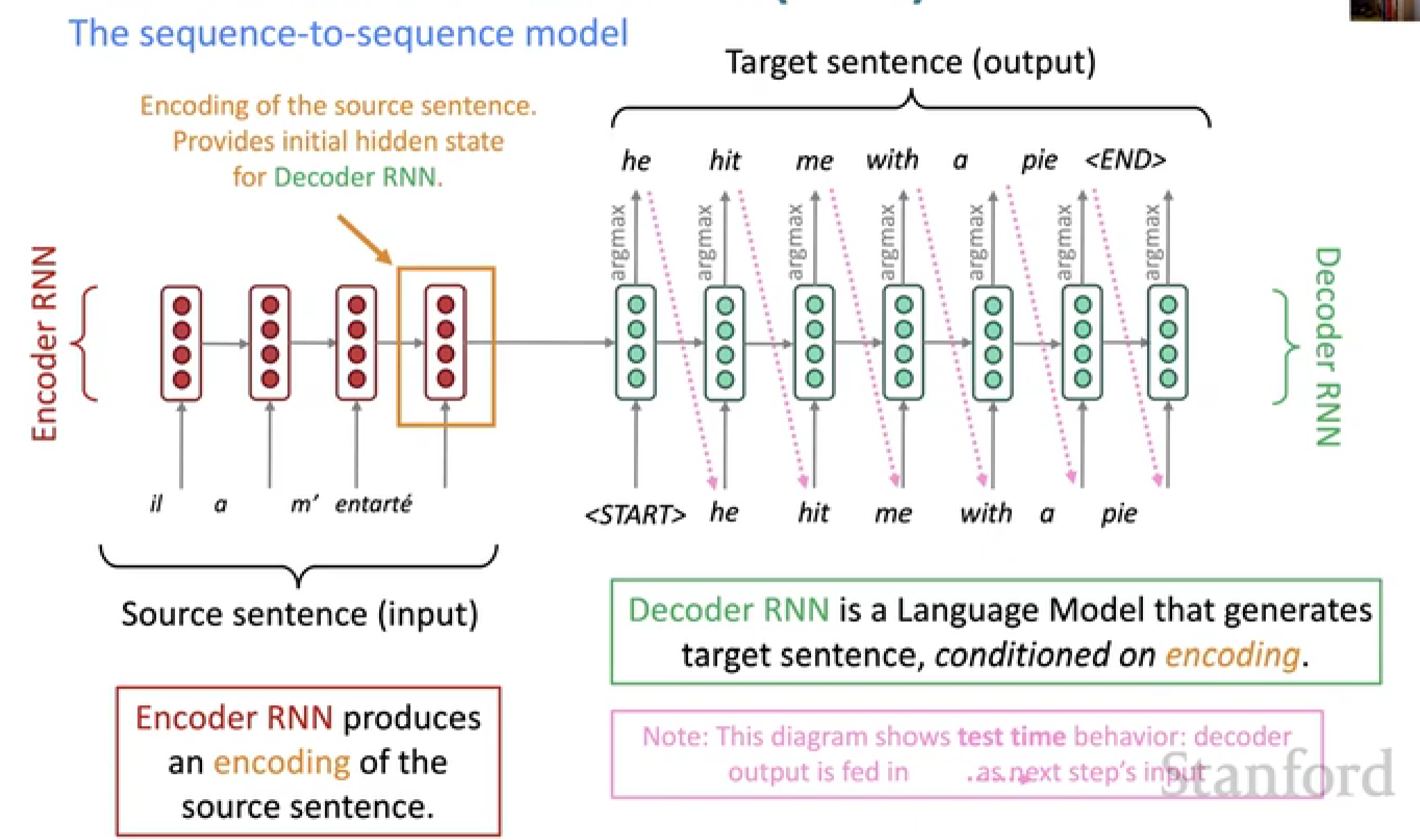
## 机器学习课程

Lecture7

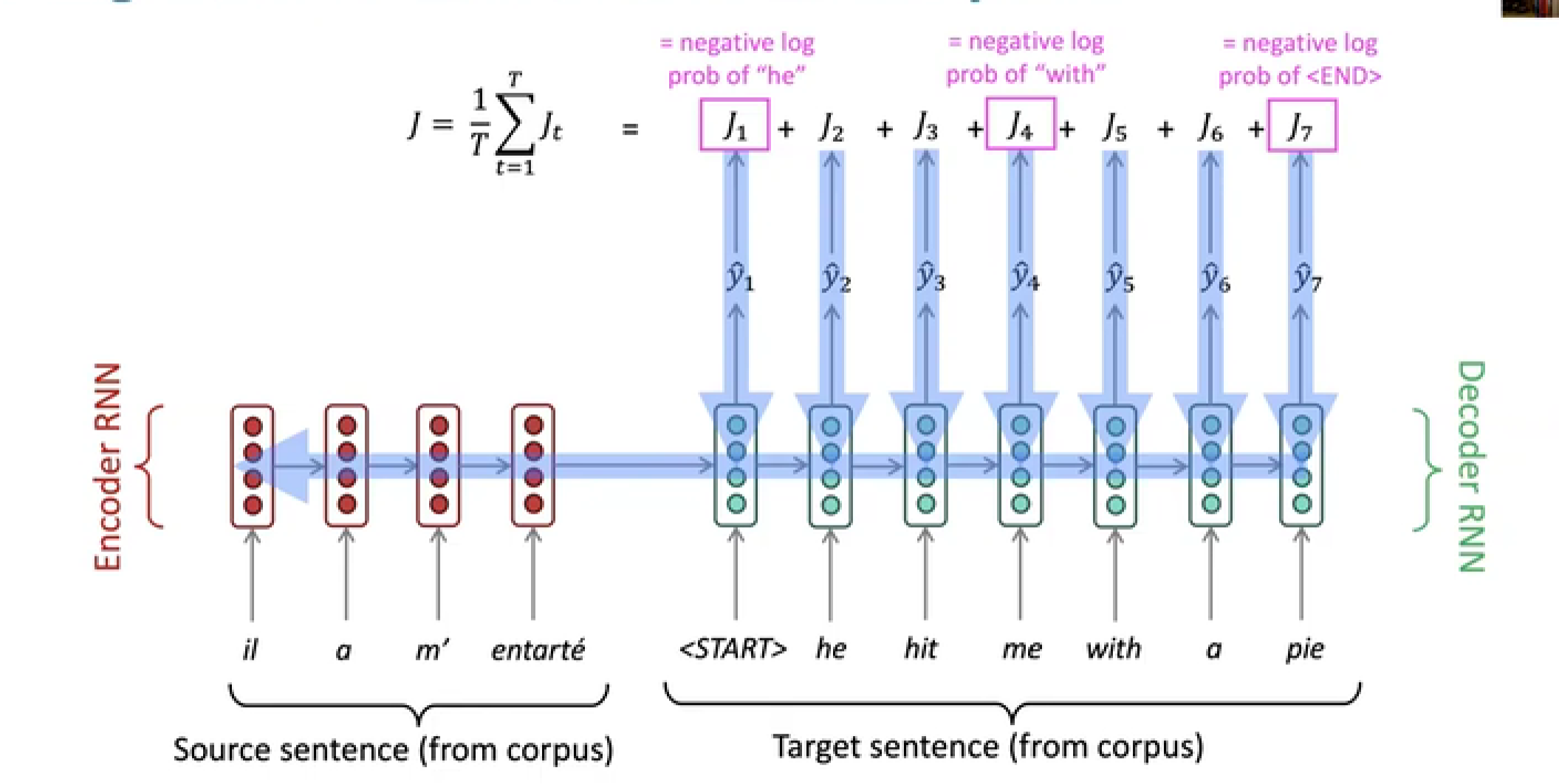
① 神经网络机器翻译（Neural Machine Translation）是2014年以后来发展起来的一种先进的机器翻译方法，利用神经网络模型来模拟人类翻译的过程，实现自动的文本翻译。是一种序列到序列的模型（seq2seq）。如RNN就包含其中。

②一个常见的seq2seq的model：



一般分为编码器和解码器。编码器通过把输入的源语言转换为隐藏的序列表示，输入给解码层，解码层根据这些隐藏状态来生成目标语言。

③训练这个模型的过程可简化为：



模型首先经过encoder进行编码，在decoder产生的单词处，我们计算实际产生的单词和目标词汇之间的loss，基于这个loss对参数反向传播进行更新。作为一个端到端的model，模型的encodel和decoder的参数在这个过程中都被更新了。

