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

第三节课：封闭领域的聊天机器人模型

- 基于检索的聊天模型
  - 用于封闭领域内回复检索的机器学习方法
  - 句子特征表达， 模型的损失函数与衡量指标
- 聊天模型**最基本功能**的代码演示
- 以email smart replay为例看模型的实际应用

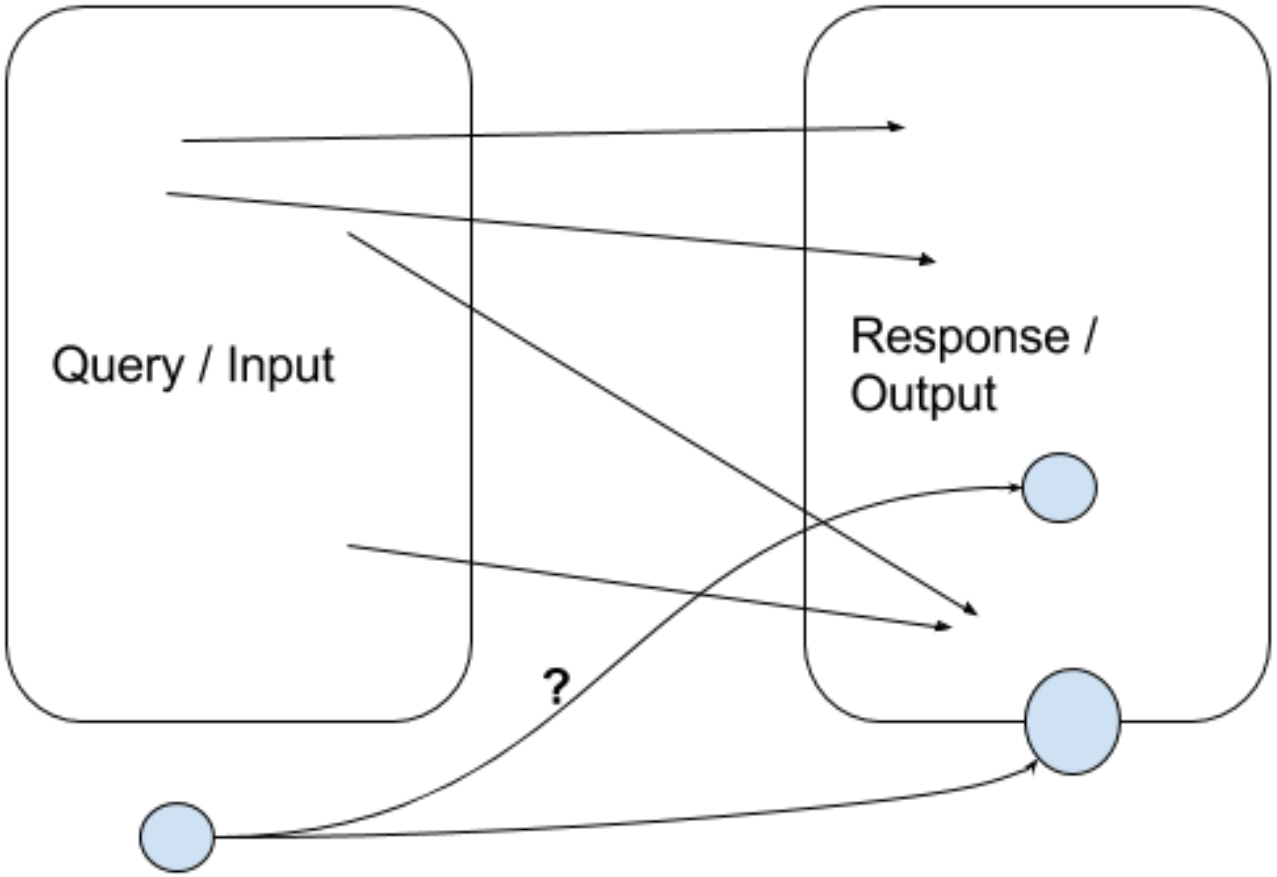
参考文献

- 基于检索的聊天模型原理
  - 聊天数据的提取和处理
    - The Unbuntu Dialogue Corpus: A large dataset for research in unstructured multi-turn dialogue systesm (2015)
    - Github DeepQA and Ubuntu-ranking-data-creator
  - Training end-to-end dialogue system with the UDC (McGill, 2017)
  - Blog: Deep Learning for Chatbots
- 基于检索的聊天模型的代码演示
  - Github DeepQA and chatbot-retrieval
- 基于检索的聊天模型的实际应用
  - Smart reply, automated response suggestion in email. (2016)

