法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

第四节课：seq2seq模型与generative 聊天机器人

一个原理非常简单，

然而实践起来非常多细节和麻烦的模型

本节内容

- Seq2seq产生模型
  - 在聊天机器人领域的应用
  - Seq2seq模型的evaluation Metric
  - 代码演示
- Beam search: Seq2seq模型的解码过程
  - Beam search 的原理
  - Beam search 的应用
  - 代码演示

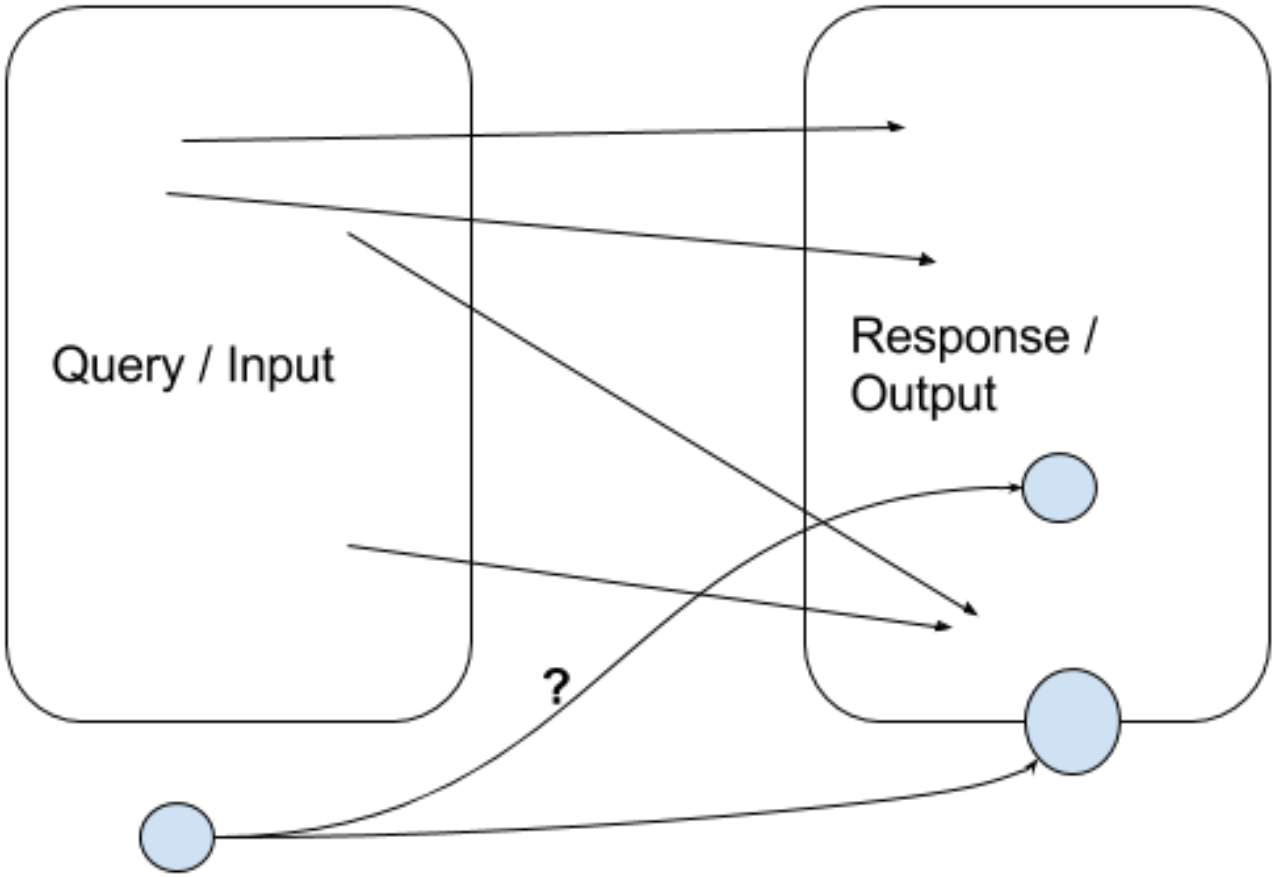
参考文献

- Seq2seq产生模型
  - 模型结构：
    - Sequence to sequence learning with Neural Networks (Sutskever 2014, cited > 1700)
    - A Neural Conversational Model (2015)
  - 模型evaluation metric:
    - How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation (2017)
    - Training End-to-End Dialogue Systems with the Ubuntu Dialogue Corpus (2017)
    - Towards an automatic turing test: learning to evaluate dialogue responses (2017)

