- 第15课：自然语言生成器 (Natural Language Generation, NLG)
- 第16课：指代消解 Coreference Resolution

第15课：自然语言生成器 (Natural Language Generation, NLG)

回顾：

- 语言模型
- RNN-LM用来生成下一个词
- Conditional LM：根据前面词预测下一个词，并且有其他输入x。（机器翻译、摘要等）
- 在训练过程中，在decoder中输入已知最优方法，而不是decoder前面的预测结果，这种方法叫Teacher Forcing
- 解码搜索方法：greedy search、beam search
- beam search的k选择遇到的挑战：太小会更加类似greedy search，太大计算量较大，并且可能降低BLEU score，或者让输出过于常见（在聊天对话任务中）。

新的搜索方式：基于 采样 Sample 的解码

- 纯采样：每一步采用概率分布进行随机采样
- top-n采样：每一步采用概率分布对top-n的可能词汇进行随机采样

配合各类解码：Softmax 温度参数（temperature）

$$P_t(w) = \frac{\exp(s_w/\tau)}{\sum_{w' \in V} \exp(s_{w'}/\tau)}$$

[NLG任务]：文本摘要

- 文本摘要任务定义：给定输入文本x，形成一个摘要y，它更短但包括了x的主要部分。
- 文本摘要可以分为单文档摘要和多文档摘要。
- 单文档摘要的一些数据集：Gigaword、LCSTS、NYT、CNN/DailyMail、Wikihow，文档简化数据集Simple Wikipedia、Newsela
- github.com/mathsyouth/a...

文本摘要两种主要方式

- 抽取式摘要
- 抽象摘要

神经网络之前的文本摘要

- 绝大部分是抽取式摘要
- 整体流程：内容选取->信息重排->句子实现
- 内容选取算法
  - 句子打分算法：根据关键词、tf-idf，还有句子出现位置等特征
  - 基于图的算法：计算句子之间的相关性，用图算法寻找中心句子

文本摘要评价方法：**ROUGE** (Recall-Oriented Understudy for Gisting Evaluation)

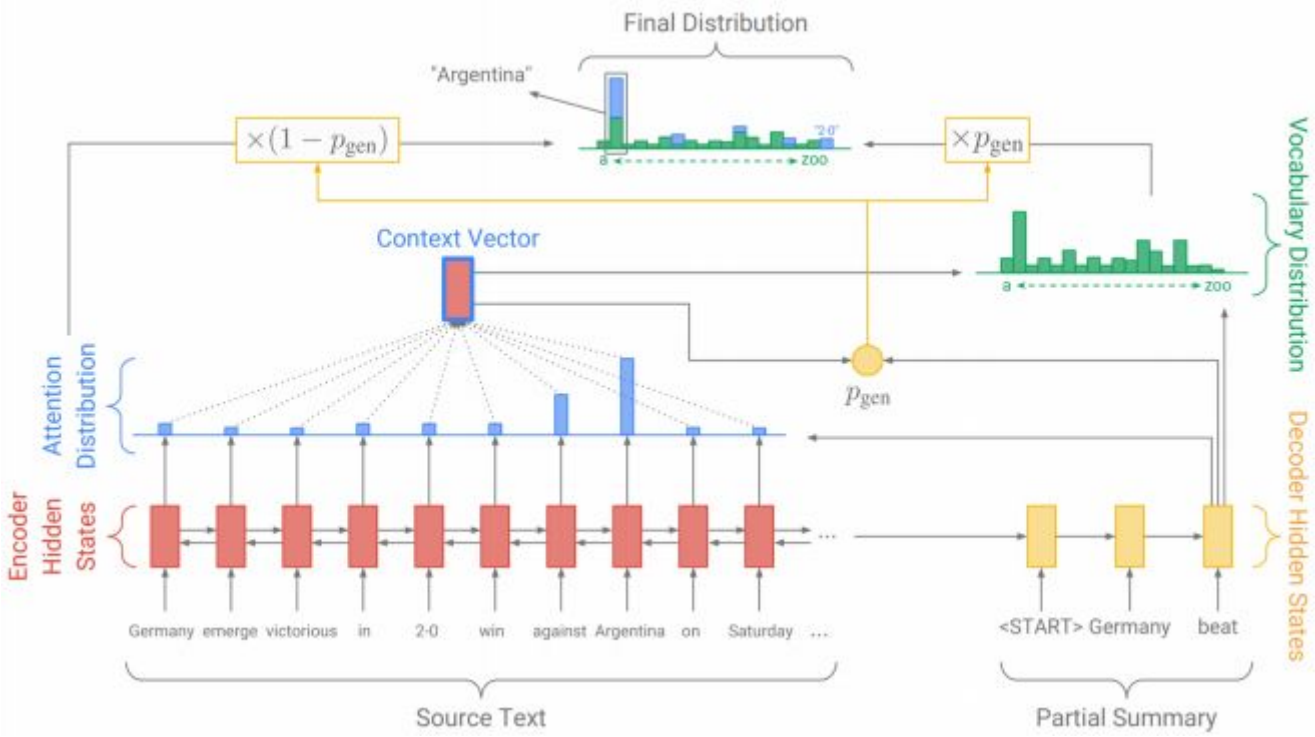
- 和BLEU一样基于n-gram重叠，但没有简短惩罚
- 基于recall，不像BLEU基于precision
- 分别提供基于不同n-gram的分数，比如ROUGE-1（unigram）、ROUGE-2（bigram）、ROUGE-L（LCS）

