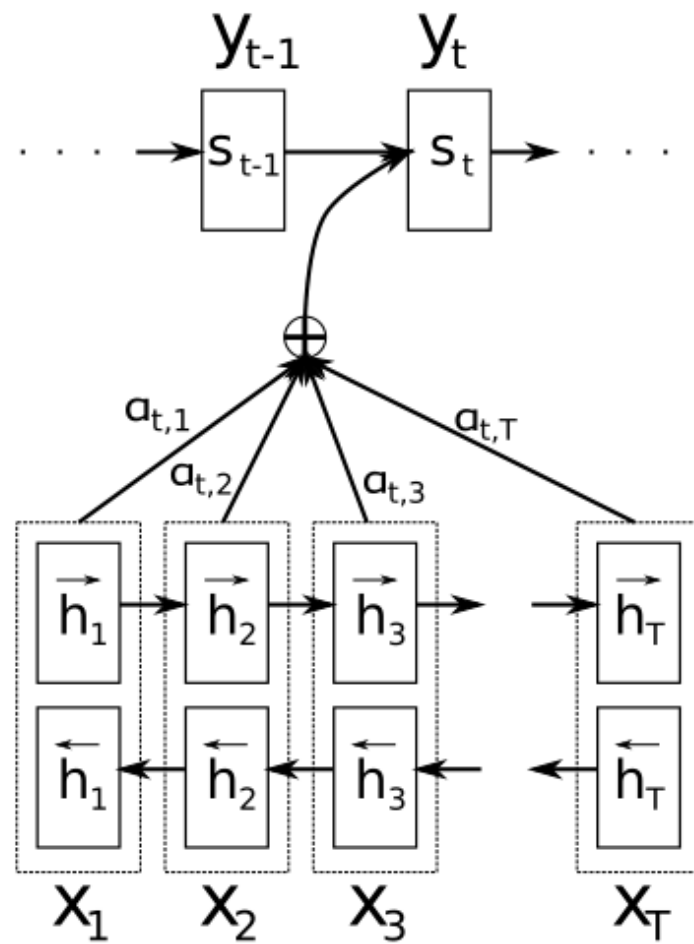


TASK:自然语言答案生成

seq2seq+attention



- generation model:

I: Hello jack, my name is Chandralekha.

R: Nice to meet you.

- copynet:

I: Hello Jack, my name is Chandralekha.

R: Nice to meet you, Chandralekha.