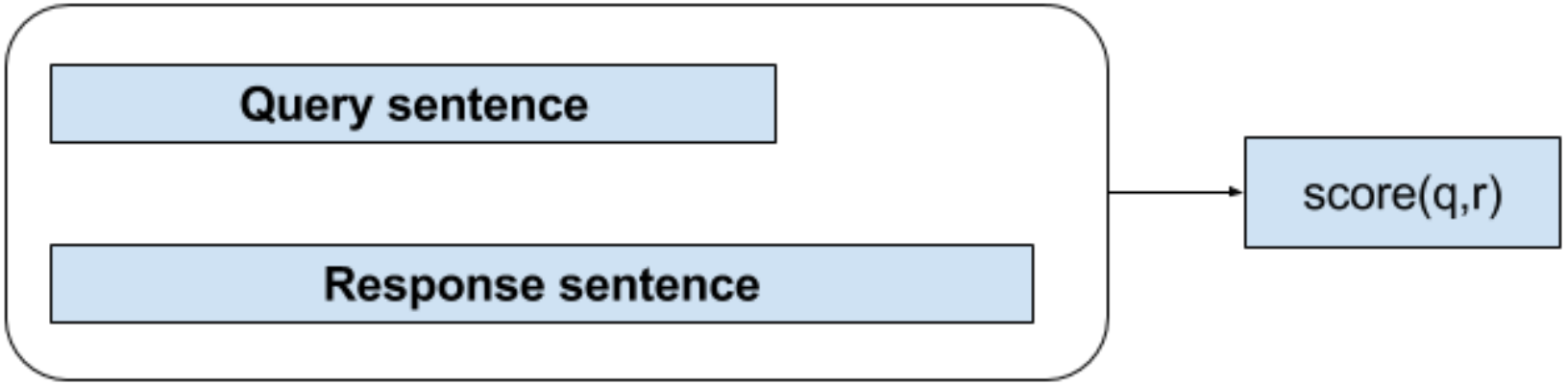
基于检索的聊天模型

模型原理

开放领域和封闭领域



封闭领域聊天模型



最基本的模型结构：

给定问题query,所有的候选response句子里面，真实/正确的回复有最高的分数

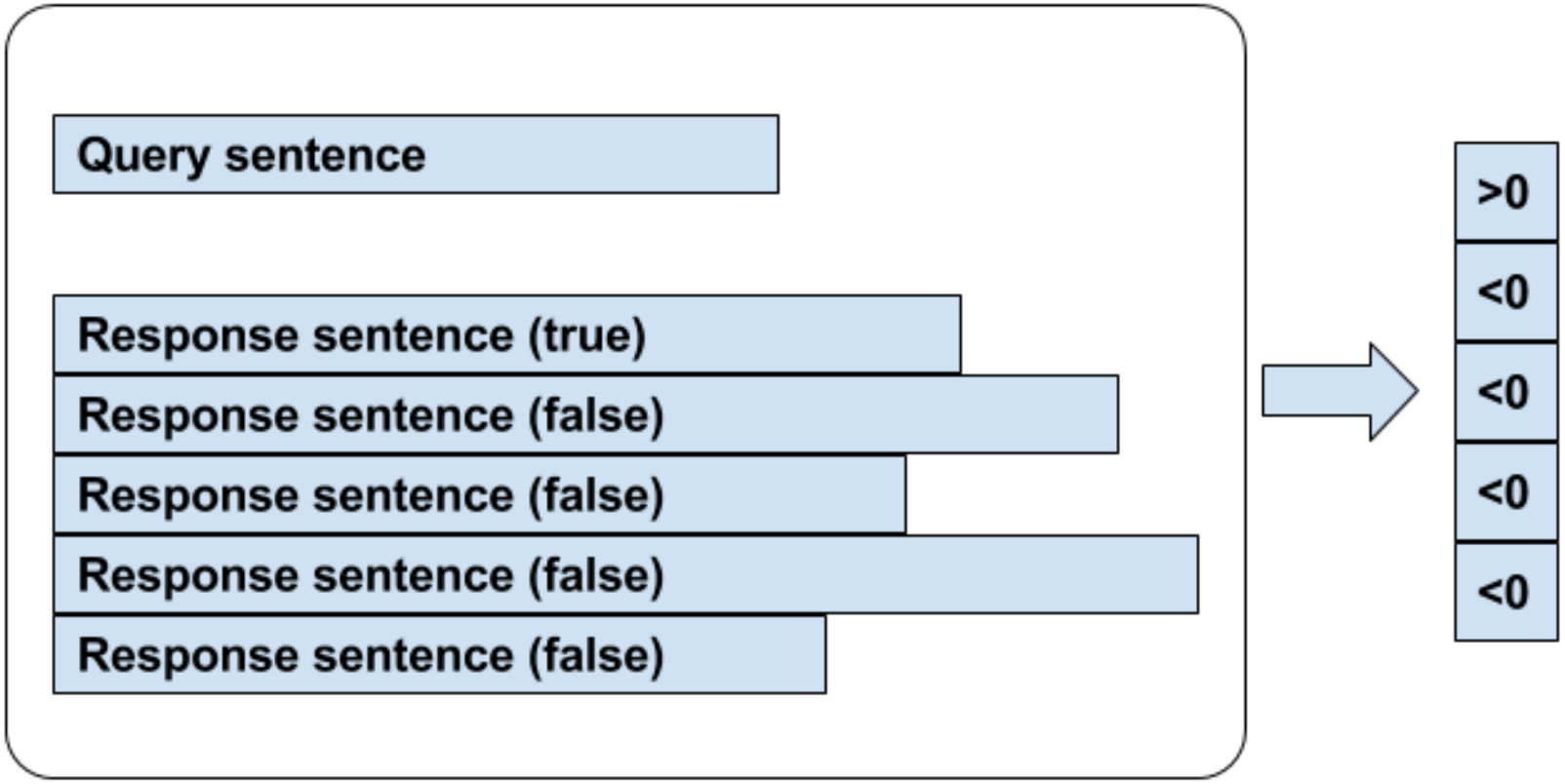
Response 分类问题

Context	Response	Flag
well, can I move the drives? __eot__ ah not like that	I guess I could just get an enclosure and copy via USB	1
well, can I move the drives? __eot__ ah not like that	you can use “ps ax” and “kill (PID #)”	0

Response分类问题

- 模型的预测值
  - Score(Query， Response)
  - 衡量问题和回复的合适程度
  - Score越高，认为Response越可能是一个合适的回复
- 模型的学习目标，binary分类器
  - $\sigma\left(score(Query, Response_{\text{true}})\right) \rightarrow 1$
  - $\sigma\left(score(Query, Response_{\text{false}})\right) \rightarrow 0$

