论文阅读笔记

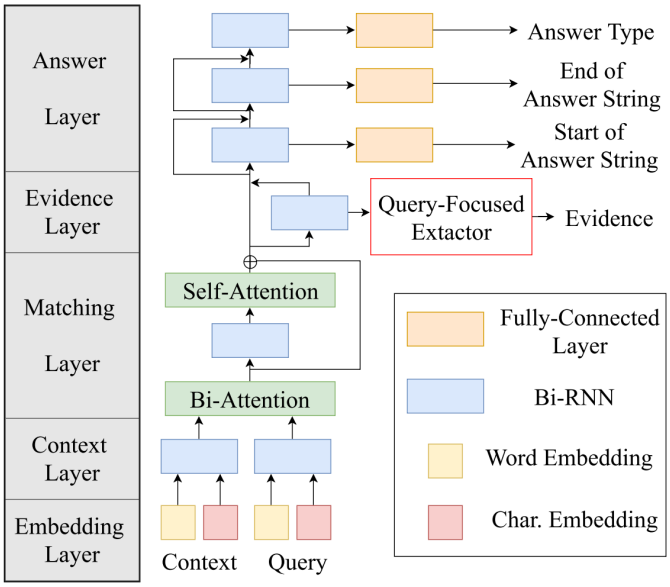
Step7

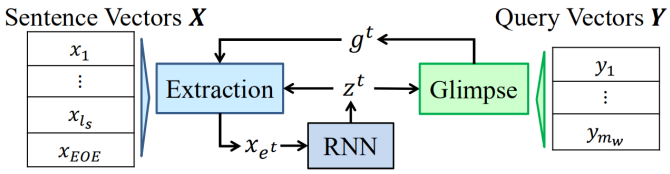
MF1833063, 史鹏, spwannasing@gmail.com

2019 年 8 月 10 日

1 Answering while Summarizing:Multi-task Learning for Multi-hop QA with Evidence Extraction

本文基于HotpotQA，提出了一个叫Query Focused Extractor (QFE)的模型，基于 extractive summarization模型，另外相比已有的模型，克服了预测evidence sentence之间相互独立的问题（通过使用RNN），以及引入了Multi-Task，可以和任意的RC模型结合。亮点是即使QFE和一个简单的baseline RC模型结合起来，也能达到evidence extraction score上的SOTA效果。





$$z^t = \mathrm{RNN}\left(z^{t-1}, x_{e^t}\right) \in \mathbb{R}^{2d_c} \tag{1.1}$$

$$\Pr\left(i; E^{t-1}\right) = \mathrm{softmax}_i\left(u_i^t\right) \tag{1.2}$$

$$u_i^t = \begin{cases} v_p^\top \tanh\left(W_{p1}x_i + W_{p2}g^t + W_{p3}z^t\right) & (i \notin E^{t-1}) \\ -\infty & (\text{ otherwise }) \end{cases} \tag{1.3}$$

$$\begin{aligned} g^t &= \sum_j \alpha_j^t W_{g1} 1_j \in \mathbb{R}^{2d_c} \\ \alpha^t &= \mathrm{softmax}\left(a^t\right) \in \mathbb{R}^{m_w} \\ a_j^t &= v_g^\top \tanh\left(W_{g1}y_j + W_{g2}z^t\right) \end{aligned} \tag{1.4}$$

$$L_E = -\sum_{t=1}^{|E|} \log\left(\max_{i \in E \setminus E^{t-1}} \Pr\left(i; E^{t-1}\right)\right) + \sum_i \min\left(c_i^t, \alpha_i^t\right) \tag{1.5}$$