参考文献

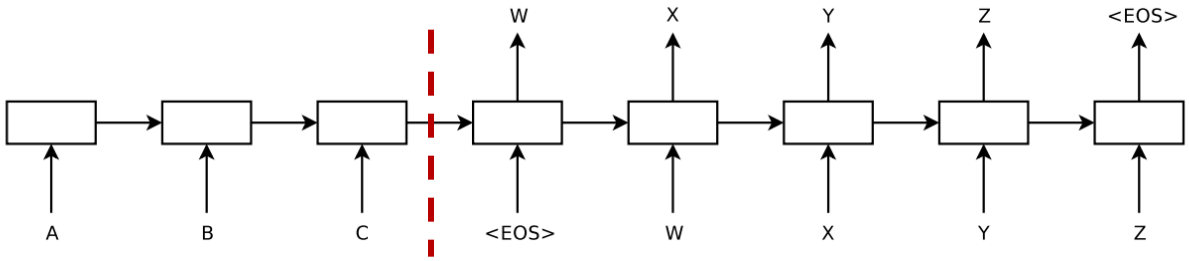
- Beam search
  - Sequence to sequence learning with Neural Networks (Sutskever 2014, cited > 1700)
  - A Neural Conversational Model (2015)
  - Sequence-to-Sequence Learning as Beam-Search Optimization (2016)
  - Training End-to-End Dialogue Systems with the Ubuntu Dialogue Corpus (2017)
  - https://github.com/tensorflow/tensorflow/issues/654#issuecomment-169009989

开放领域和封闭领域



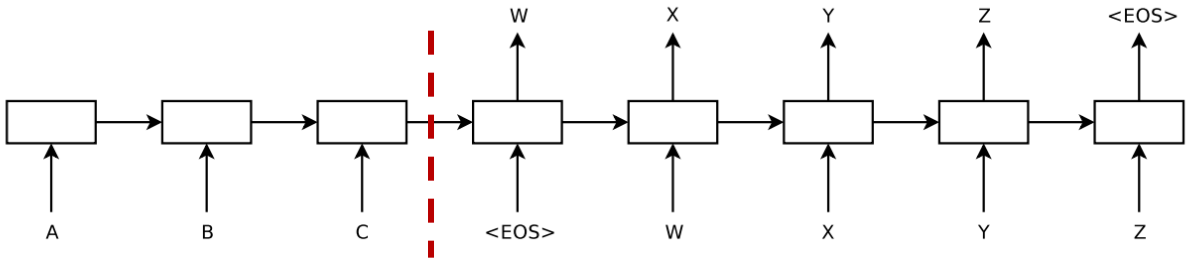
Sequence to Sequence Learning with NN

- 提出简单而有效的seq2seq模型并应用于机器翻译问题
- 实验研究LSTM模型的性质，并给出一些可能提高模型表现的经验性(empirical)方法
  - 实验表明LSTM可以处理非常长的句子(did not suffer on very long setnences)