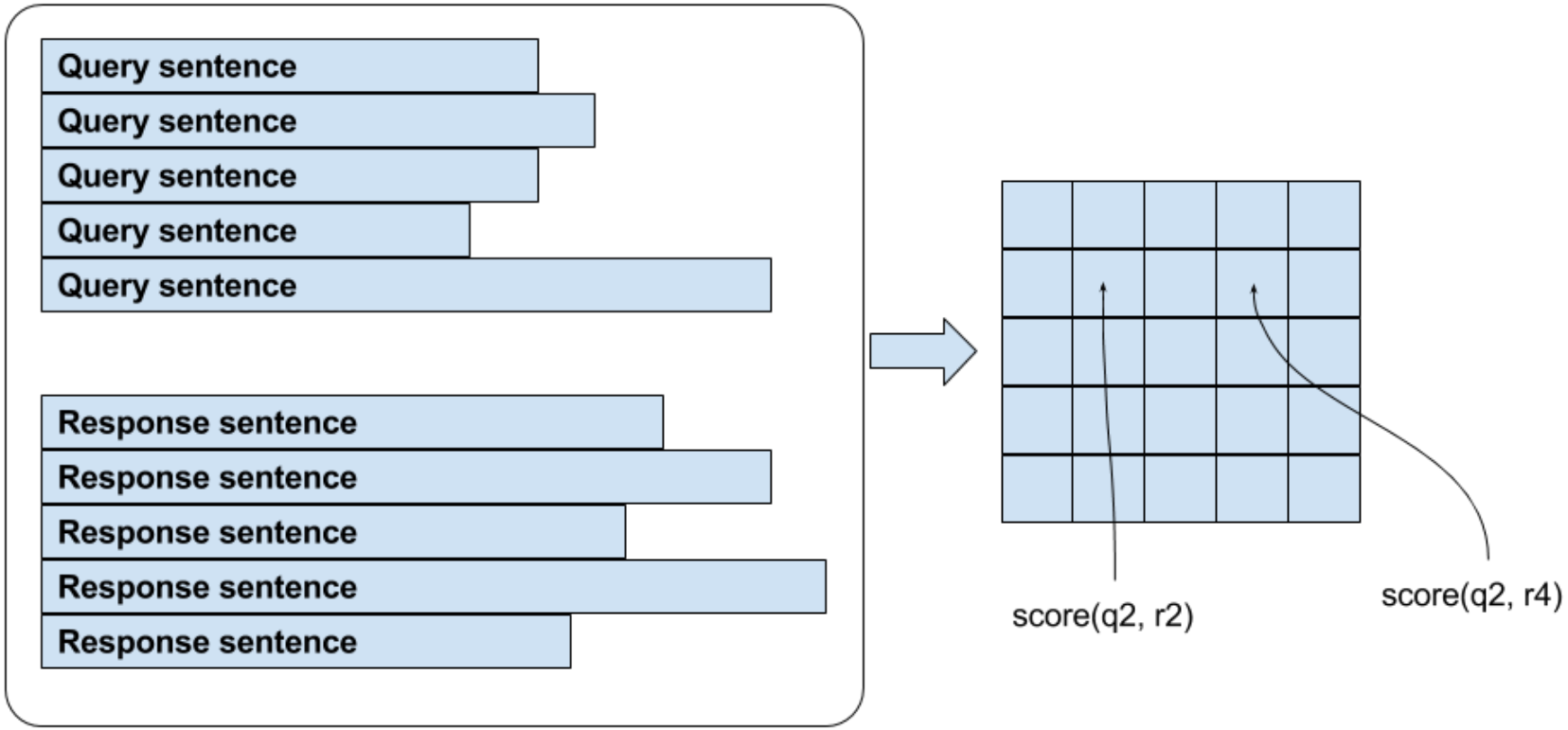
Response分类问题



Response分类问题

	Retrieval Metrics			
Method	1 in 2 R@1	1 in 10 R@1	1 in 10 R@2	1 in 10 R@5
TF-IDF	74.9%	48.8%	58.7%	76.3%