$$L = L_A + L_E \tag{1.6}$$

2 Do you know that Florence is packed with visitors?Evaluating state-of-the-art models of speaker commitment

这篇论文展示出了带有语言学知识的模型的巨大潜力

对基于规则的和双向LSTM这两种最先进的说话人承诺模型进行了系统的评价

论文中的语言学分析给人启发，也展现出了系统的优势和劣势

当一个人，比如 Mary，问你「你知不知道佛罗伦萨全都是游客？」，我们会认为她相信佛罗伦萨全都是游客；但如果她问「你觉得佛罗伦萨游客多吗？」，我们就不会这样认为。推断说话人承诺（或者说事件真实度）是问答和信息提取任务中的关键部分。在这篇论文中，作者们探索了这样一个假说：语言学信息的缺乏会影响说话人承诺模型中的错误模式。他们的验证方式是在一个有挑战性的自然语言数据集上分析模型错误的语言学关联性。作者们在 CommitmentBank 这个由自然英语对话组成的数据集上评价了两个目前最好的说话人承诺模型。CommitmentBank 数据集已经经过了说话人承诺标注，方式是在 4 种取消蕴含的环境中向着时态嵌入动词（比如知道、认为）的补充内容进行标注。作者们发现，一个带有语言学知识的模型能展现比基于 LSTM 的模型更好的表现，这表明如果想要在这样的有挑战性的自然语言数据中捕捉这些信息的话，语言学知识是必不可少的。对语言学特征的逐项分析展现出了不对称的错误模式：虽然模型能在某些状况下得到好的表现（比如否定式），但它很难泛化到更丰富的自然语言的语言学结构中（比如条件句式），这表明还有很大提升的空间。

(1)	**Context**	The answer is no, no no. Not now, not ever.
	Target	I *never* **believed** I could wish anyone dead but last night changed all that.
		Gold:1.56, Rule-based:3.0, Hybrid: 0.50
(2)	**Context**	Revenue is estimated at $18.6 million. The maker of document image processing equipment said the state procurement division had declared FileNet in default on its contract with the secretary of state uniform commercial code division.
	Target	FileNet said it *doesn't* **believe** the state has a valid basis of default and is reviewing its legal rights under the contract ,
		Gold: -0.47, Rule-based: 3.0, Hybrid: 1.08
		but said it can't predict the outcome of the dispute.
(3)	**Context**	**A**: Yeah, that's crazy. **B**: and then you come here in the Dallas area, um,
	Target	I *don't* **believe** that people should be allowed to carry guns in their vehicled .
		Gold: -2.64, Rule-based: 3.0, Hybrid: 1.40

3 Emotion-Cause Pair Extraction:A New Task to Emotion Analysis in Texts

情绪原因提取（Emotion cause extraction ， ECE）是一项旨在提取文本中某些情绪背后潜在原因的任务，近年来由于其广泛的应用而受到了很多关注。然而，它有两个缺点：1）情绪必须在ECE原因提取之前进行标注，这极大地限制了它在现实场景中的应用；2）先标注情绪然后提取原因的方式忽略了它们是相互指示的事实。