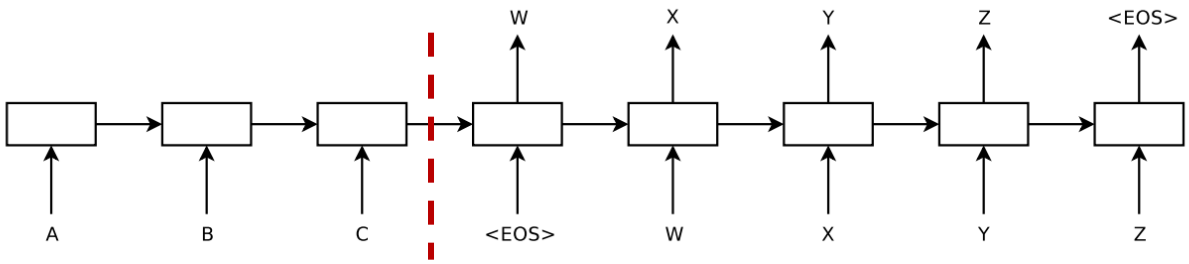
Sequence to Sequence Learning with NN

- seq2seq模型结构
  - 使用LSTM 将不定长度的句子映射到固定维度的表达向量(vector representation)
  - 使用beam search从句子的表达向量出发，使用LSTM产生另一个句子，使用另外一种语言表示同样的意义



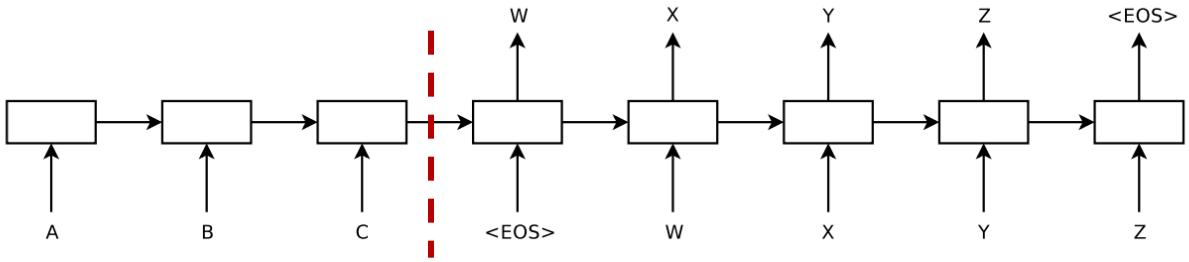
Sequence to Sequence Learning with NN

- seq2seq模型结构
  - 相对于RNNLM(input & output perfectly aligned)，seq2seq模型的一个挑战是，input和output句子的长度不需相同，而且两者长度的关系是未知的。



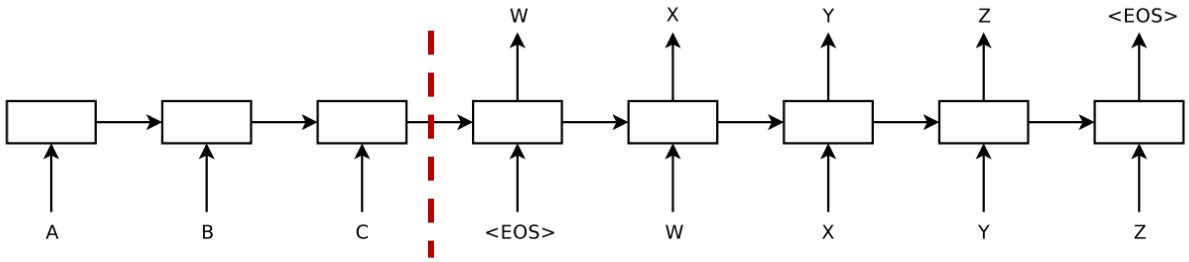
Sequence to Sequence Learning with NN

- seq2seq模型结构
  - 将decoding过程描述成一个LSTM-LM
  - 训练过程中最大化$(x_1 \cdots x_T) \rightarrow (y_1 \cdots y_{T'})$的条件概率
    - $P(y_1 \cdots y_{T'} | x_1 \cdots x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \cdots y_{t-1})$,
    - $y_0 \equiv < EOS >$