Dual Encoder w/RNN units	77.7%	37.9%	56.1%	83.6%
Dual Encoder w/LSTM units	**86.9%**	**55.2%**	**72.1%**	**92.4%**

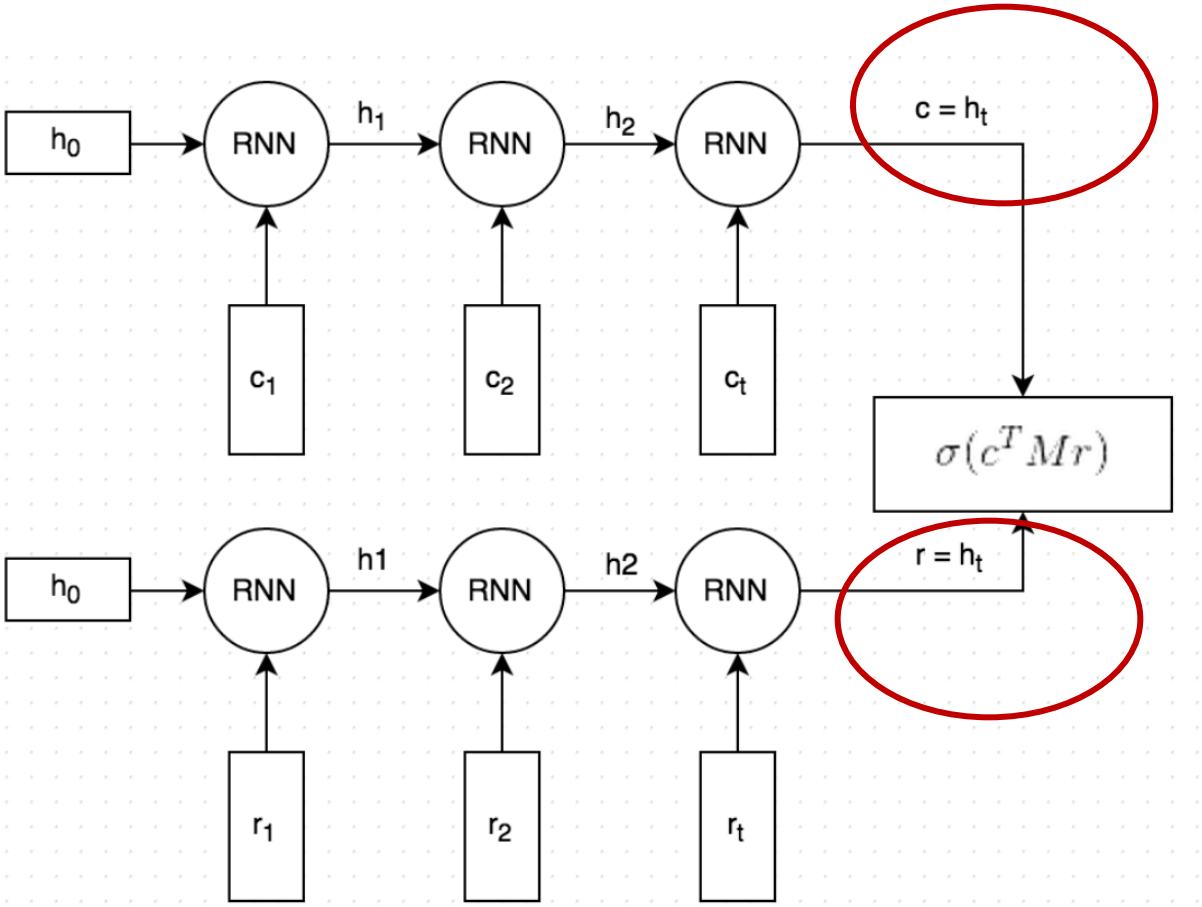
Response排序问题



句子的特征(feature)