神经网络文本摘要的发展（2015~）

- 第一篇seq2seq摘要论文， Rush et al.，2015
- 更倾向于复制，但避免过多复制
- 层次、多层attention
- 更全局、高层次内容选择
- 使用增强学习直接最大化ROUGE
- 将过去的方法（比如图算法）融合到神经网络系统中

神经网络文本摘要：**复制机制**

- seq2seq+attention系统擅长流畅输出，但缺乏对一些细节的复制（比如罕见词）
- 复制机制的引入，让seq2seq能容易复制一些词和短语
- 方法之一：Get To The Point: Summarization with Pointer-Generator Networks, See et al, 2017
  - 通过 p_{gen} 来混合生成词分布和复制词（attention）分布。
- 复制机制的最大问题是可能复制太多，几乎变成了一个抽取式模型



Get To The Point: Summarization with Pointer-Generator Networks

神经网络文本摘要：**更好的内容选取**

- 神经网络之前的文本摘要系统明确区分了内容选取、句子实现这两个阶段
- 而标准seq2seq+attention文本摘要模型混合了两阶段，缺乏全局性的内容选取策略
- 方法之一：自下而上的文本摘要（**bottom-up summarization**）
  - 内容选择阶段：采用神经网络sequence-tagging模型标注哪些词要选择
  - 自下而上注意力阶段：仅仅针对被选择的词进行注意力计算。
  - 优点：足够好的全局内容选择策略，避免长序列的过多复制。

神经网络文本摘要：**增强学习**

- A Deep Reinforced Model for Abstractive Summarization, Paulus et al, 2017
- RL能够获得更高的ROUGE分数，但人工评价可读性不佳，ML+RL混合模型会更好

[NLG任务]：对话

对话任务

- 任务导向：助理Assistive、联合Co-operative、对抗Adversarial
- 社交对话：聊天机器人、精神治疗

对话系统发展

- 由于开放式自由NLG的困难性，神经网络之前的对话系统通常使用预设模板，或者从语料集中检索

- 2015年后开始出现seq2seq神经网络对话系统
- 标准的seq2seq+attention模型用于聊天机器人，遇到的困难
  - 一般性Genericness，无聊的回复
  - 不相干的回复
  - 重复
  - 缺乏上下文
  - 缺乏始终如一的人物角色

对话系统：**不相干问题**

- seq2seq通常会产生一些不相干的回复，要么是常见的通用回答，要么是内容无关
- 解决方法：最大互信息量（MMI）

$$\log \frac{p(S,T)}{p(S)p(T)}$$

$$\hat{T} = \arg\max_{T} \left\{ \log p(T|S) - \log p(T) \right\}$$

MMI

对话系统：**一般性、无聊回复**

- 简单解决：对罕见词提权、采用取样方式decode
- 采用条件解决：采用一些额外内容作为条件；采用检索-调整模型，而不是直接生成

对话系统：**重复**

- 简单解决：在decode搜索时对重复n-gram进行屏蔽
- 复杂方案：训练一个覆盖机制，避免注意力机制多次注意到相同的词。

对话系统：**缺乏始终如一的人物角色**

- A Persona-Based Neural Conversation Model, Li et al 2016
- Personalizing Dialogue Agents: I have a dog, do you have pets too?, Zhang et al, 2018