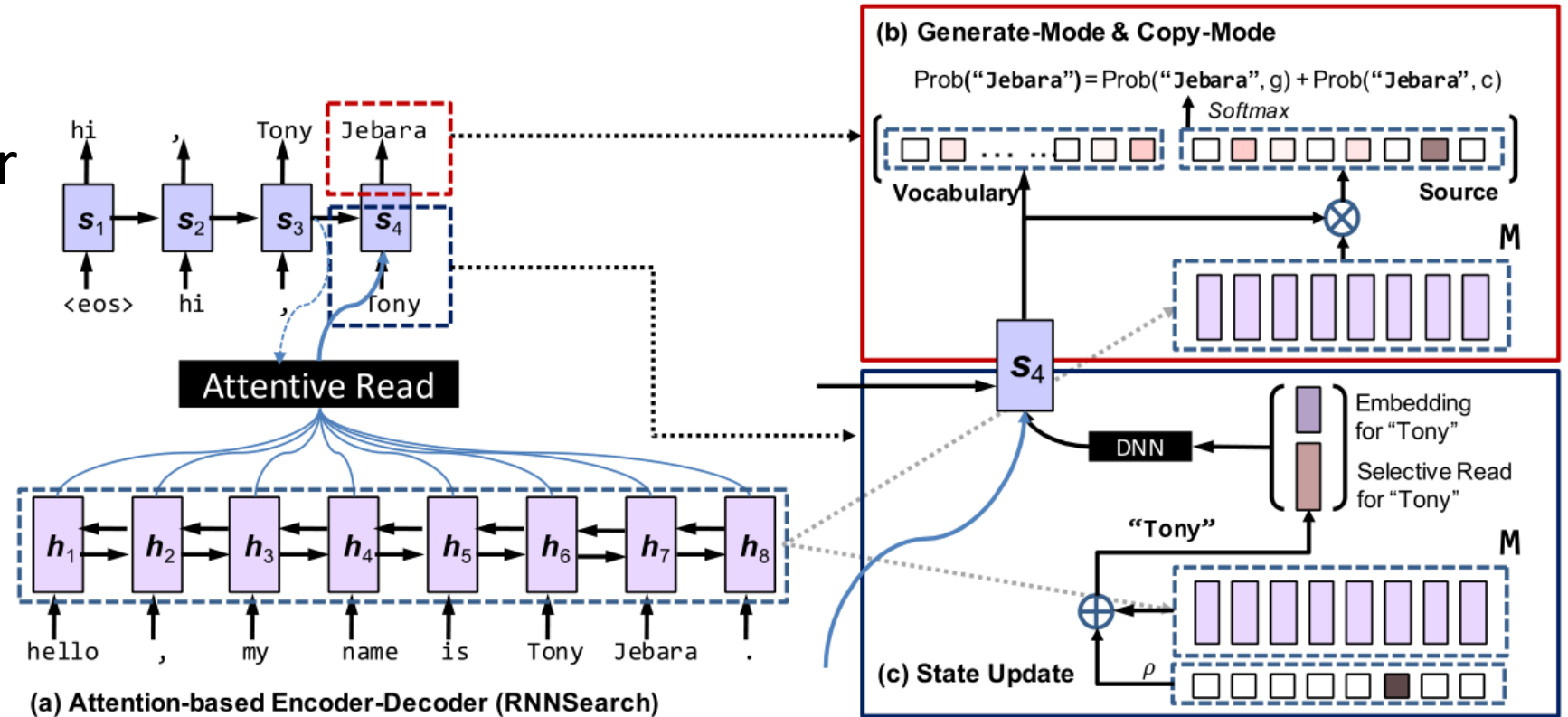
1.Out Of Vocabulary

2.Which to copy

3.Where to patse

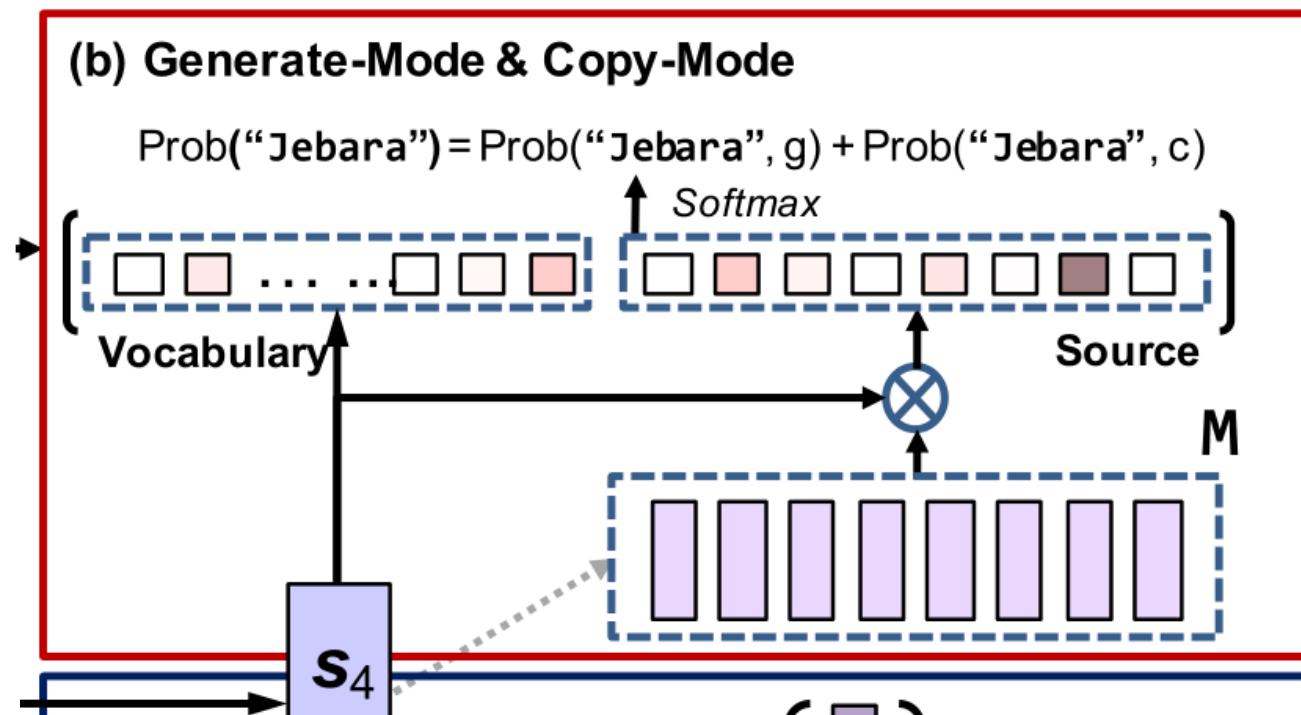
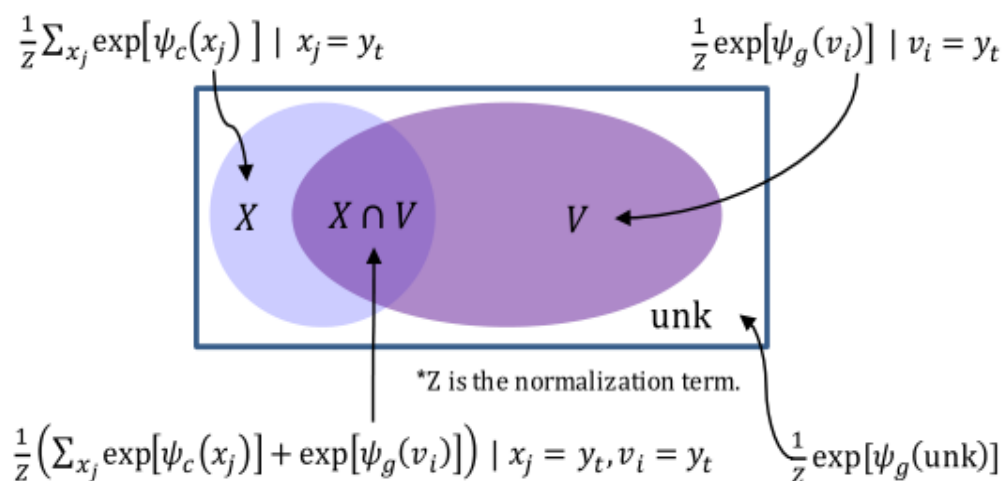
Incorporating Copying Mechanism in Sequence-to-Sequence Learning

- main structure: attention-based encoder-decoder
- differences:
 1. Prediction
 2. State Update



Prediction

$$p(y_t | s_t, y_{t-1}, c_t, M) = p(y_t, g | s_t, y_{t-1}, c_t, M) + p(y_t, c | s_t, y_{t-1}, c_t, M) \quad (4)$$



I: Hello Jack, my name is Chandralekha.

R: Nice to meet you, Chandralekha.

- Generate-Mode

- 对于在输出词表 \mathcal{V} 中的单词 v_i , 有

$$\varphi(y_t = v_i) = \mathbf{W}_o \mathbf{s}_t, \quad v_i \in \mathcal{V} \cup \text{UNK}$$

- 上面 $\mathbf{W}_o \in \mathbb{R}^{(N+1) \times d_s}$, $\mathbf{s}_t \in \mathbb{R}^{d_s}$, d_s 是 \mathbf{s}_t 的维度大小。