- BOW， TFIDF
- Sentence embedding

Dual-LSTM模型


h0
RNN
h1
RNN
h2
RNN
c = ht
c1
c2
ct
σ(cTMr)
h0
RNN
h1
RNN
h2
RNN
r = ht
r1
r2
rt

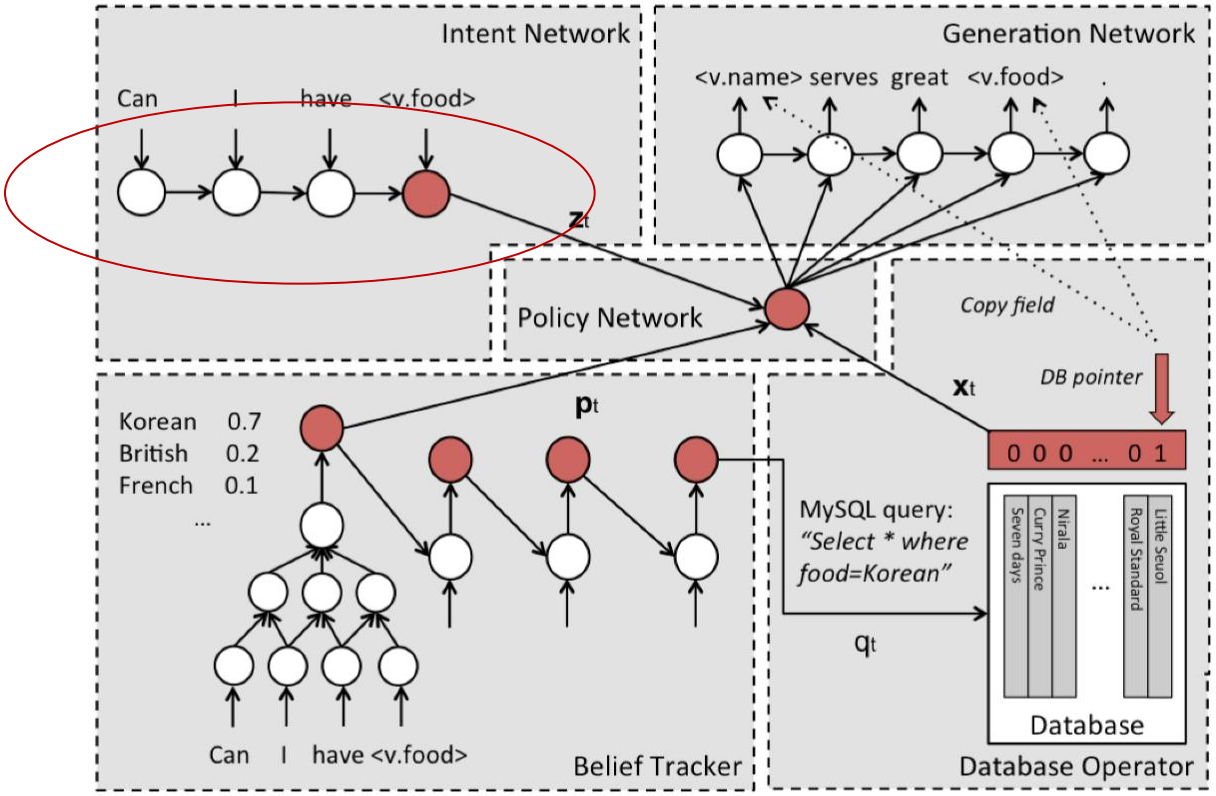
基于检索的聊天模型

排序问题代码演示

基于检索的聊天模型

实际应用

Machine learning模型和chatbot产品



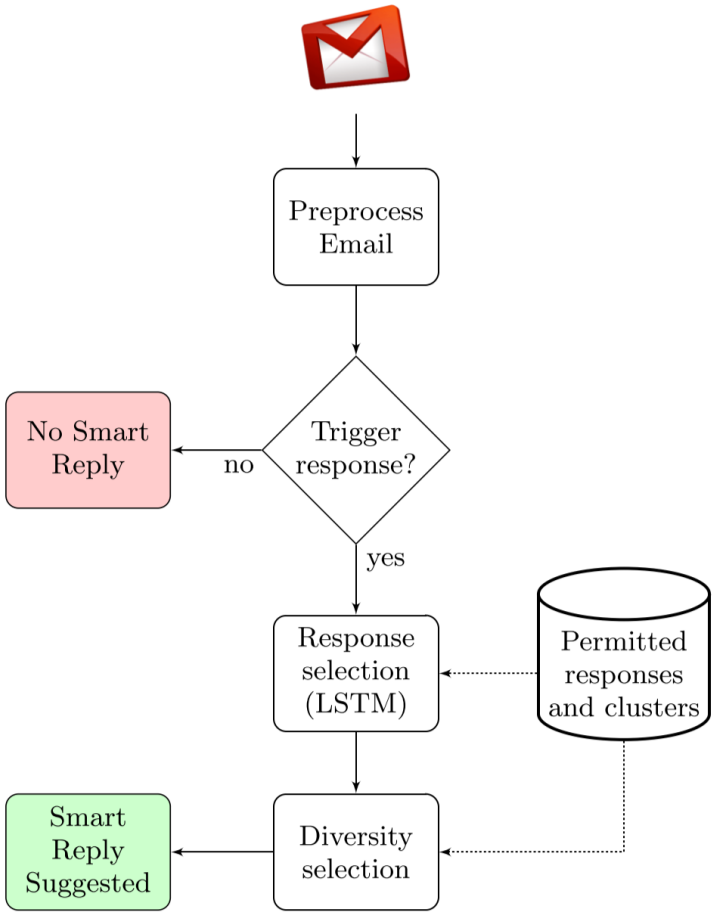
即便是小的封闭领域，达到好的效果需要很多工程上的努力

预告：最后一节课的餐馆推荐机器人

Smart Email Reply

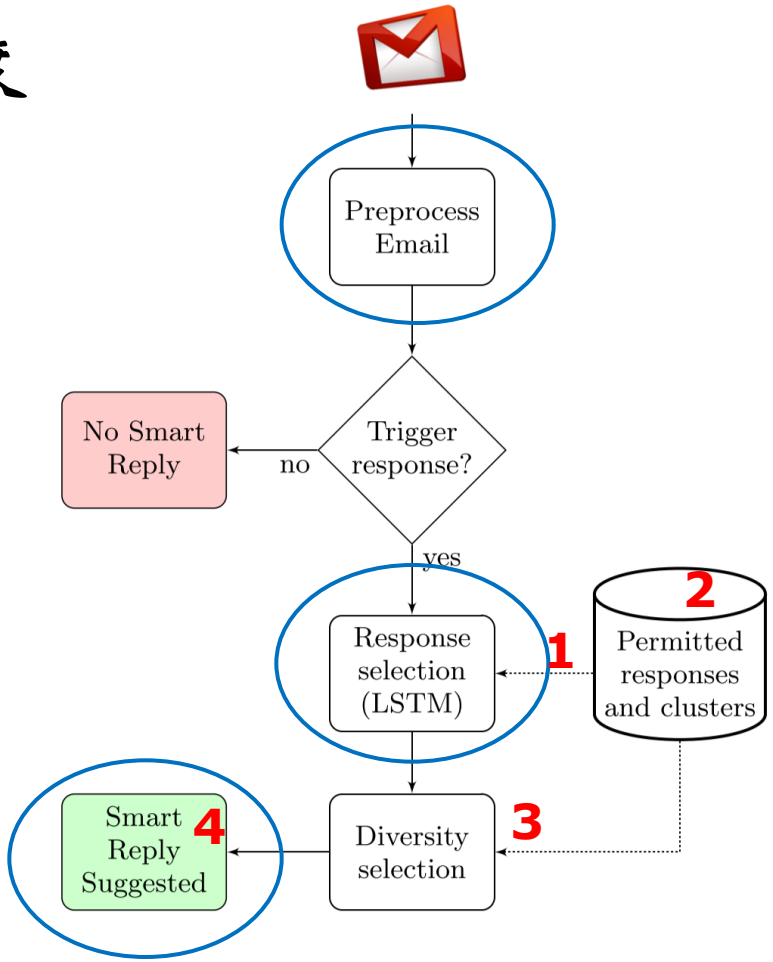
Smart reply, automated response suggestion in email. (2016)

- 一篇”Applied AI” 论文
- 关于machine learning模型在实际产品中的使用
- 具有“回复邮件”功能的聊天机器人案例



Smart Email Reply

- 从Machine learning角度
  - 输入：来信
  - 特征：LSTM
  - 回复：高分选项
- 从应用的角度
  1. 如何提高回复的速度？
  2. 如何构建模板库？
  3. 如何避免generic回复？
  4. 如何衡量模型的表现？



回复效率问题：

- 如何从大规模模板库中快速寻找最优选项？
  - 给可能有million量级的模板库中的每个选项打分过于低效
  - 将模板库存储在一个TRIE树结构中，通过LSTM模型打分在TRIE树中找到最合适的一些回复