谈判对话 Negotiation dialogue

- Deal or No Deal? End-to-End Learning for Negotiation Dialogues, Lewis et al, 2017
  - 先采用seq2seq系统和标准最大似然训练了一个流畅但缺乏策略性的对话功能
  - 然后采用增强学习，优化离散奖励

    - RL目标和ML目标进行融合（纯RL可能会为了对抗而偏离英语）
- Hierarchical Text Generation and Planning for Strategic Dialogue, Yarats et al, 2018
    - 将策略层面和NLG层面分拆
    - 每个表达x_t都有一个对应的离散潜变量z_t，z_t用于训练预测对话未来事件而不是x_t本身，也就是预估x_t在对话中的效果

对话问答：**CoQA**

- CoQA: a Conversational Question Answering Challenge, Reddy et al, 2018
- 在对话上下文中进行问答，结合了问答、阅读理解、对话三项任务

讲故事 Storytelling

- 任务：看图讲故事
    - 缺乏并行数据，需要使用常识编码空间
    - 使用skip-thought vectors，先把图片映射到这个向量空间内，然后使用目标预料，训练decoder。[Skip-Thought Vectors, Kiros 2015]
- 任务：看提示词讲故事
    - Hierarchical Neural Story Generation, Fan et al, 2018
        - 采用基于卷积的seq2seq
        - Gated multi-head multi-scale self-attention
        - 先采用一个seq2seq模型训练通用LM，然后用另外一个seq2seq根据提示词学习
- 任务：续讲故事
- 重大挑战：语言模型能够语法流畅，但缺乏故事的事件脉络
- Event2event Story Generation [Event Representations for Automated Story Generation with Deep Neural Nets, Martin et al, 2018]
- Structured Story Generation [Strategies for Structuring Story Generation, Fan et al, 2019]
- 在NLU（自然语言理解）领域有很多方法追踪事件、实体、状态，但用于NLG有更多困难，如果约束在狭窄领域内会更容易一些
    - 菜谱领域的状态追踪 [Simulating Action Dynamics with Neural Process Networks, Bosselut et al, 2018]

生成诗歌：**Hafez**

- [Hafez: an Interactive Poetry Generation System, Ghazvininejad et al, 2017]
- 用有限状态接收器（FSA）来定义所有符合韵脚的可能序列，然后驱动RNN-LM的输出

非自回归神经网络机器翻译模型（**Non-autoregressive Neural Machine Translation**）

- [Non-Autoregressive Neural Machine Translation, Gu et al, 2018]