④多层堆叠的RNN/Lstm往往比单层的效果更好，能够预测更高级的语言特征。

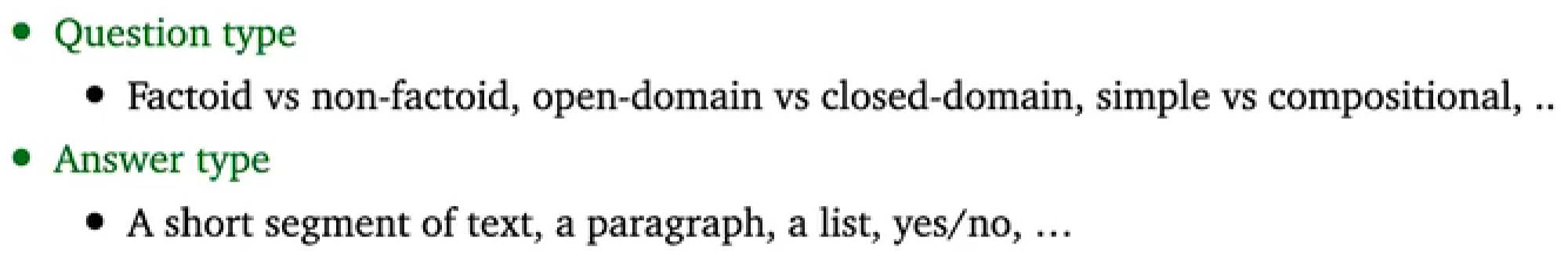
⑤Greedy Decoding：一种在序列生成任务中常用的解码方法。其基本原理是在每个时间步选择当前条件下概率最高的词语或符号作为输出，然后基于这个输出继续预测下一个时间步的输出。

Beam Search Decoding：在每一步解码时，不是简单地选择当前概率最高的词汇，而是保留概率最高的前K个候选序列，基于这K个候选序列，分别计算它们扩展后的新序列的概率，并再次保留概率最高的前K个新序列。

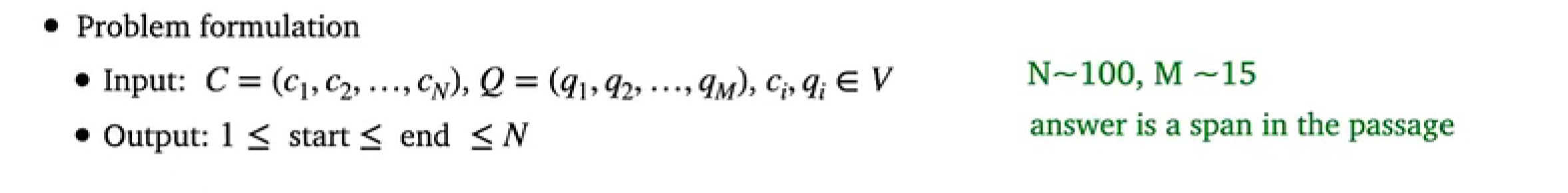
⑥评估机器翻译效果好坏的方法BLUE：BLEU是一种基于n-gram的自动评价指标。通过比较机器翻译输出与一个或多个参考译文的相似度来评估翻译的质量。

Lecture11

①Question answering:建立一个能够自动回答问题的系统。其中question和answer的种类如下所示：



②用于阅读理解的神经网络模型的问题范式：



输入一个段落，有n个词，给出M个问题，输出是段落中的一个子序列或一个词语。

③双向注意力流模型（BiDAF）用于阅读理解任务：

Encoder：

 将段落中的每个词转化为词向量表示。常用的词嵌入方法包括预训练的词向量（如GloVe或Word2Vec）。将问题中的每个词转化为词向量表示。使用卷积神经网络（CNN）或双向长短期记忆网络（BiLSTM）等方法，对段落和问题的词嵌入进行特征提取，得到段落和问题的高维特征表示。