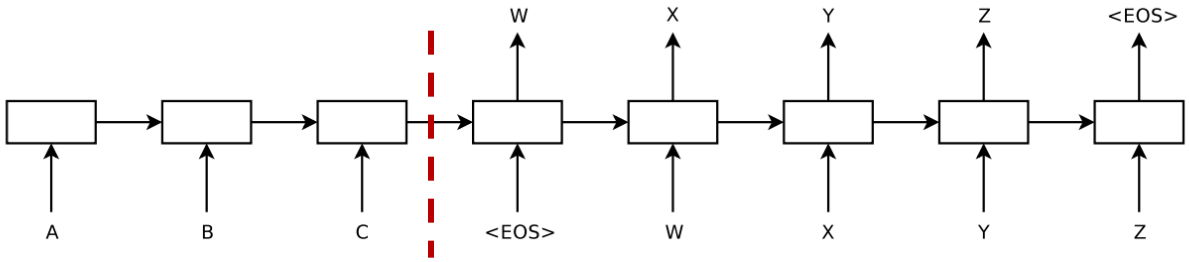
Sequence to Sequence Learning with NN

- seq2seq模型学习与测试
  - 训练：最大化$\frac{1}{|D|}\sum_{(T,s)\in D}\log p(T|S)$
  - 测试(inference)：寻找$\hat{T} = \text{argmax}_T P(T|s)$
    - Intractable full search space
    - 使用beam search启发式地搜索本地最优解
    - 当遇到<EOS>的时候，结束解码



Sequence to Sequence Learning with NN

- 一些有用的经验（至少在MT，机器翻译，任务上）
  - 在编码过程中，将输入的句子反向有效地提高任务上的表现
  - 编码和解码阶段使用不同的LSTM模型
    - 增加模型参数而不会增加时间复杂度
    - 至少在MT任务上非常容易理解
  - Deep LSTM 明显好过Shallow LSTM（本文使用4层）



A Neural Conversational Model

- 将前述的MT问题中引入的的 seq2seq (simplicity, generality)模型应用于对话问题上
- 给出(query, response)以后，seq2seq结构方便End-to-End训练，不需要人工标注任务和中间过程
- 通过优化一个和“对话”这个目标并不完全一致的目标函数，可以获得不错的对话效果(the **objective function** being optimized does not capture the **actual objective achieved through human communication**, which is typically longer term and based on exchange of information rather than next step prediction.)

A Neural Conversational Model

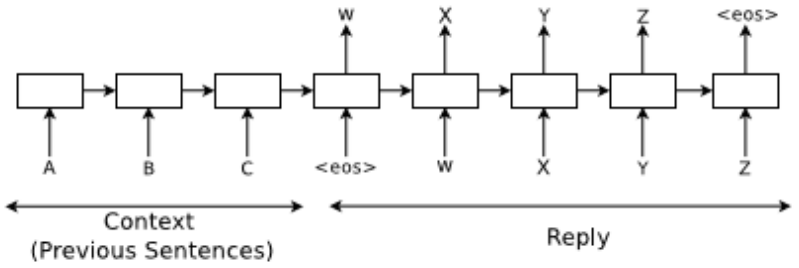


Figure 1. Using the *seq2seq* framework for modeling conversations.

- MT
  - Sentence-English
  - Sentence-French
- Conversation
  - Query sentence
  - Response sentence

A Neural Conversational Model

测试阶段，对产生的机器回复的定性观察：

Human: *what do you think about tesla ?*
Machine: *he 's a good conductor .*
Human: *what do you think about bill gates ?*
Machine: *he 's a good man .*
Human: *What do you think about messi ?*
Machine: *he 's a great player .*
Human: *what do you think about cleopatra ?*
Machine: *oh , she 's very regal .*
Human: *what do you think about england during the reign of elizabeth ?*
Machine: *it was a great place .*

Seq2seq模型可以在一定程度上记住知识，理解语境，进行简单的推理

Human: *what is your job ?*
Machine: *i 'm a lawyer .*
Human: *what do you do ?*
Machine: *i 'm a doctor .*

Seq2seq模型上不能保留记忆和性格，对相同语意的不同表达会返回不同的答复

Evaluation Metric

- 训练阶段
  - 关注真实response出现的概率和最大化这个概率的模型参数
  - 通过计算真实句子出现的概率，训练阶段的目标函数容易定义（例如，LSTM-LM decoding产生的句子和训练样本中的真实句子的cross entropy）
- 测试阶段
  - 然而，考虑到回复的灵活性（一个query可以有多种得体的回复）和产生模型的自由度（不需要在模板库里面选择回复，可以是全新的句子）
  - 使用合适的Metric衡量产生的句子是非常困难的问题