- 同步产生翻译结果，而不是从左往后逐步生成。

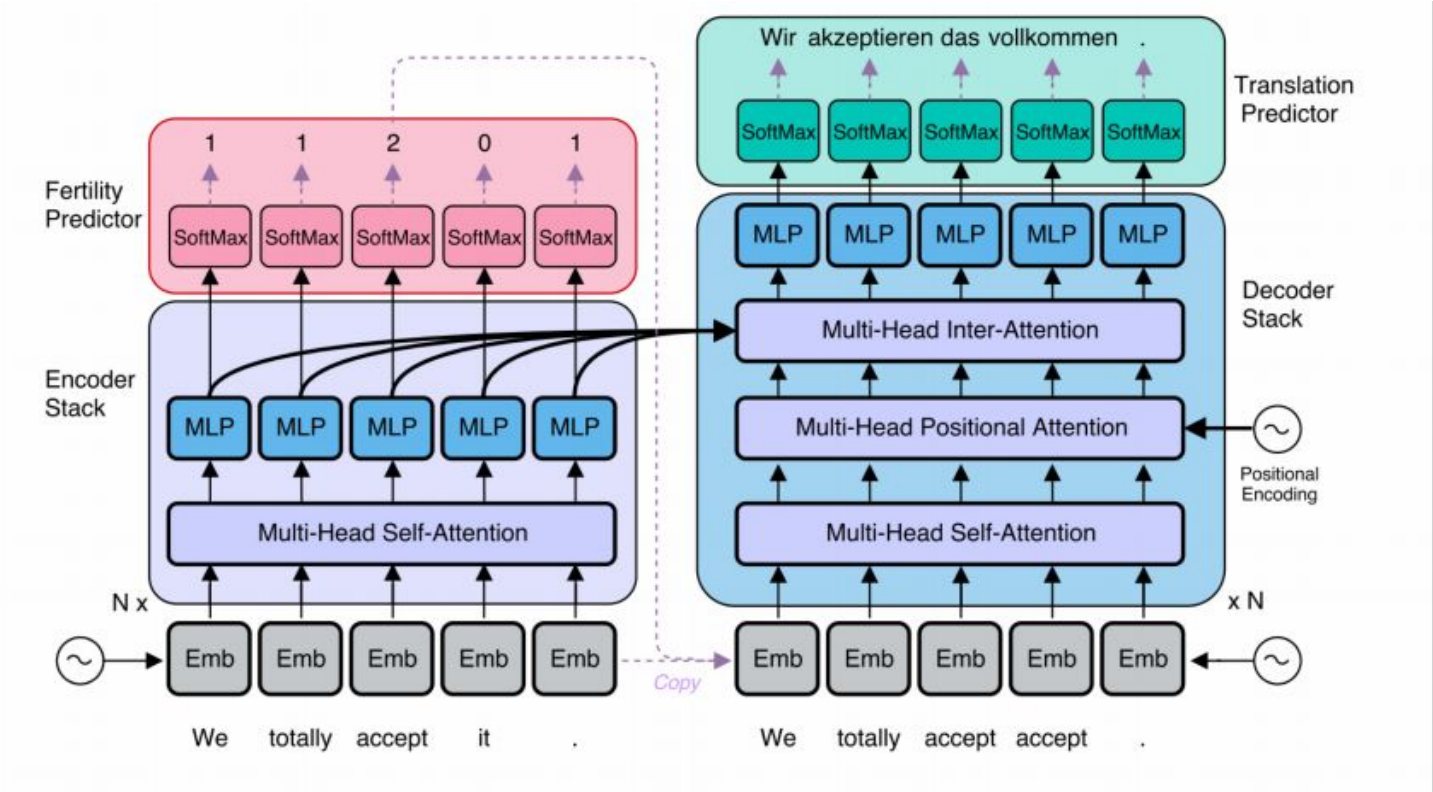


Figure 2: The architecture of the NAT, where the black solid arrows represent differentiable connections and the purple dashed arrows are non-differentiable operations. Each sublayer inside the encoder and decoder stacks also includes layer normalization and a residual connection.

Non-autoregressive Neural Machine Translation

[NLG评估]

自动评价方式

- 基于词汇覆盖的评价方式： (BLEU, ROUGE, METEOR, F1, etc.)
  - 对于机器翻译并不够好
  - 对于摘要和对话可能更糟
- [How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, Liu et al, 2017]
- [Why We Need New Evaluation Metrics for NLG, Novikova et al, 2017]
- 困惑度 Perplexity，只能看出LM的强大性，但无法用于生成
- 基于word embedding的评价方式，仍然无法更好在开放式任务上与人工评价比拼

无法找到自动的通用评价标准，只能有**一些角度**：

- 流畅度（训练好的LM给出的概率）
- 正确风格（针对目标预料训练的LM给出的概率）
- 多样性
- 与输入的相关性

- 简单指标：长度、重复性等
- 其他与任务相关的指标，比如摘要的压缩比

人工评估并没有解决所有问题

- 前后矛盾、不合逻辑、注意力不集中、误解问题、一些不可解释性
- （他们团队在做的一些解决尝试）

第16课：指代消解 Coreference Resolution

应用场景

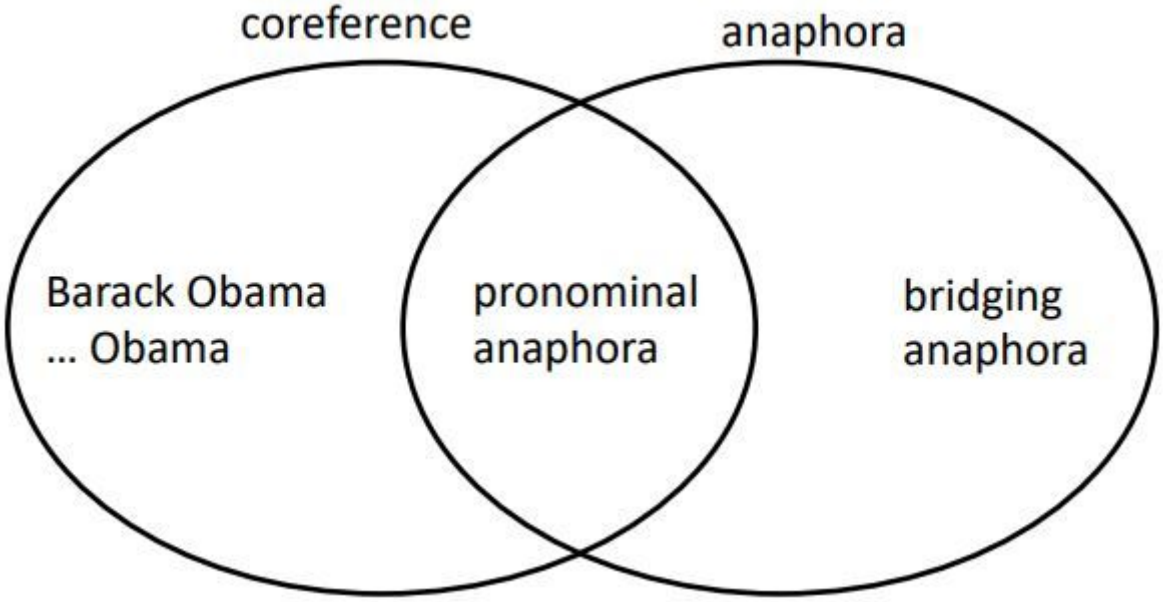
- 全文理解：信息抽取、问答、摘要
- 机器翻译
- 对话系统

指代消解的步骤

- 识别出提及词语
  - 代词：词性标注 PoS tag
  - 名称：命名实体识别NER系统
  - 名词词组：专用解析器
  - 对于一些不好的提及词语，可以分类过滤掉，也可以作为候选在后面聚类时去除
- 对提及词语进行聚类
- （当然也有端到端一次性的模型解决，而不用分成两步）

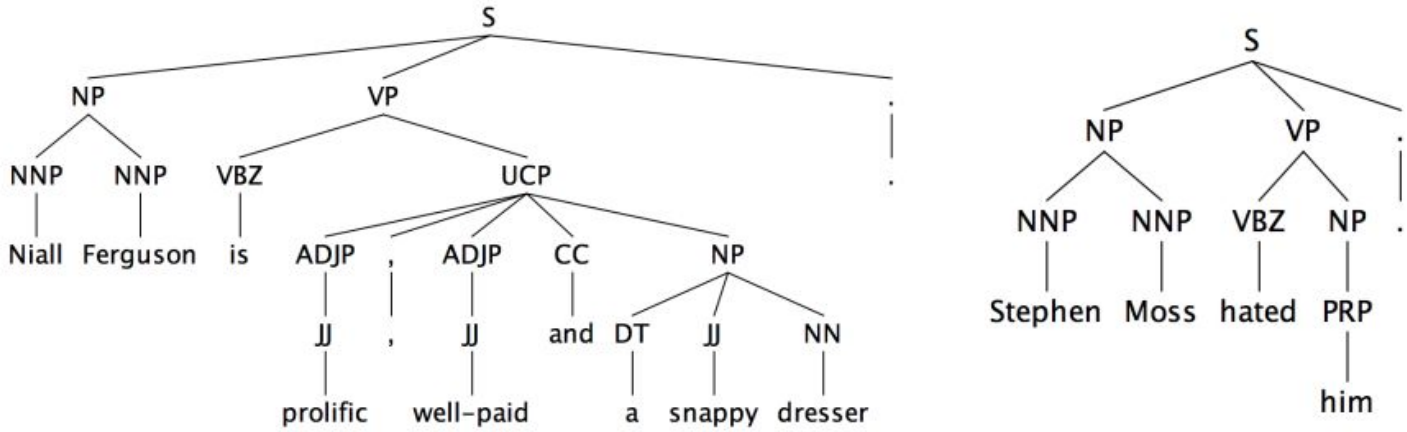
语言学定义

- 指代 **Coreference**：多个词汇表示现实中同一样东西
- 首语重复法 **Anaphora**，用回指词指向前面的先行词



指代模型1：Rule-based (pronominal anaphora resolution)