- Copy-Mode

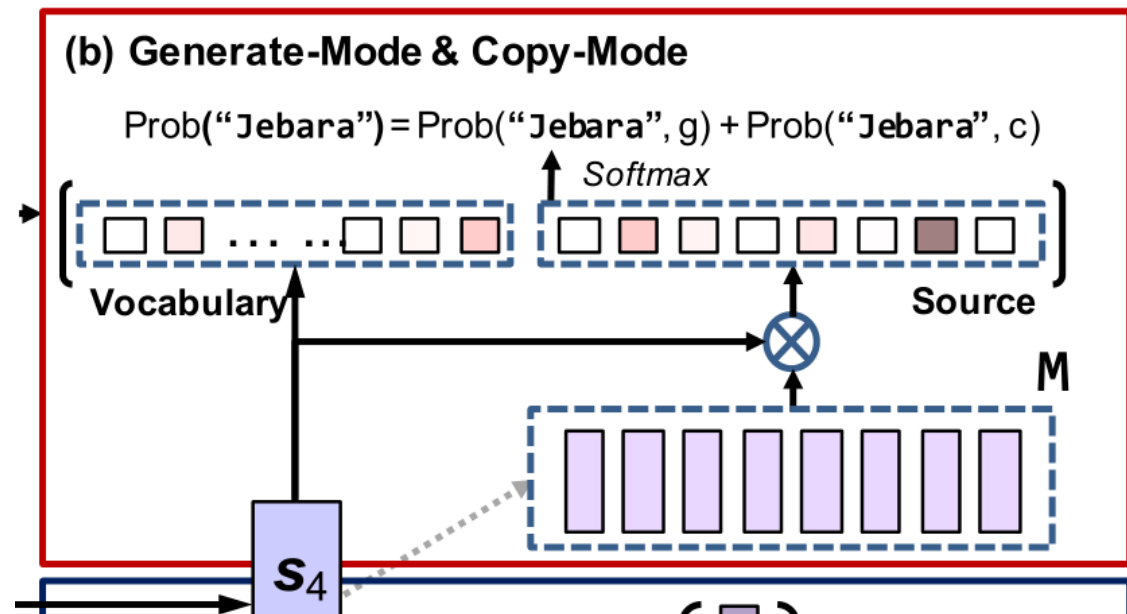
- 对于在输入中出现的单词, 预测的概率有

$$\varphi(y_t = x_j) = \sigma(\mathbf{M} \mathbf{W}_c) \mathbf{s}_t, \quad x_j \in \mathcal{X}$$

- $\mathbf{M} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{T_s}\}$ 是 encoder hidden state, T_s 是输入的单词的个数

- 上面的 $\mathbf{M} \in \mathbb{R}^{T_s \times d_h}$, $\mathbf{W}_c \in \mathbb{R}^{d_h \times d_s}$, $\mathbf{s}_t \in \mathbb{R}^{d_s}$, d_h, d_s 分别是 $\mathbf{h}_t, \mathbf{s}_t$ 的维度

- 为什么要加 σ ? 论文里说用 tanh non-linearity 比线性映射效果好, 但是没有解释为啥



简易版：

- 把预测词的概率由原来的只有生成模式变成混合的2种：
- 1.生成模式
- 2.拷贝模式(这里引入了输入序列的词语，相当于扩大了生成词的范围)

StateUpdate

Original:

$$\mathbf{s}_t = f(y_{t-1}, \mathbf{s}_{t-1}, \mathbf{c})$$

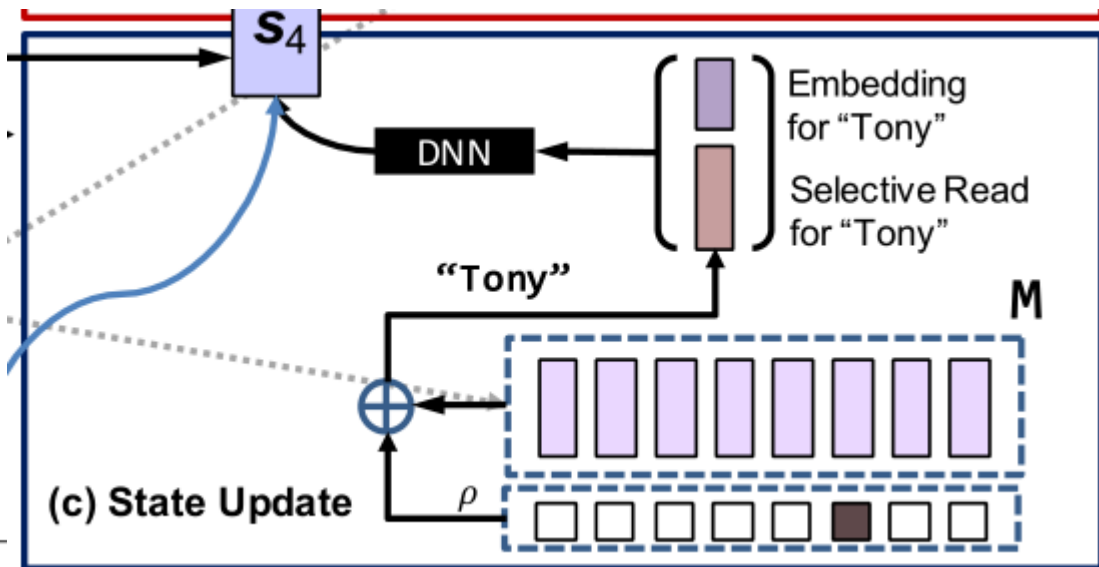
Now:

y_{t-1} will be represented as $[\mathbf{e}(y_{t-1}); \zeta(y_{t-1})]^\top$

$$\zeta(y_{t-1}) = \sum_{\tau=1}^{T_S} \rho_{t\tau} \mathbf{h}_\tau$$

$$\rho_{t\tau} = \begin{cases} \frac{1}{K} p(x_\tau, \mathbf{c} | \mathbf{s}_{t-1}, \mathbf{M}), & x_\tau = y_{t-1} \\ 0 & \text{otherwise} \end{cases}$$


即这个隐层的更新与它前一个词所在的 hidden state 有关。考虑到前一个词可能在多个地方出现，所以才有了 K 这个归一化。



蓝色的是attention得到的 \mathbf{c}_t 进行更新。

在用 y_{t-1} 更新 \mathbf{s}_t 时，CopyNet 不仅仅考虑了词向量，还使用了 \mathbf{M} 矩阵中特定位置的 hidden state。或者说， y_{t-1} 的表示中就包含了这两个部分的信息 $[\mathbf{e}(y_{t-1}); \zeta(y_{t-1})]$ ， $\mathbf{e}(y_{t-1})$ 是词向量， $\zeta(y_{t-1})$ 和 attention 的形式差不多，是 \mathbf{M} 矩阵中与 y_{t-1} 相对应的词的位置的 hidden state 的加权和。

简易版：

- 在隐层更新的时候，由原来的 $\mathbf{s}_t = f(\mathbf{y}_{t-1}, \mathbf{s}_{t-1}, \mathbf{c})$ 变成 $\mathbf{s}_t = f(\mathbf{y}_{t-1}, \mathbf{s}_{t-1}, \mathbf{c})$

$$\zeta(y_{t-1})$$

而 $\zeta(y_{t-1})$ 隐含了它的位置信息。

状态更新：用t-1时刻的预测出的词来更新t时刻的状态，COPYNET 不仅仅词向量，而且使用M矩阵中特定位置的hidden state。

Text Summurization

Input(3): 工厂，大门紧锁，约20名工人散坐在树荫下。“我们就是普通工人，在这里等工资。”其中一人说道。7月4日上午，记者抵达深圳龙华区清湖路上的深圳愿景光电子有限公司。正如传言一般，愿景光电子倒闭了，大股东邢毅不知所踪。

The door of factory is locked. About 20 workers are scattered to sit under the shade. "We are ordinary workers, waiting for our salary" one of them said. In the morning of July 4th, reporters arrived at Yuanjing Photoelectron Corporation located at Qinghu Road, Longhua District, Shenzhen. Just as the rumor, Yuanjing Photoelectron Corporation is closed down and the big shareholder Xing Yi is missing.