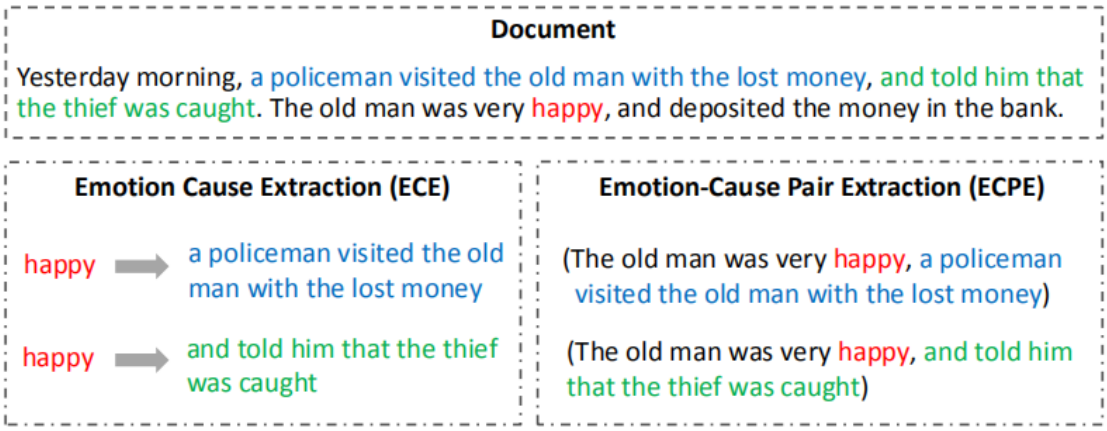
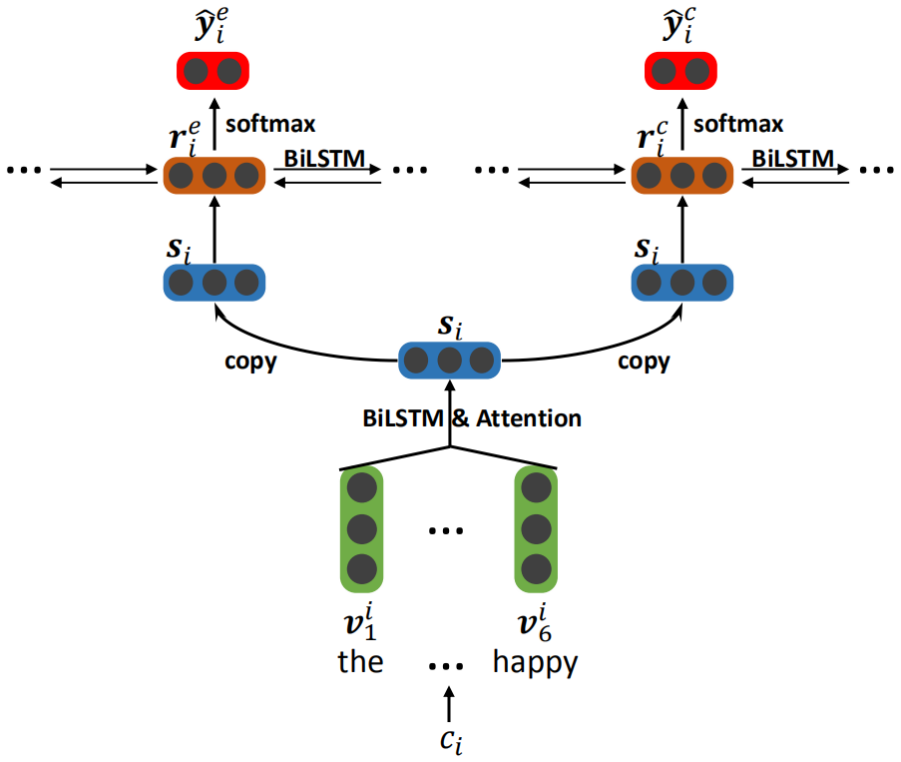


Figure 1: An example showing the difference between the ECE task and the ECPE task.



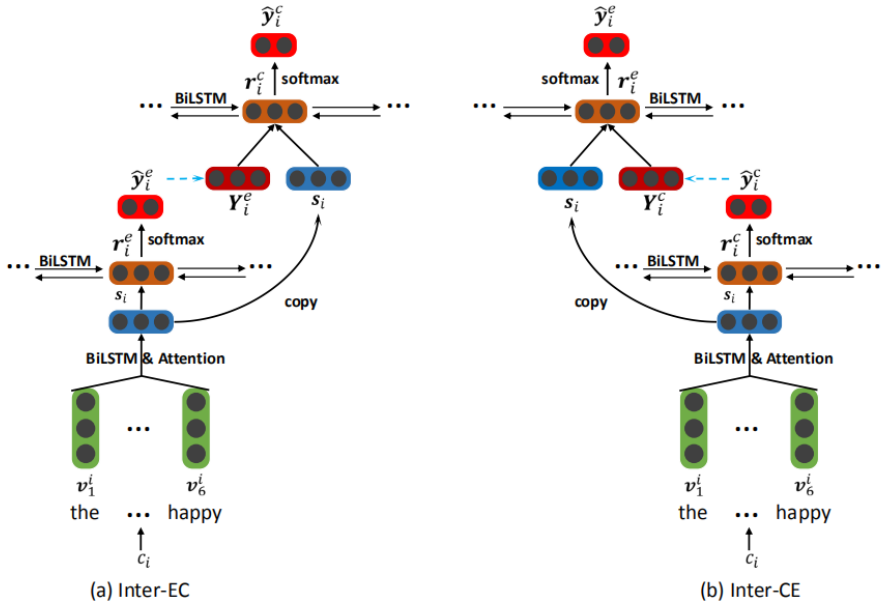


Figure 3: Two Models for Interactive Multi-task Learning: (a) Inter-EC, which uses emotion extraction to improve cause extraction (b) Inter-CE, which uses cause extraction to enhance emotion extraction.

这篇文章的创新点就是提出了一个新的任务 emotion-cause pair extraction (ECPE)，相较于传统的ECE不用依赖于事先的情感标注。其次使用2-step的结构，第一步提取出E和C（有两种结构，一种E和C之间相互独立，另一种是有联系的），第二步进行筛选。

总体算法很简单，只使用了Bi-LSTM和attention结构，核心是提出一个新的Task。

4 Bridging the Gap between Training and Inference for Neural Machine Translation

ACL19最佳论文

神经机器翻译（NMT）是以上下文为条件来预测下一个词，从而顺序地生成目标词。在训练时，它以ground truth词汇作为上下文进行预测；而在推理时，它必须从头开始生成整个序列。反馈上下文信息的这种差异会导致误差累积。此外，词级训练要求所生成的序列与ground truth序列之间严格匹配，这导致对不同的但合理的翻译的过度校正。在本文中，我们在模型训练中不仅从ground truth序列还从预测序列中来采样上下文，其中预测序列是用句子级最优来选择的。

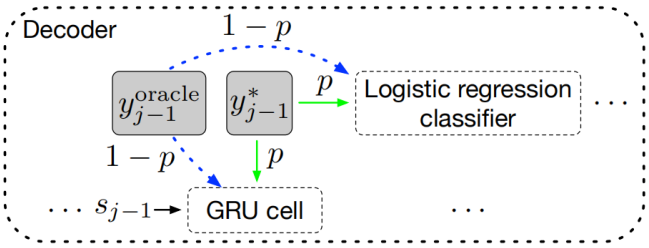


Figure 1: The architecture of our method.

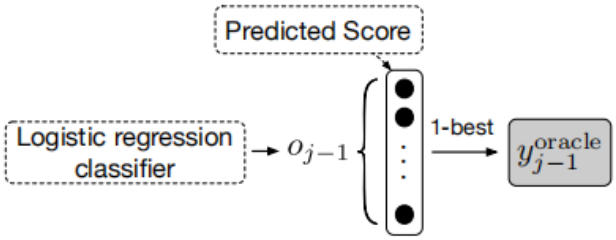


Figure 2: Word-level oracle without noise.

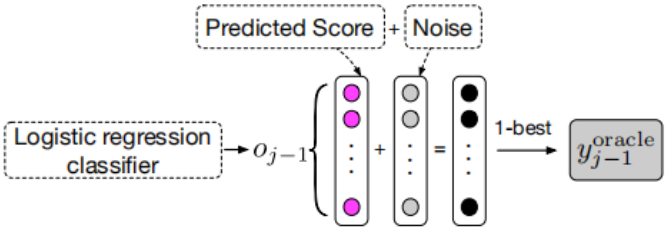


Figure 3: Word-level oracle with Gumbel noise.

5 Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems

过度依赖领域本体和缺乏跨领域知识共享是对话状态跟踪的两个实际存在但研究较少的问题。现有方法通常在在推理过程中无法跟踪未知slot 值，且通常很难适应新领域。在本文中，我们提出了一个可转换对话状态生成器（Transferable Dialogue State Generator，TRADE）它使用复制机制从话语中生成对话状态，当预测在训练期间没有遇到的（domain，slot，value）三元组时可以促使知识转移。我们的模型由一个话语编码器、slot gate、状态生成器组成，它们跨域共享。实验结果表明，TRADE在人类对话数据集MultiWOZ的五个领域中实现了最先进的联合目标准确率48.62%。此外，我们通过模拟针对未见过的领域的zero-shot和few-shot对话状态跟踪，证明了其传输性能。在其中一个zero-shot域中TRADE实现了60.58%的联合目标准确率，并且能够适应少数几个案例而不会忘记已经训练过的域。

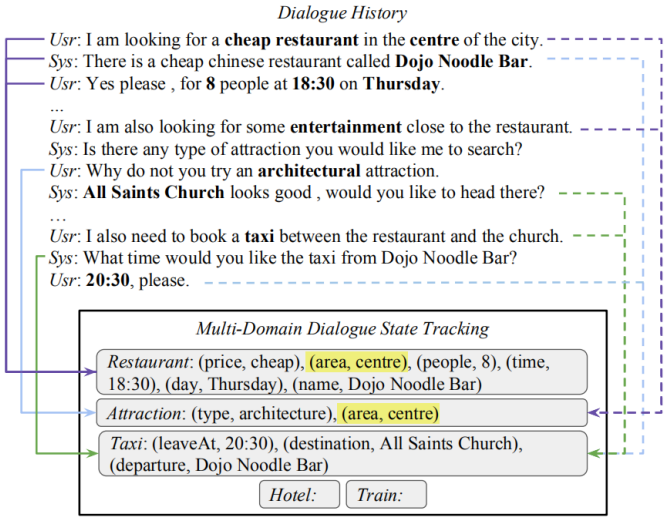


Figure 1: An example of multi-domain dialogue state tracking in a conversation. The solid arrows on the left are the single-turn mapping, and the dot arrows on the right are multi-turn mapping. The state tracker needs to track slot values mentioned by the user for all the slots in all the domains.

$$
\begin{aligned}
P_{jk}^{\text{vocab}} &= \text{Softmax}\left(E \cdot \left(h_{jk}^{\text{dec}}\right)^{\top}\right) \in \mathbb{R}^{|V|} \\
P_{jk}^{\text{history}} &= \text{Softmax}\left(H_t \cdot \left(h_{jk}^{\text{dec}}\right)^{\top}\right) \in \mathbb{R}^{|X_t|}
\end{aligned} \tag{5.1}
$$

$$
\begin{aligned}
P_{jk}^{\text{final}} &= p_{jk}^{\text{gen}} \times P_{jk}^{\text{vocab}} \\
&\quad + \left(1 - p_{jk}^{\text{gen}}\right) \times P_{jk}^{\text{history}} \in \mathbb{R}^{|V|}
\end{aligned} \tag{5.2}
$$

$$
\begin{aligned}
p_{jk}^{\text{gen}} &= \text{Sigmoid}\left(W_1 \cdot \left[h_{jk}^{\text{dec}}; w_{jk}; c_{jk}\right]\right) \in \mathbb{R}^1 \\
c_{jk} &= P_{jk}^{\text{history}} \cdot H_t \in \mathbb{R}^{d_{hdd}}
\end{aligned} \tag{5.3}
$$

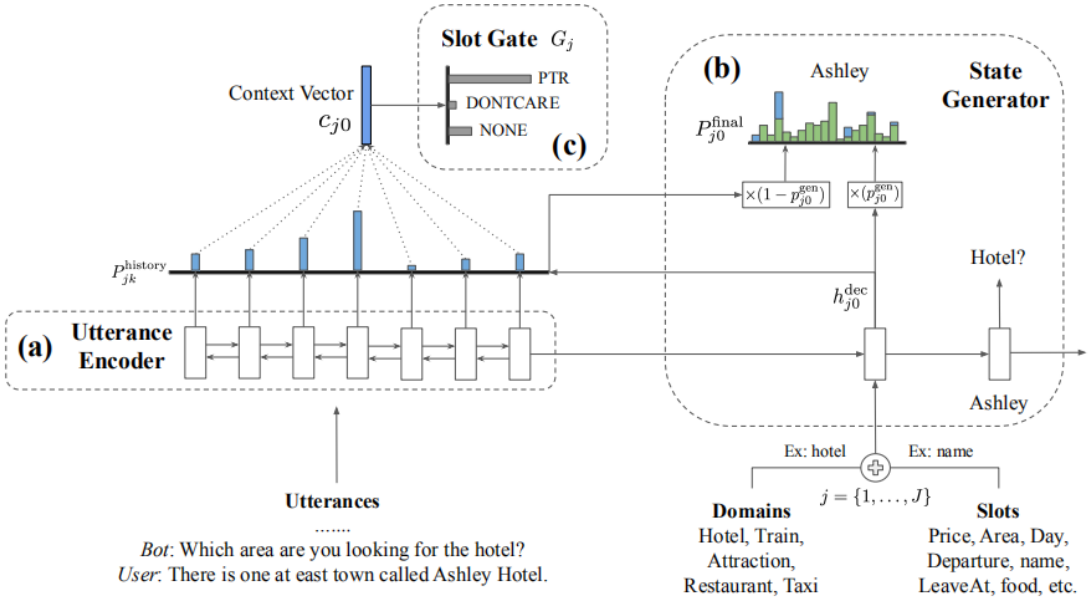


Figure 2: The architecture of the proposed TRADE model, which includes (a) an utterance encoder, (b) a state generator, and (c) a slot gate, all of which are shared among domains. The state generator will decode J times independently for all the possible *(domain, slot)* pairs. At the first decoding step, state generator will take the j-th *(domain, slot)* embeddings as input to generate its corresponding slot values and slot gate. The slot gate predicts whether the j-th *(domain, slot)* pair is triggered by the dialogue.

6 A Simple Theoretical Model of Importance for Summarization

这篇文章讨论了自动文本摘要中长期存在的深层问题：

如何衡量摘要内容的适用性？

提出了「内容重要性」的三部分理论模型

提出了建设性的评估指标

文章中还与标准指标和人类判断进行了比较

摘要研究主要由经验方法驱动，手工精心调制的系统在在标准数据集上表现良好，但其中的信息重要性却处于隐含状态。我们认为建立重要性（Importance）的理论模型会促进我们对任务的理解，并有助于进一步改进摘要系统。为此，我们提出了几个简单但严格定义的概念：冗余（Redundancy），相关性（Relevance）和信息性（Informativeness）。这些概念之前只是直观地用于摘要，而重要性是这些概念统一的定量描述。此外，我们提供了建议变量的直观解释，并用实验证明了框架的潜力以知道后续工作。

7 We need to talk about standard splits

语音和语言技术的标准做法是根据在一个测试集上的性能来对系统进行排名。然而很少有研究人员用统计的方法来测试性能之间的差异是否是由偶然原因造成的，且很少有人检查同一个数据集中分割出不同的训练-测试集时的系统排名的稳定性。我们使用了2000年至2018年间发布的九个词性标注器进行复现实验，这些标注器每个都声称在广泛使用的标准的分割方式上获得了最佳性能。然而当我们使用随机生成的训练-测试集分割时，根本无法可靠地重现某些排名。我们在此建议使用随机生成的分割来进行系统比较。

本文质疑了评估NLP模型时公认且广泛运用的方法；

本文提出了几种关于数据集的标准拆分方法；

本文使用POS标记说明了问题；

本文建议系统排名应当基于使用随机分组的重复评估方法

8 Zero-Shot Entity Linking by Reading Entity Descriptions

提出了zero-shot实体链接任务，其中mentions必须链接到没有域内标记数据的未曾见过的实体。这样做的目的是实现向高度专业化的领域的鲁棒迁移，也因此我们不会假设有元数据或别名表。在这种设置中，实体仅通过文本描述进行标记，并且模型必须严格依赖语言理解来解析新实体。首先，我们表明对大型未标记数据进行预训练的阅读理解模型可用于推广到未曾见过的实体。其次，我们提出了一种简单有效的自适应预训练策略，我们将其称为域自适应预训练（domain-adaptive pre-training ，DAP），DAP可以解决与在新域中链接未见实体的域迁移问题。我们在为此任务构建的新数据集上进行的实验，显示了DAP在强预训练基线（包括BERT）上有所改进。

本文提出了一种新颖的词义消歧系统，专门用于提高稀少的和未见过的词上的表现；

本文提出的感知选择任务被视为连续任务，并且使用了资源的组合；

本文的结果富有洞察力，并且改善了现有水平。

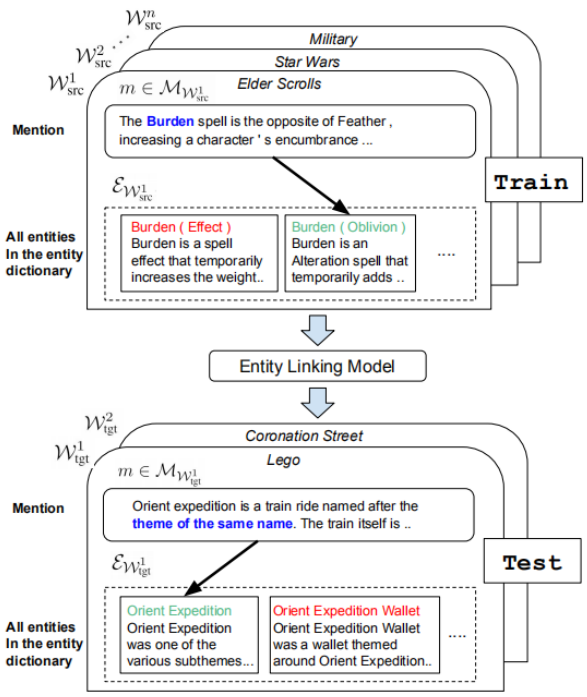
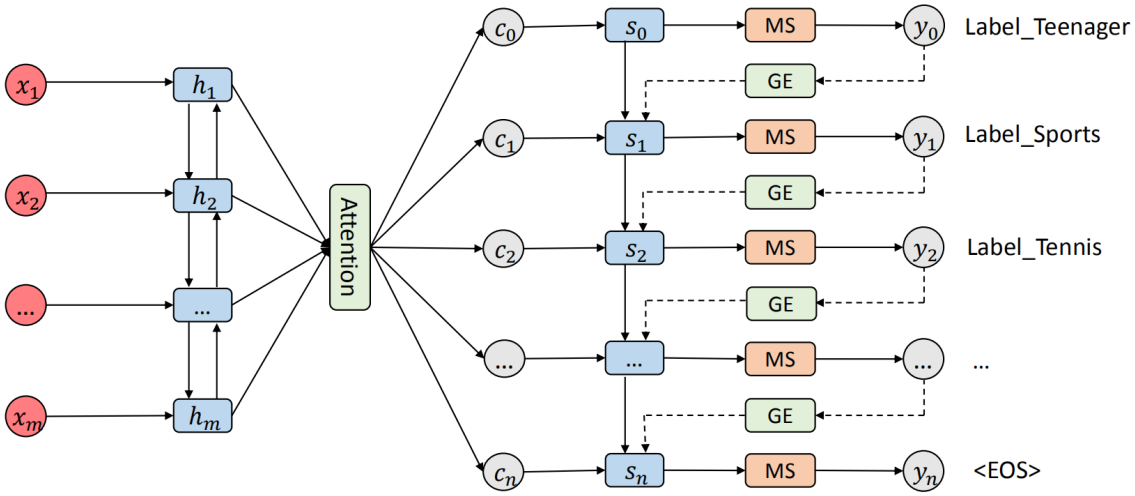


Figure 1: **Zero-shot entity linking.** Multiple training and test domains (worlds) are shown. The task has two key properties: (1) It is **zero-shot**, as no mentions have been observed for any of the test world entities during training. (2) Only **textual** (non-structured) information is available.

9 SGM: Sequence Generation Model for Multi-Label Classification

本文是文本多标签分类领域的一个工作，首次在这个领域引入了seq2seq的结构，并且引入了global embedding解决了exposure bias。通过decoder中的lstm输出，建立了标签之间的相关性。



$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p\left(y_i|y_1, y_2, \cdots, y_{i-1}, \boldsymbol{x}\right) \tag{9.1}$$

Attention：

$$\begin{aligned} e_{ti} &= \boldsymbol{v}_a^T \tanh\left(\boldsymbol{W}_a \boldsymbol{s}_t + \boldsymbol{U}_a \boldsymbol{h}_i\right) \\ \alpha_{ti} &= \frac{\exp\left(e_{ti}\right)}{\sum_{j=1}^{m} \exp\left(e_{tj}\right)} \end{aligned} \tag{9.2}$$

$$c_t = \sum_{i=1}^{m} \alpha_{ti} h_i \tag{9.3}$$

Decoder：

$$s_t = \mathrm{LSTM}\left(s_{t-1}, \left[g\left(\boldsymbol{y}_{t-1}\right); \boldsymbol{c}_{t-1}\right]\right) \tag{9.4}$$

$$\begin{aligned} \boldsymbol{o}_t &= \boldsymbol{W}_o f\left(\boldsymbol{W}_d \boldsymbol{s}_t + \boldsymbol{V}_d \boldsymbol{c}_t\right) \\ \boldsymbol{y}_t &= \mathrm{softmax}\left(\boldsymbol{o}_t + \boldsymbol{I}_t\right) \end{aligned} \tag{9.5}$$

$$\left(I_t\right)_i = \begin{cases} -\infty & \text{if the label } l_i \text{ has been predicted at previous } t-1 \text{ time steps.} \\ 0 & \text{otherwise.} \end{cases} \tag{9.6}$$

GLobal Embedding:

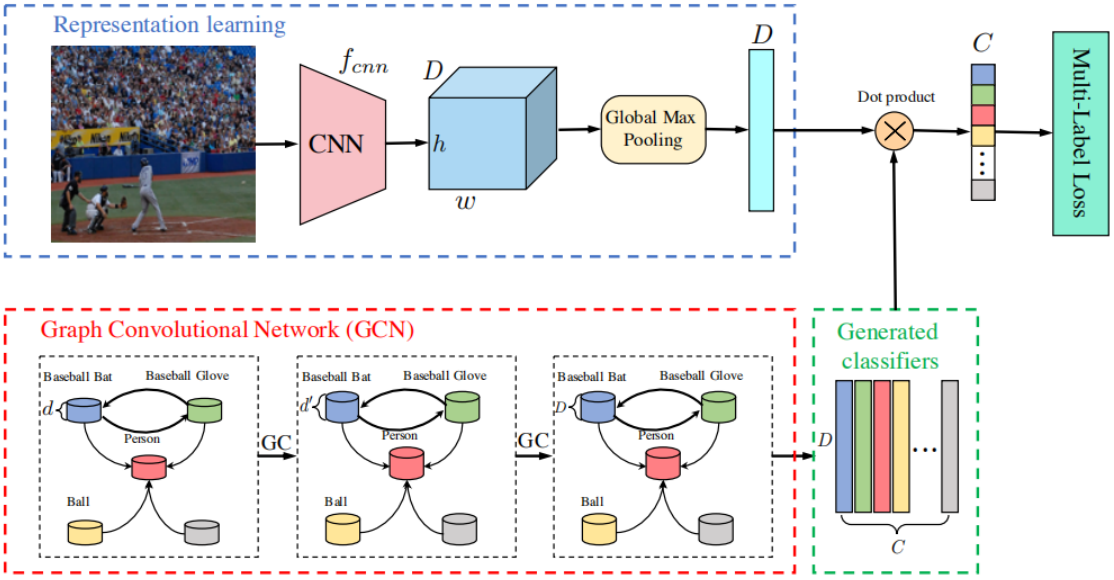
$$\overline{e} = \sum_{i=1}^{L} y_{t-1}^{(i)} e_i \tag{9.7}$$

$$g\left(\boldsymbol{y}_{t-1}\right) = (\boldsymbol{1} - \boldsymbol{H}) \odot \boldsymbol{e} + \boldsymbol{H} \odot \overline{\boldsymbol{e}} \tag{9.8}$$

$$\boldsymbol{H} = \boldsymbol{W}_1 \boldsymbol{e} + \boldsymbol{W}_2 \overline{\boldsymbol{e}} \tag{9.9}$$

10 Multi-Label Image Recognition with Graph Convolutional Networks

这篇文章是CV领域的多标签分类问题。亮点是对label用stacked GCN来处理，从而解决其相关性问题。



$$\boldsymbol{H}^{l+1} = h\left(\widehat{\boldsymbol{A}}\boldsymbol{H}^{l}\boldsymbol{W}^{l}\right) \tag{10.1}$$

其中$\boldsymbol{W}^{l}$是一个学习参数，$\widehat{\boldsymbol{A}}$是相关性矩阵的正则化后的结果。

stacked GCN的第一层输入是$Z \in \mathbb{R}^{C\times d}$，最后一层输出是$\boldsymbol{W} \in \mathbb{R}^{C\times D}$

于是，可以得到C个label的预测结果：$\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{x}$

GCN中的C个node即为C个label上的分类器。