- 传统的代词首语重复法识别算法： Hobbs' naive algorithm (1976)
- 代词指代存在知识背景问题，比如“市议会拒绝给示威者颁发许可，因为他们[担心/宣扬]暴力。”（Winograd Schema Challenge）



Hobbs' naive algorithm

指代模型2：Mention Pair

- 对任意一对提及词之间是否指代相同的概率p(m_i, m_j)建模
- 预测两两关系，然后用传递闭包做聚类
- 缺点：大部分提及词只会有一个明确先行词，然而这里却全部预测，可能会很多

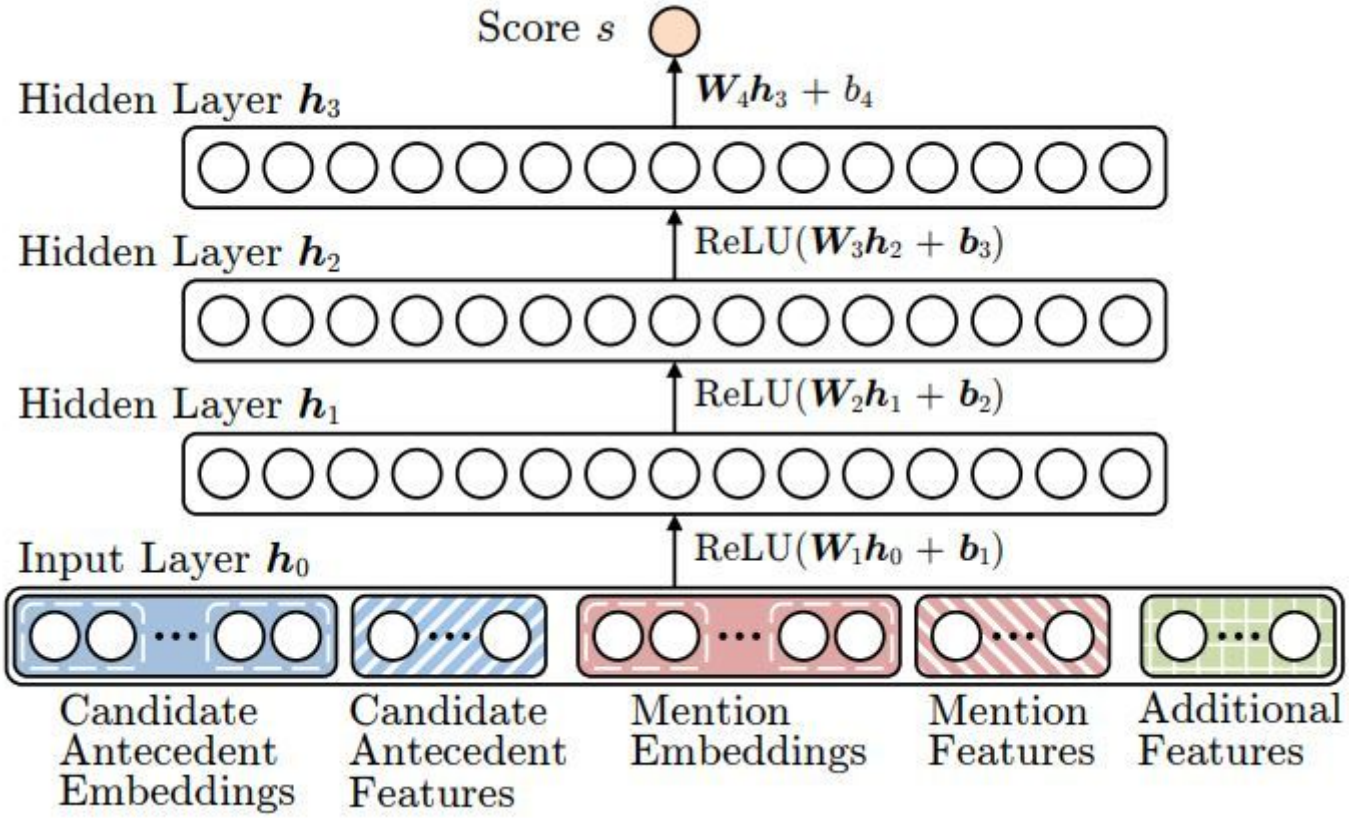
指代模型3：Mention Ranking