Golden: 深圳 亿元级 LED 企业倒闭 烈日下 工人 苦等 老板

Hundred-million CNY worth LED enterprise is closed down and workers wait for the boss under the scorching sun

RNN context: 深圳 "<UNK>": 深圳<UNK><UNK>, <UNK>, <UNK>, <UNK>

CopyNet: 愿景光电子 倒闭 20 名工人 散坐在 树荫下

Yuanjing Photoelectron Corporation is closed down, 20 workers are scattered to sit under the shade

Input(5): 中国 反垄断 调查 风暴 继续 席卷 汽车行业，继德国 车企 奥迪 和美国 车企 克莱斯勒 “沦陷”之后，又有 12 家日本汽车企业 卷入漩涡。记者从业内人士获悉，丰田旗下的 雷克萨斯 近期 曾被 发改委 约谈。

Chinese antitrust investigation continues to sweep the automotive industry. After Germany Audi car and the US Chrysler "fell", there are 12 Japanese car companies involved in the whirlpool. Reporters learned from the insiders that Toyota's Lexus has been asked to report to the Development and Reform Commission recently.

Golden: 发改委 公布 汽车 反垄断 进程：丰田 雷克萨斯 近期 被 约谈

the investigation by Development and Reform Commission: Toyota's Lexus has been asked to report

RNN context: 丰田雷克萨斯遭发改委约谈：曾被约谈丰田旗下的雷克萨斯遭发改委约谈 负责人被约谈

CopyNet: 中国 反垄断 继续 席卷 汽车行业 12 家日本汽车企业 被 发改委 约谈

Chinese antitrust investigation continues to sweep the automotive industry. 12 Japanese car companies are asked to report to the Development and Reform Commission

Single-turn Dialogue

Input(1): 我出生在上海。
/I was born in Shanghai

Golden: 真巧, 我也来自上海, 老乡啊! /What a coincidence, I also come from Shanghai, we are fellow townsman!

✗ RNNSearch: 真巧, 我也来自海南, 老乡啊! /What a coincidence, I also come from Hainan, we are fellow townsman!

○ CopyNet: 真巧, 我也来自上海, 老乡啊! /What a coincidence, I also come from Shanghai, we are fellow townsman!

Input(2): 今天去爬黄山吗?
/Shall we climbing Huangshan?

Golden: 去爬泰山吧, 黄山没意思! /Climb Taishan. Huangshan is boring!

✗ RNNSearch: 去爬长白山吧, 衡山没意思! /Climb Changbaishan, Hengshan is boring!

○ CopyNet: 去爬嵩山吧, 黄山没意思! /Climb Songshan, Huangshan is boring!

Input(3): 你喜欢喝雪碧还是喜欢可乐?
/Would you like to drink Sprite or Coke?

Golden: 都不喜欢, 谢谢! /Not at all, thanks!

✗ RNNSearch: 我喜欢喝蜂蜜! /I like to drink honey!

○ CopyNet: 我喜欢喝可乐! /I like to drink coke!

Input(4): 我大青岛四季宜人
/Our great Qingdao has a good weather!

Golden: 壮我大青岛~~~ /Our Qingdao is great~~~

✗ RNNSearch: 壮我大哈尔滨~~~ /Our Harbin is great~~~

○ CopyNet: 壮我大青岛~~~ /Our Qingdao is great~~~

Input(5): 女儿会游蝶泳
/Daughter can swim butterfly?

Golden: 我连游泳都不会T^T /I cannot swim at all T^T

○ RNNSearch: 我连游泳都不会T^T /I cannot swim at all T^T

○ CopyNet: 我连游泳都不会T^T /I cannot swim at all T^T

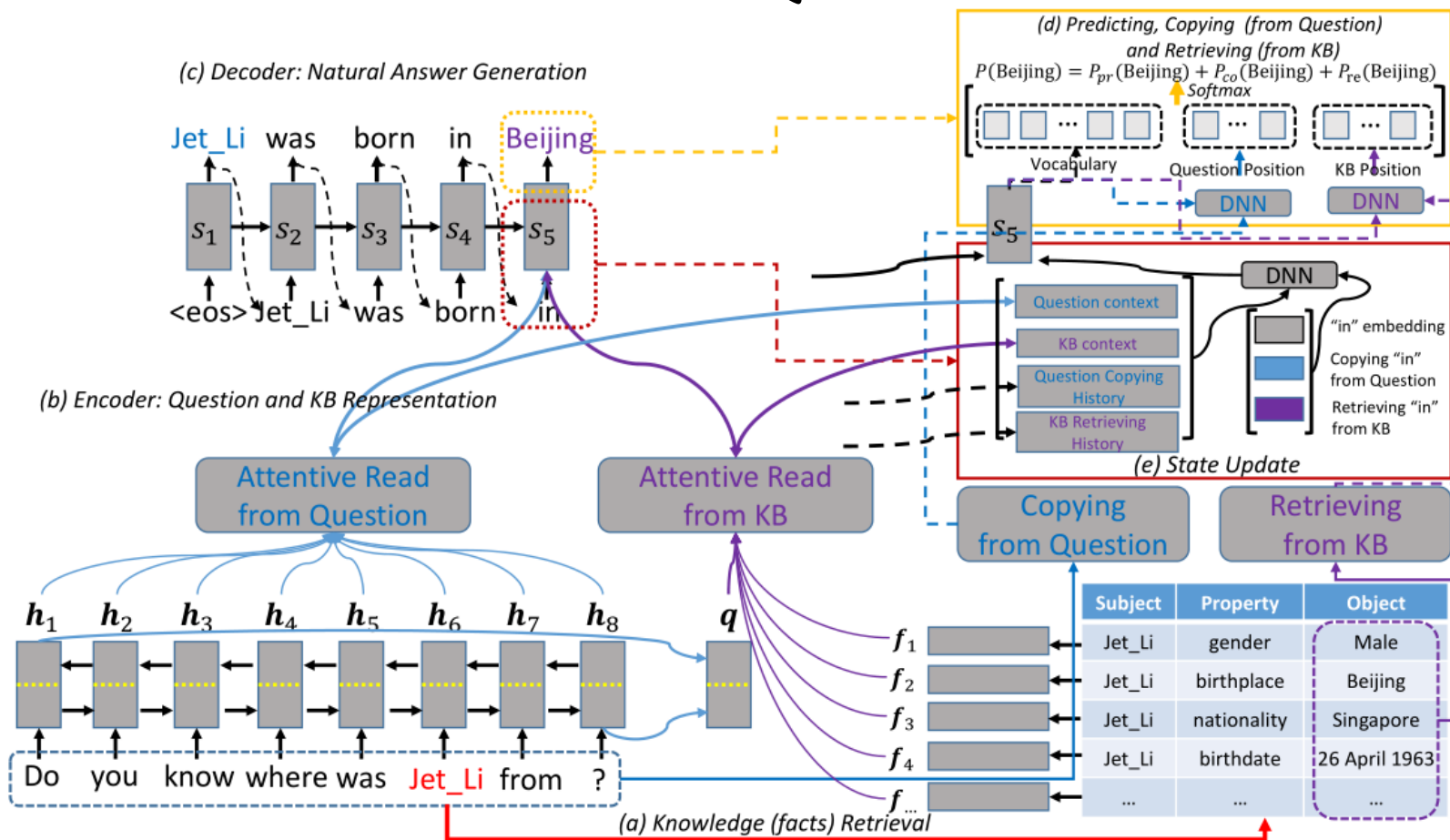
Input(6): 苏有朋是谁?
/Who is Su You Peng?

Golden: 苏有朋是一个男明星。 /Su You Peng is a male star.

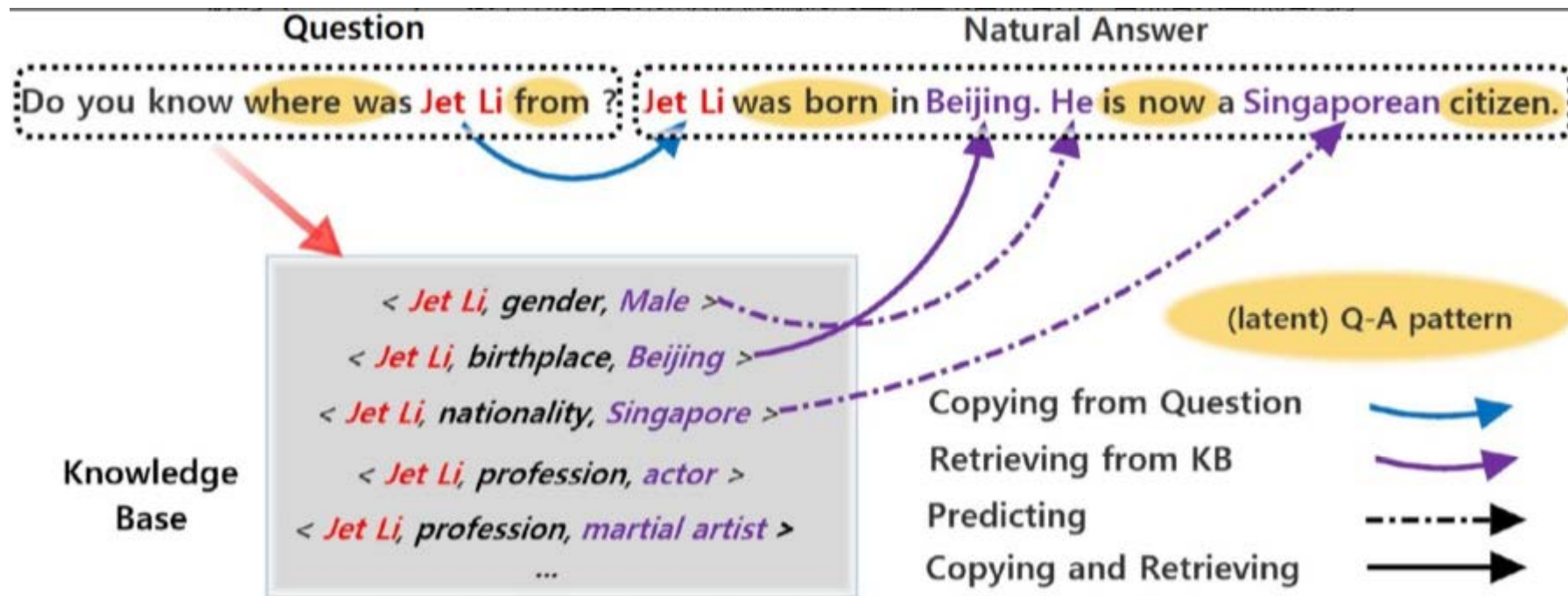
✗ RNNSearch: 吴亦凡是一个男明星。 /Wu Yifan is a male star.

✗ CopyNet: 苏有是一个男明星。 /Su You is a male star

COREQA



COREQA 利用 拷贝(copy)、检索(retrieval)和预测(prediction) 从不同来源获取不同类型的词汇，产生复杂问句的自然答案。



- 单词预测有三种模式，predict-mode, copy-mode 和 retrieve-mode，predict-mode 和普通 seq2seq 原理相同，生成词汇表中的单词，copy-mode 从问句中复制单词，retrieve-mode 从知识库中选取单词。过程和 CopyNet 差不多，也有两种读取方式，一种是读取语义，一种是读取位置。

- 何世柱博士表示：“**COREQA** 模型目前还是过于依赖学习数据。从实验结果可以看出，在模拟的人工数据上几乎可以有完美的表现，但是在真实的数据上还是差强人意。”

THANKS