在BiDAF模型中，注意力机制起到了关键作用。双向注意力机制包括以下两个部分：

1.问题到段落注意力：

计算段落中每个词对问题中每个词的注意力权重，生成问题到段落的注意力表示。

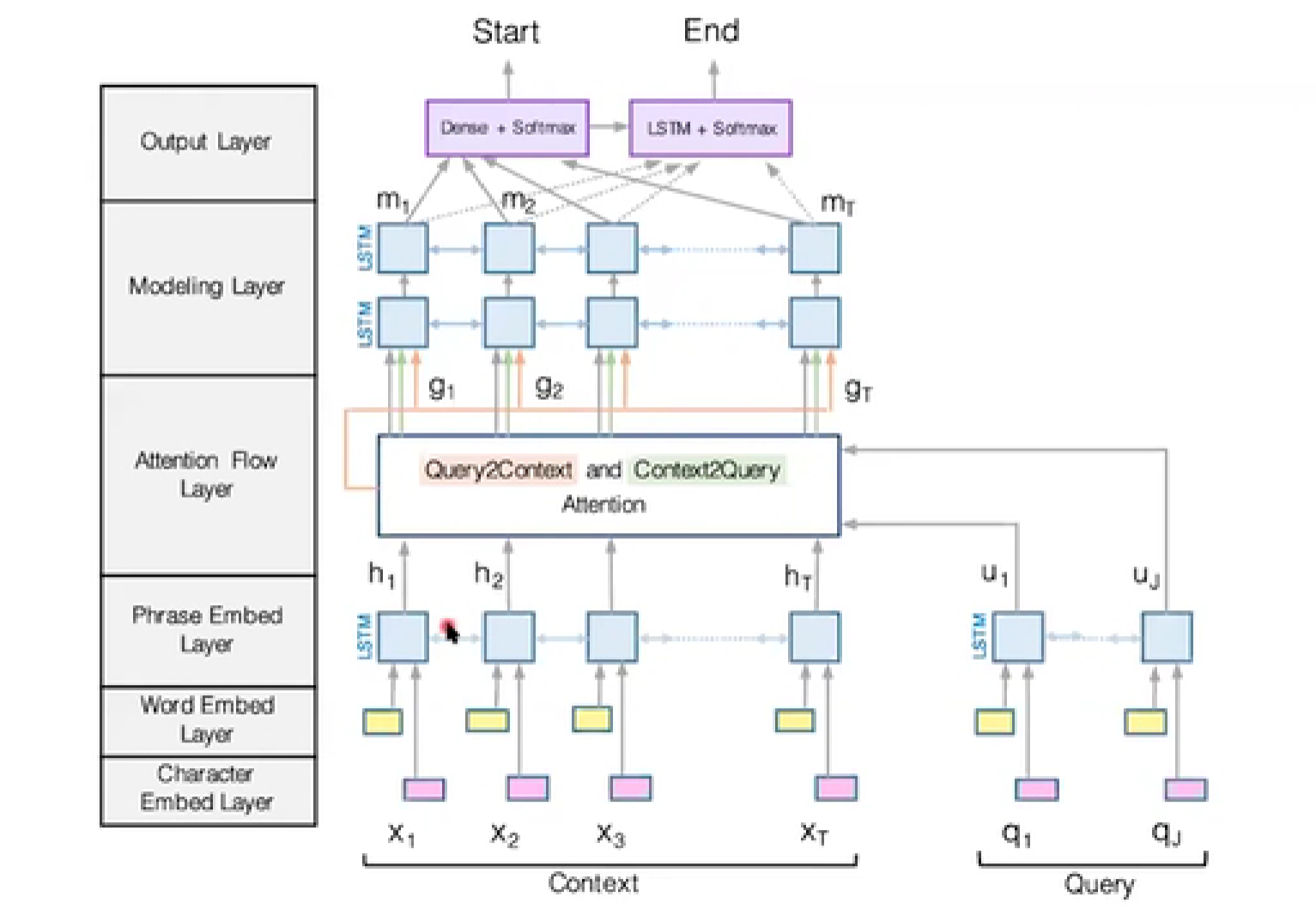
2. 段落到问题注意力：

计算问题中每个词对段落中每个词的注意力权重，生成段落到问题的注意力表示。

最后将原始段落表示与注意力表示结合，形成最终的上下文表示。

Decoder:

将上下文表示输入到双向LSTM中进行进一步编码，得到最终的上下文表示。预测答案在段落中的起始位置和结束位置。BiDAF模型通过指针网络来实现这一点。



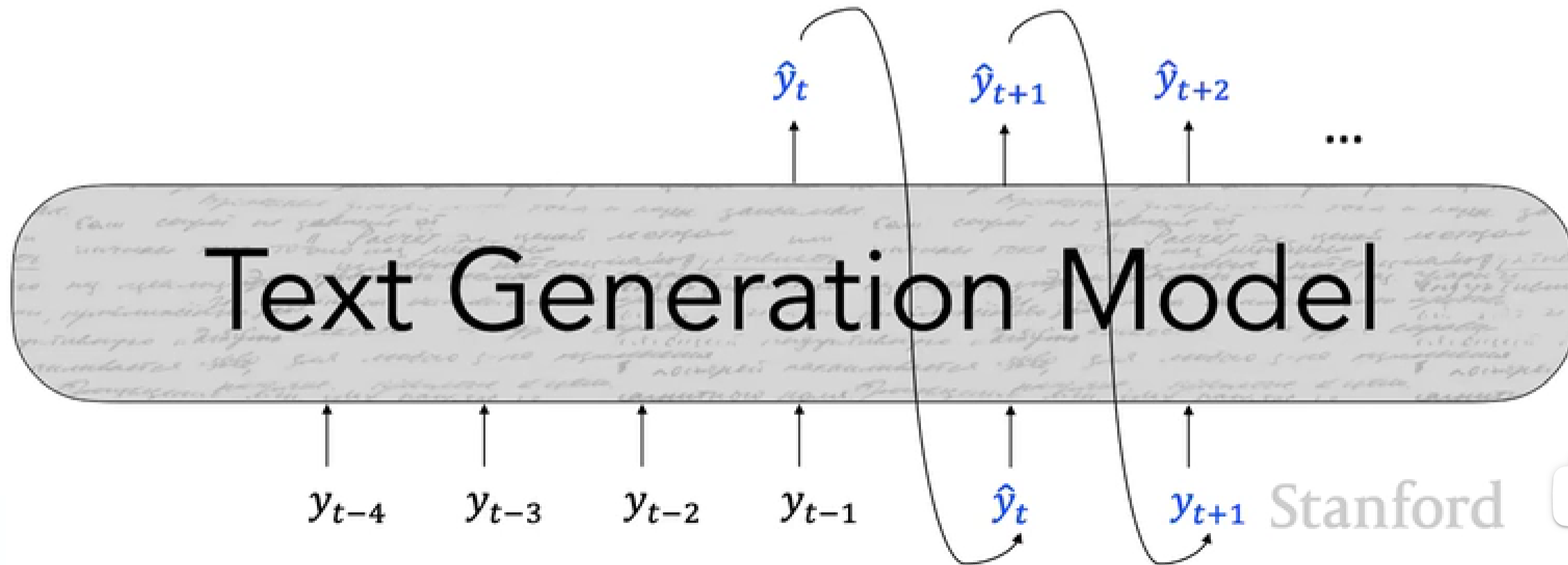
④Bert用于阅读理解任务：效果远优于BiDAF，这主要归功于其强大的语言表示能力和Transformer结构的双向性。这使得BERT能够更准确地理解问题的意图，并在文章中定位到相关的答案信息。

⑤开放域问题回答（Open-Domain Question Answering）：在让机器能够理解和回答来自开放领域的问题。这种类型的问题回答系统通常需要处理大量文本数据，并从中提取相关信息以生成答案。

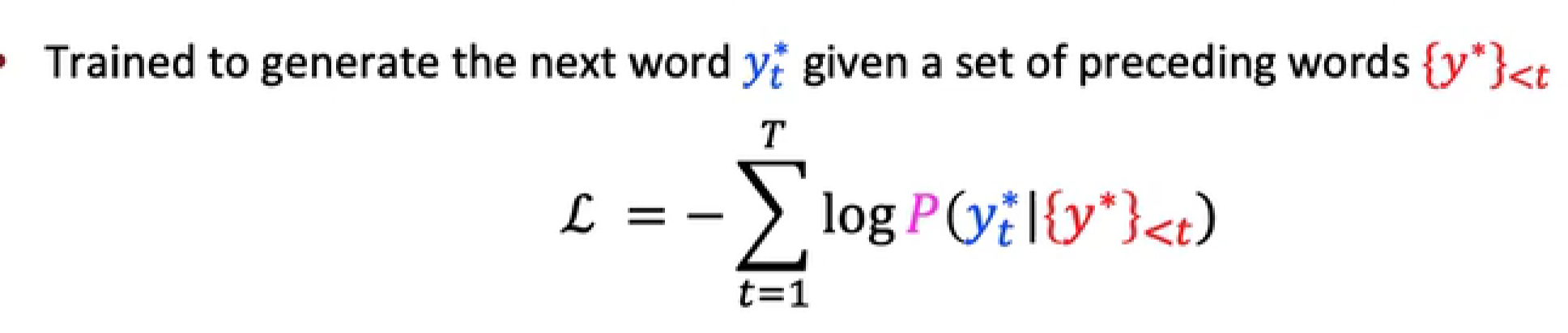
Lecture12

①自然语言生成（NLG）是自然语言处理（NLP）中的一个重要领域，其目的是通过计算机生成自然语言文本。NLG可以应用于许多场景，如新闻自动生成、对话系统、报告生成和内容摘要等。

②NLG系统：输入为数个token，输出的token返回输入。



③优化预测下一个词为正确词汇的概率，取其对数似然的相反数。

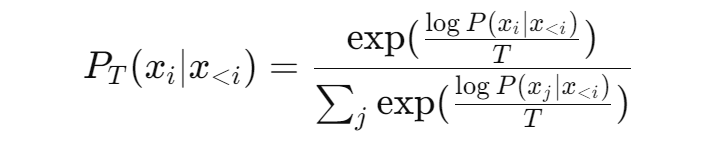


在decoder时，选取最大的概率词，通常称作argmax decoder。这是较为常用的方法。但这样的方法也存在问题：如可能会出现持续的重复词汇。要消除这样的重复情况，可以检查生成文本中的N-gram，如果检测到重复的N-gram，则通过调整生成策略或重新生成部分文本来避免重复。

另一种采样方法是Top-k采样。它是一种在自然语言生成（NLG）任务中常用的采样方法，用于增加生成文本的多样性并减少生成低概率词的可能性。Top-k采样通过在每一步生成时仅从概率最高的k个词中进行采样，从而限制了候选词的范围。由于Top-k采样仅考虑概率最高的前k个词，这意味着低概率的词被排除在外，多样性受限，并且很收到K值的影响。

Top-p采样比Top-k采样更加灵活，通过根据累积概率选择候选词集合，从而在增加生成文本多样性的同时，避免生成低概率词。与Top-k采样相比，它根据概率分布动态调整候选词的数量。

④温度缩放（Temperature Scaling）：是一种用于调整生成模型输出概率分布的技术。温度缩放通过调整softmax函数中的温度参数 TTT 来改变概率分布的“平坦度”或“尖锐度”。



**T<1**：降低温度，使概率分布更加尖锐，模型倾向于生成高概率词，文本生成更加确定性和连贯，但多样性减少。

**T>1**：提高温度，使概率分布更加平坦，模型倾向于生成更多样的词，文本生成的多样性增加，但可能会引入低概率词，生成文本的连贯性可能降低。

⑤Rerank来提高decoder效果：基本思想是对初始生成或检索到的候选结果进行二次处理，通过某种评价指标或额外信息重新排序，以选出最符合目标的结果。通常，初始候选集是由一个生成模型或检索系统产生的，然后再通过rerank模型或算法进行调整。

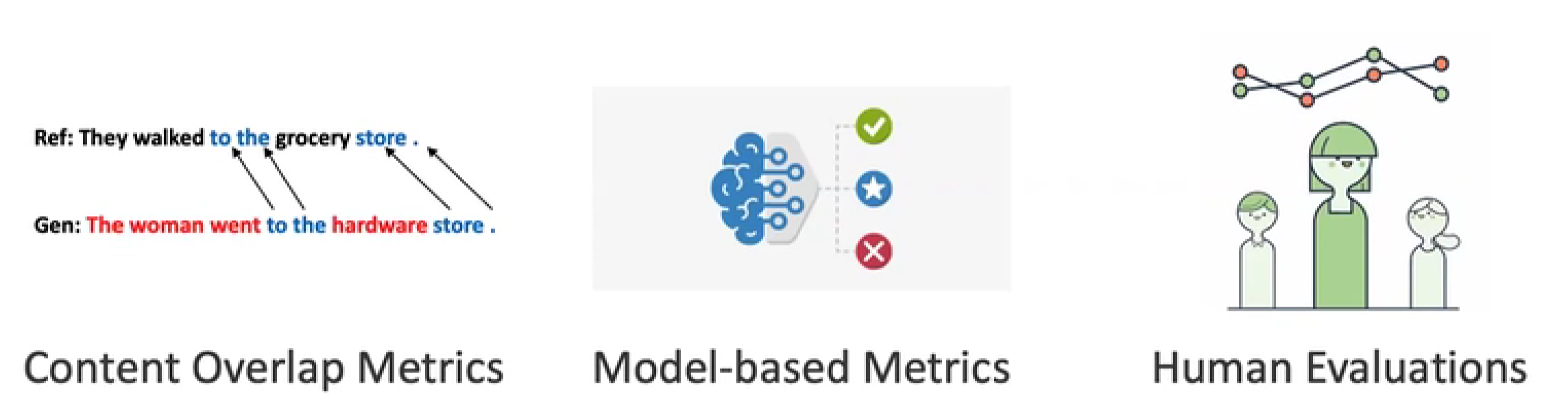
⑥Unlikehood training：通过在损失函数中引入负对数似然项，专门对重复的或不期望的词施加惩罚，从而降低这些词的生成概率。相比于传统的最大似然估计方法，Unlikehood training 更加注重生成文本的多样性和避免重复。

⑦暴露偏差（Exposure Bias）是自然语言处理（NLP）和生成任务中常见的问题，特别是在使用基于最大似然估计进行训练时。暴露偏差指的是模型在训练和推断过程中遇到的不一致性，从而导致生成结果质量下降。

⑧reward的设计：



⑨常见的三种评估模型的方法：

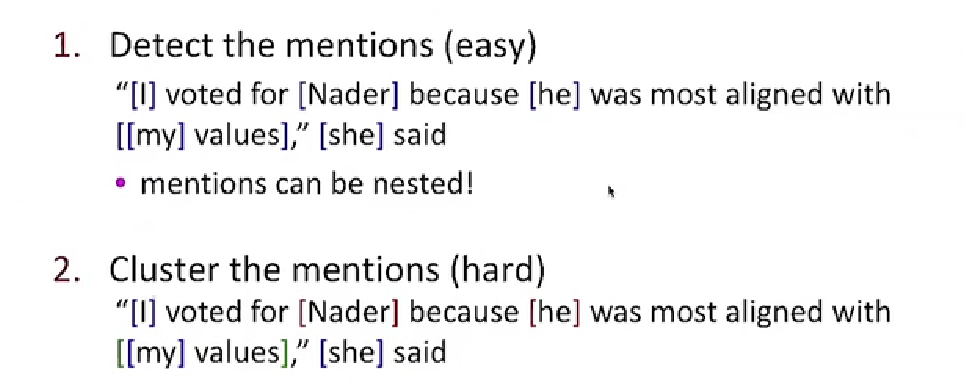


内容重叠矩阵（Content Overlap Matrix）是一种用于量化两个或多个文本之间重叠程度的方法。它可以用于评估生成的文本与参考文本之间的相似性，在nlp中，内容重叠矩阵可以帮助理解文本间的相似性和差异。

Lecture13

①共指消解任务（Coreference Resolution）：旨在识别文本中指代同一实体的所有表达。简单来说，就是找出文本中哪些词或短语指的是同一个事物。

②共指消解的两个步骤：

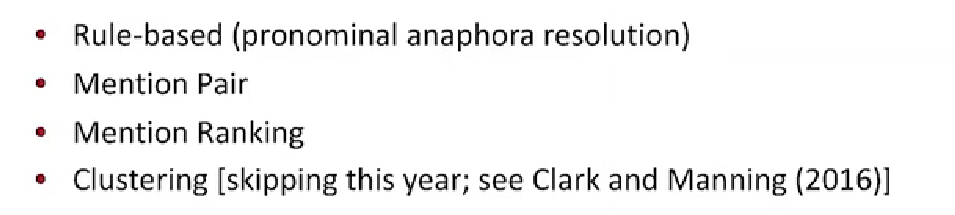


检测指示词，对指示词进行分类，指示同一实体的词汇是同一类。

提及检测（Mention Detection）是指识别文本中所有可能的指代实体，即所有可能成为共指消解对象的词或短语。

③前指（Cataphora）是与照应（Anaphora）相反的一种语言现象。前指指的是在句子或段落中，使用一个词或短语来指代在后面出现的词或短语。照应是指在句子或段落中使用一个词或短语来指代前面提到的词或短语。

④共指模型（Coreference Model）是用于自动化处理自然语言中的共指消解任务的机器学习模型。常见的模型有四种：



⑤提及对（mention pair）是指两个文本片段（mention）之间的关系，这两个文本片段指代同一个实体。共指消解的目标就是识别出这些提及对，从而确定文本中所有指代同一实体的词或短语。

⑥提及排序（Mention Ranking）：给每个可能的先行词（antecedent）打分，来确定最可能的先行词。特别是当代词和先行词之间的距离较远时效果较好。