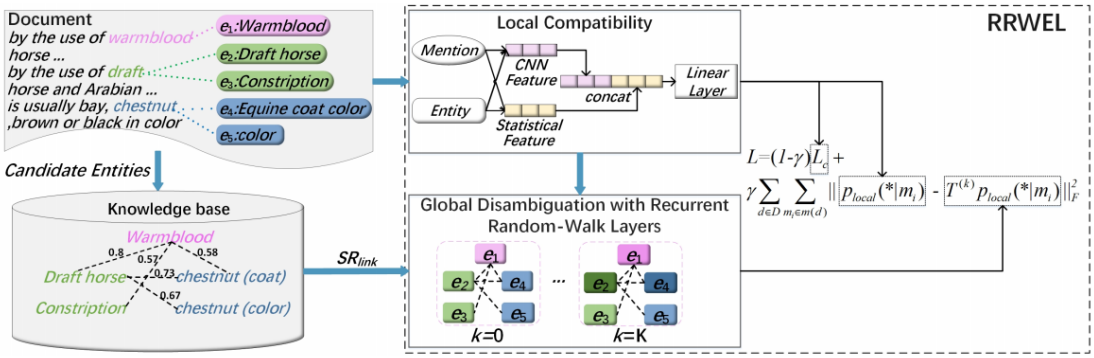


Figure 1: The architecture of our proposed RRWEL model.

本文从另外一个角度解决 Collective Entity Linking，其中的 SRlink 还是通过第二篇文章的方法进行计算，利用随机游走策略替代之前的多头注意力模型。

《Neural Collective Entity Linking》

这是刘知远老师组在 2018 年末发表的一篇论文，主要思想是利用 GCN 将局部上下文特征和全局语义一致性结合起来进行实体链接，对相邻实体提及的子图进行近似的图卷积来获取最终结果。

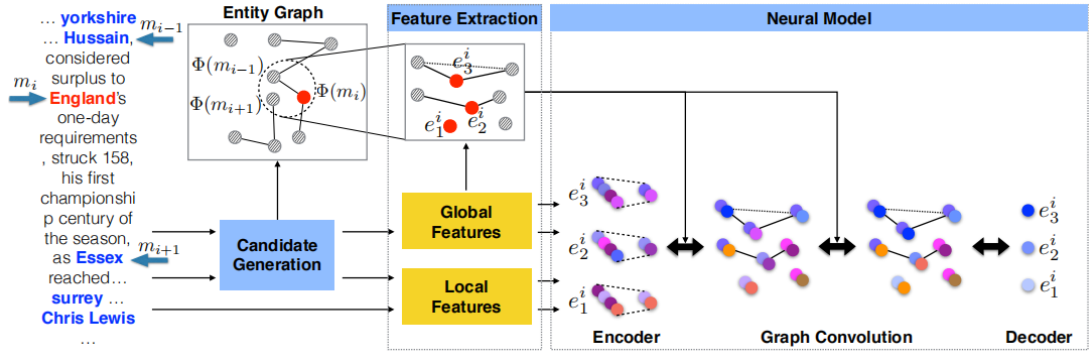
现有的实体链接主要依赖于本地上下文（提及的相近 tokens）独立进行解析，而忽略了全局特征，即文档中所有提及的目标实体在主题上应该保持一致性，具体示例见下图：

Figure 1: Illustration of named entity disambiguation for three mentions *England*, *Hussain*, and *Essex*. The nodes linked by arrowed lines are the candidate entities, where red solid lines denote target entities.

如图所示，传统 EL 模型不能从周围的词中找到足够的消歧线索，但是利用全局相关性则会得到最终的正确结果。此文提出了一个 Neural Collective Entity Linking model (NCEL)模型，将深度神经网络与图卷积网络(GCN)结合进行全局 EL，允许对实体图进行灵活编码。集成了局部上下文信息和文档中提及的全局相互依赖性，并且可以以 end-to-end 方式有效地进行培训。同时将注意机制引入到局部上下文信息的鲁棒模型中，通过选择信息词和过滤噪声来实现，利用 GCNs 来改善候选实体的判别信号，利用正确实体下的丰富结构。为了减少全局计算，对相邻提及的子图进行卷积，整体的连贯性通过文档上的滑动窗口以链状的方式实现。

这篇文章 EL 的实体候选列表生成是通过大型的语料库来计算先验概率，得到一个预定义的“提及--实体”字典来表示候选列表，同时作为其局部特征应用于模型，但是这种方法的问题就是如果在一个非限定域的条件下进行实体链接，即得到的提及不曾出现于训练数据中，应该如何去生成候选实体列表？所以此模型的泛化能力仍需进行讨论。

主要的模型结构如下图：



结构也比较容易看懂，大体分为三个部分。一，候选实体列表生成，主要是依赖一个离线的训练数据得到；二，特征抽取，主要为两方面：局部特征和全局特征；三，神经网络模型，利用 GCN 来得到每个提及的最大概率实体。

对于我们工作的借鉴意义我认为主要在全局特征的应用上，这种一段话中所有实体的连贯相似性也可以应用于我们的之后的模型上面。

《VisualBERT: A Simple and Performant Baseline for Vision and Language》

这篇论文里作者们提出了 VisualBERT，这是一个可以对一系列不同的视觉-语言任务进行建模的框架，而且简单灵活。VisualBERT 包含了一组层叠的 Transformer 层，借助自我注意力把输入一段文本中的元素和一张相关的输入图像中的区域隐式地对齐起来。除此之外，作者们还提出了两个在图像描述数据上的视觉-语言关联学习目标，用于 VisualBERT 的预训练。作者们在 VQA、VCR、NLVR2 以及 Flickr30K 这四个视觉-语言任务上进行了实验，结果表明 VisualBERT 以明显更简单的架构在所有任务中都达到了做好的表现或者和竞争者相当的表现。作者们的进一步分析表明 VisualBERT 可以在没有任何显式监督的情况下建立语言元素和图像中区域之间的联系，而且也对句法关系和追踪（根据描述建立动词和图像区域之间的关系）有一定的敏感性。

《Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training》

作者们提出了 Unicoder-VL，这是一个以预训练的方式学习视觉和语言的联合表征的通用编码器。这个模型借鉴了 XLM 和 Unicoder 等跨语言、预训练模型的设计思路，视觉和语言内容都会被传入一个多层 transformer 中，作为跨模态预训练阶段；预训练阶段使用三个任务，包括掩蔽语言建模、掩蔽对象标签预测以及视觉-语言匹配。前两个任务会让模型学习从基于语言和视觉内容输入的联合 token 学习到内容相关的表征；后一个任务尝试预测一张图像和一段文本描述之间是否相符。在大量的图像-描述对上预训练之后，作者们把 Unicoder-VL 迁移到了图像-文本检索任务上，只添加了一个额外的输出层，就在 MSCOCO 和 Flicker30K 两个数据集上都取得了目前最佳的表现。

《 VL-BERT: Pre-training of Generic Visual-Linguistic Representations》

作者们设计了一种新的用于视觉-语言任务的可预训练的通用表征，名为 VL-BERT。VL-BERT 把简单有效的 Transformer 模型作为主干并进行拓展，视觉和语言嵌入特征可以同时作为输入。输入中的每个元素可以是来自句子的一个单词，也可以是输入图像中的一个感兴趣区域。模型的设计也为了能够和所有视觉-语言的下游任务兼容。作者们在大规模的 Conceptual Captions 上对模型进行预训练，三个预训练任务为：带有视觉线索的掩蔽文字建模、带有语言线索的感兴趣区域分类、句子-图像关系预测。作者们通过大量的实证分析表明预训练阶段可以更好地对齐视觉-语言线索，并为视觉问答、视觉常识推理、代指词汇理解等下游任务带来收益。值得一提的是 VL-BERT 在 VCR 排行榜上取得了单一模型的最好成绩。

5.6 月计划

1.完成四门专业课程作业，这学期主要是在线教学，所以大部分课程都有一个较为庞大的课程设计，需要花费较多时间去完成四门课程设计。
2.继续多模态网站的更新任务，和超宇讨论一下尝试把现有的 VEL 模型部署到线上，并且寻找一个外网服务器存储图片信息，使得我们的网站可以在外网访问到。
3.阅读相关论文，去考虑一下漆老师提出的 fewshot 情况下的 VEL 任务，做一些文献调研，考虑一下现有相关模型如何应用到我们的任务中来。
4. 继续对现有模型发展进行学习和代码实践，养成日常习惯。