生成Smart reply模板库I

- 搜集邮件数据后的预处理
  - 去掉非英语的样本
  - Tokenization（分词） 话题和正文
  - 将内容分割成句子为单位
  - 使用特殊符号替换不常用的单词(e.g. 人名，url，邮件地址)
  - 去掉引用和转发的邮件部分
  - 去掉问候和致敬部分

生成Smart reply模板库II

- 在预处理后的数据中，选择短的，最常出现的，匿名的，短回复。O(million)量级
- 语法分析处理语言中的灵活性问题
  - 将类似的句子, e.g. “Thanks for your kind update”, “Thank you for updating!”, and “Thanks for the status update” 转换为**canonical形式**，即，“Thanks for the update.”
- 将上述处理过后的回复数据集做语义聚类，每个cluster对应一个意图（intent）.
  - 初始化～100个类别，每个类别～3个人工选择的样本

生成Smart reply模板库III

- 使用graph上的半监督学习方法对回复数据集做语义聚类
  - **初始化:** 标记～100个类别(cluster)，每个类别～3个人工选择的样本
  - **Graph:** 使用（来信，回复），（回复1，回复2，特征）信息对回复数据集里面的样本建立联系
  - **Iteration**：
    - Inference: 推测未标记样本的cluster类别
    - Update: sample～100个cluster类别不清的样本，人为标记
  - **验证**：提取每个cluster的top-k个回复样本，人工验证

生成Smart reply模板库III

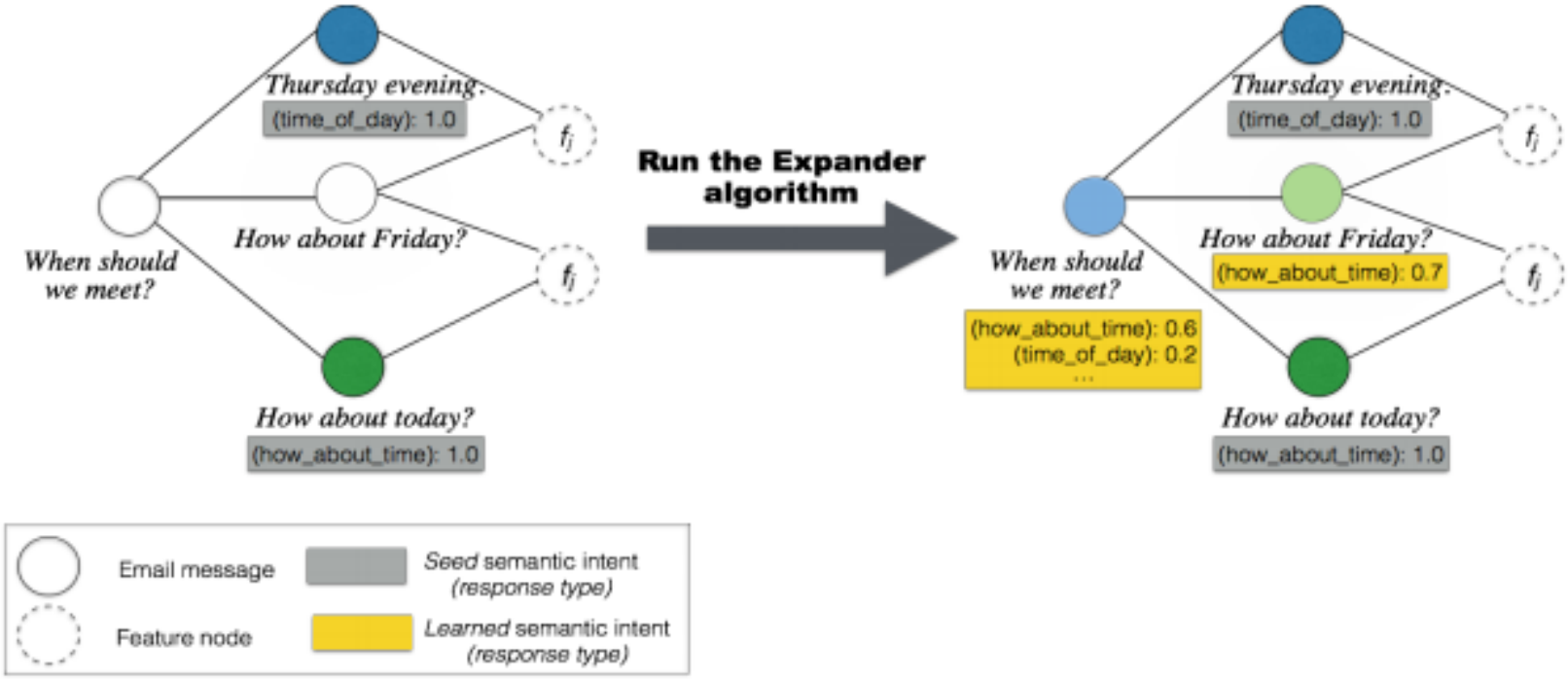


Figure 4: Semantic clustering of response messages.

Inference阶段，每个回复样本会有一个intent分布，将最可能的intent作为这个回复的意图（intent）标记

生成Smart reply模板库IV

- 一个为回复样本标记了intent类别的回复模板库
  - 回复的diversity
  - 选择回复的speed

回复的多样性（diversity）

过滤掉过于generic的回复
e.g. “Yes!”

兼顾正面与负面回复

Unnormalized Responses	Normalized Responses
Yes, I'll be there.	Sure, I'll be there.
Yes, I will be there.	Yes, I can.
I'll be there.	Yes, I can be there.
Yes, I can.	Yes, I'll be there.
What time?	Sure, I can be there.
I'll be there!	Yeah, I can.
I will be there.	Yeah, I'll be there.
Sure, I'll be there.	Sure, I can.
Yes, I can be there.	Yes. I can.
Yes!	Yes, I will be there.

Normalized Negative Responses
Sorry, I won't be able to make it tomorrow.
Unfortunately I can't.
Sorry, I won't be able to join you.
Sorry, I can't make it tomorrow.
No, I can't.
Sorry, I won't be able to make it today.
Sorry, I can't.
I will not be available tomorrow.
I won't be available tomorrow.
Unfortunately, I can't.

Final Suggestions
Sure, I'll be there.
Yes, I can.
Sorry, I won't be able to make it tomorrow.

Table 2: Different response rankings for the message *“Can you join tomorrow's meeting?”*

回复的多样性（diversity）

- 模板库中的回复根据意图（intent）进行聚类
- 在得分最高的一些回复中，对每个（intent）只选择一个回复
  - e.g. score排名前五的选项是(reponse1, intent=1), (reponse2, intent=1), (reponse3, intent=3), (reponse4, intent=1), (response5, intent=2)
  - 选择的前三个选项是reponse1, response3, response5

模型/产品的衡量标准（metric）

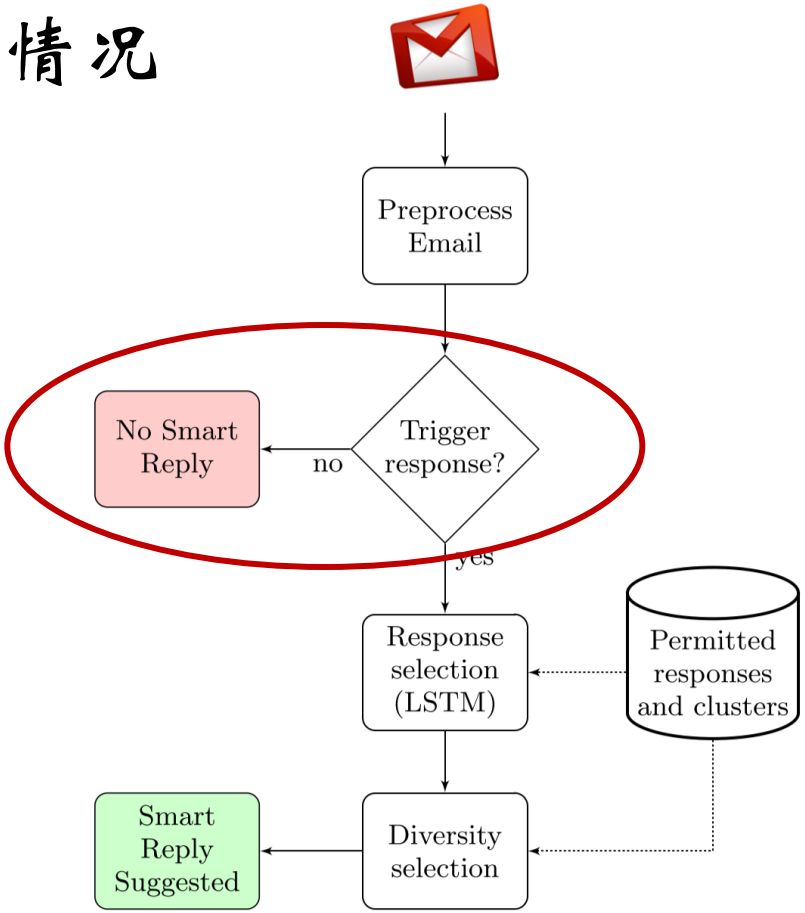
$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$$

Model	Precision@10	Precision@20	MRR
Random	$5.58e-4$	$1.12e-3$	$3.64e-4$
Frequency	0.321	0.368	0.155
Multiclass-BOW	0.345	0.425	0.197
Smart Reply	0.483	0.579	0.267

根据random+Precision的推测，18000左右的模板库
根据mrr的推测，30000左右的模板库

其他

- 适用于简短回复的情况



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