- 认为对于每个词，其他词是否指代相同是一个整体概率分布

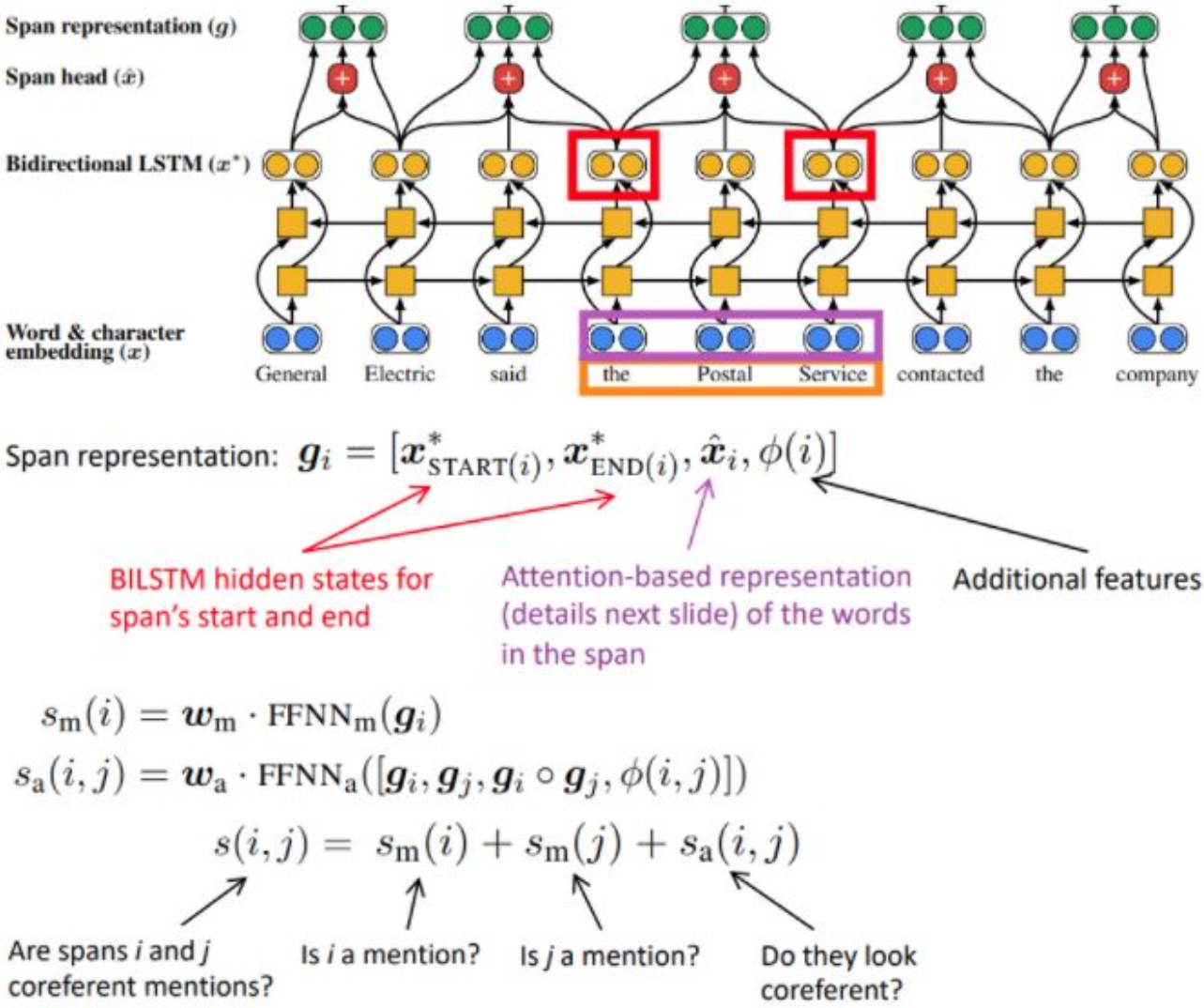
- 选择其中概率最大的确定为先行词，再通过传递闭包做聚类

概率计算模型

1. 基于各类特征的非神经网络简单机器学习
2. 简单的前馈神经网络模型
3. 端到端的神经网络模型（目前行业最好方法）[Kenton Lee et al. from UW, EMNLP 2017]
   1. 基于前馈神经网络，使用LSTM、attention
   2. 不需要做提及词的识别，把每一个词都作为候选
   3. 采用所有的1-gram, 2-gram, ..., n-gram范围作为span计算范围



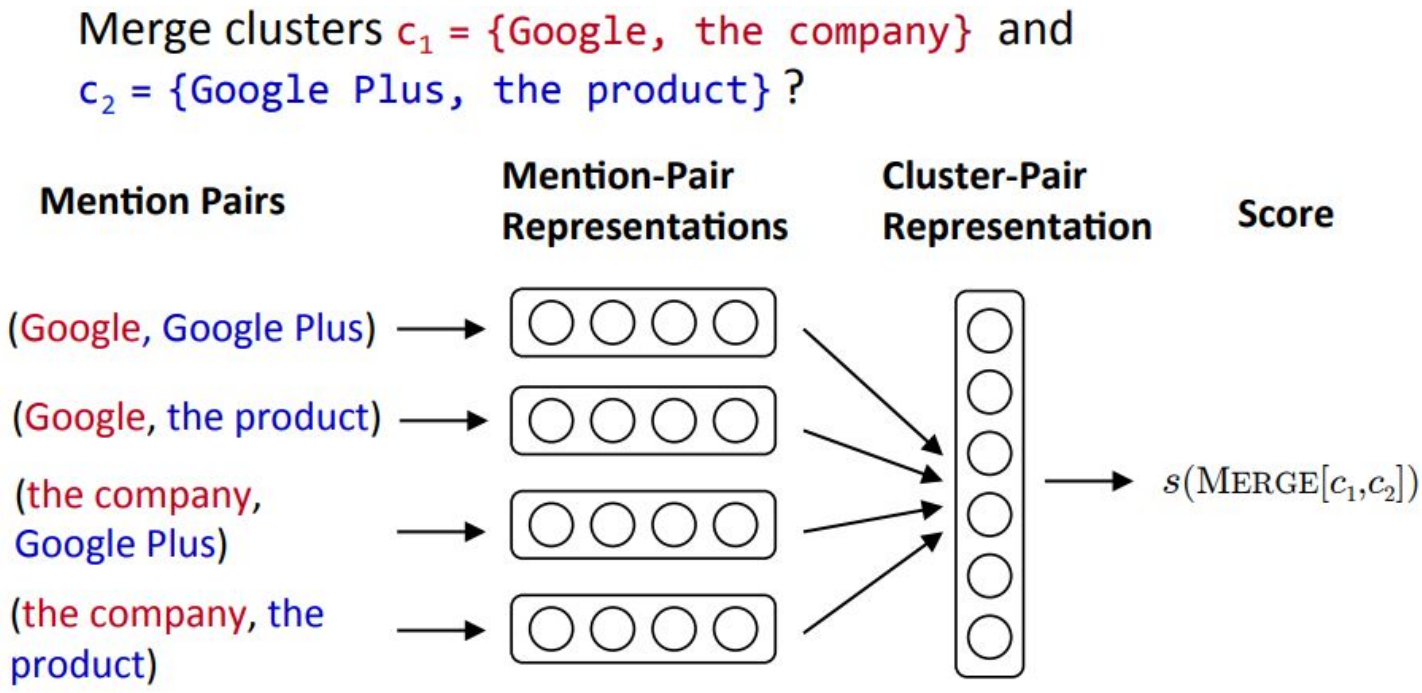
简单的前馈神经网络模型



端到端的神经网络模型

指代模型4：Clustering

- 以每一个提及词作为一个单独聚类开始
- 通过计算任意两个聚类之间是否有指代，来聚合聚类
- mention-pair的表示分别进行最大池化、平均池化，然后两者拼接成为cluster-pair表达



指代模型的评估方法

- 计算理想聚类和实际聚类之间的关系
- MUC, CEAF, LEA, B-CUBED, BLANC等等