Evaluation Metric

- 使用generative metric 计算生成或者检索的回复和真实回复之间的符合程度。其设计目标是：metric的判断和人为判断尽量相似
- BLEU (**Bi**L**L**ingual **E**valuation **U**nderstudy)
  - 最初用于衡量机器翻译效果： "the closer a machine translation is to a professional human translation, the better it is“
  - 适用于衡量数据集量级的表现，不适用于句子级别的表现
    “BLEU is designed to approximate human judgement at a corpus level, and performs badly if used to evaluate the quality of individual sentences.” ——wikipedia

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

$$P = \frac{m}{w_t} = \frac{7}{7} = 1$$

$$P = \frac{2}{7}$$

How **NOT** To Evaluate Your Dialogue System

Context of Conversation
Speaker A: Hey John, what do you want to do tonight?
Speaker B: Why don't we go see a movie?
Ground-Truth Response
Nah, I hate that stuff, let's do something active.
Model Response
Oh sure! Heard the film about Turing is out!

Table 1: Example showing the intrinsic diversity of valid responses in a dialogue. The (reasonable) model response would receive a BLEU score of 0.

1. 相对于机器翻译，对话中回复的选择空间大很多；看起来完全无关的两句话都可以是合适的回复
2. 两个回复如果不看context的话，无论是从词频还是语义来看都是不相关不想似的句子

Training End-to-End Dialogue System with UDC

- Human judgement
- Word-overlap metrics
  - BLEU, METEOR, ROUGE
- word-embedding metrics / Vector-base metrics
  - Embedding average score
  - Greedy matching score
  - Vector extrema score

Training End-to-End Dialogue System with UDC

- Vector-base metrics: 侧重比较生成的句子和真实样本的语意相似度
  - Embedding average score： 将句中单词的词向量的平均作为句子的特征，计算生成的句子和真实句子的特征的cosine similarity
  - Greedy matching score：寻找生成的句子和真实句子中最match的一对单词，单词的similarity近似句子的similarity
  - Vector extrema score：提取句中单词的词向量的dimension-wise最大（小）值作为句子向量对应维度的数值，然后计算cosine similarity
- 使用Glove词向量，词向量包含语义信息，认为词向量的简单组合也包含语义信息

How **NOT** To Evaluate Your Dialogue System

- 在和人工判断关联这一点上，上述所有的metric都是辣鸡 “We find that all metrics show either weak or no correlation with human judgements, despite the fact that word overlap metrics have been used extensively in the literature for evaluating dialogue response models”
- 在闲聊性质的数据集上，上述metric和人工判断有一定微弱的关联 (only a small positive correlation on chitchat oriented Twitter dataset)
- 在技术类的数据集上，上述metric和人工判断完全没有关联(no correlation at all on the technical UDC)
- 当局限于一个特别具体的领域时，BLEU会有不错的表现

Learning to Evaluate Dialogue Responses

- 使用机器学习学习一个好的metric

$$score(c, r, \hat{r}) = (\mathbf{c}^T M \hat{\mathbf{r}} + \mathbf{r}^T N \hat{\mathbf{r}} - \alpha)/\beta$$

$$\mathcal{L} = \sum_{i=1:K} [score(c_i, r_i, \hat{r}_i) - human\_score_i]^2 + \gamma ||\theta||_1$$

代码演示

解码 (decoding) 阶段生成回复 (response)

BEAM SEARCH

Beam search 流程

$maxP(T|S)$ 全局最优不可能（intractable）

Beam search 流程

$maxP(T|S)$ greedy解法累积误差

Beam search 流程

Beam search 流程

Beam search 流程

Beam search 流程

Beam search 流程

Beam search 流程

Beam search 流程

Beam search 应用

完全搜索树

Beam search 应用

模板库的前缀树

Beam search 应用

模板库的前缀树

Beam search 代码演示

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop