对话方面几篇论文的笔记

2016年2月20日 星期六　　下午4:46

一、简介

总结最近看的几篇对话（dialogue）方面的论文。

对话方面最流行方向的是retrieval-based 和 generation-based，其他方向本文

介绍。

下面是目录：

二、方法

1、基于规则（传统方法，不基于数据/基于很少的数据）

A rule based

相关工作：Weizenbaum，1966

B learning based

简介：reinforcement learning based

相关工作：Litman，2000

2、基于数据（data-driven）

简介：基于大规模数据，而不是规则/推理，利用机器学习、DNN、知识库

A knowledge base

简介：特定知识领域，基于知识，自动问答，question-answer类型

文中不重点

一篇论文
文，还有

等。

型对话，对

数据规模要求不高，领域不需要很宽泛

工作：

Leuski et al. 2006 and Leuski and Traum 2011

statistical language model in cross-lingual information ret

Chen et al. 2011 and Nouri et al. 2011

question generation tool

B retrieval based（popular）

简介：基于检索/召回的对话系统，根据上一句对话，从语料库的现有

配出最合适的回复，借鉴信息检索技术

相关工作：

Jafarpour et al. 2010 learning to chat (L2C)

Zongcheng Ji 2014 Arxiv An Information Retrieval Approach to S

Conversation （华为诺亚方舟）

C generation based（popular）

简介：基于生成的对话系统，根据上一句对话直接生成回复，借鉴机

言模型技术

相关工作：

Koehn et al., 2007 phrase-based SMT

Ritter et al. 2011 conducting short text conversation using

Lifeng shang2015 Neural Responding Machine 基于GRU的

decoder（华为诺亚方舟）

Oriol Vinyals、Quoc V. Le 2015 arxiv A Neural Conversationa

（google）

三、retrieval based

1、简介

这个方向重点介绍这篇论文的工作：An Information Retrieval Approach

Text Conversation，作者Zongcheng Ji，2014发在arxiv上，还没有在会

rieval

有回复中匹

Short Text

器翻译/语

SMT
Jencoder-

l Model

h to Short
会议上发

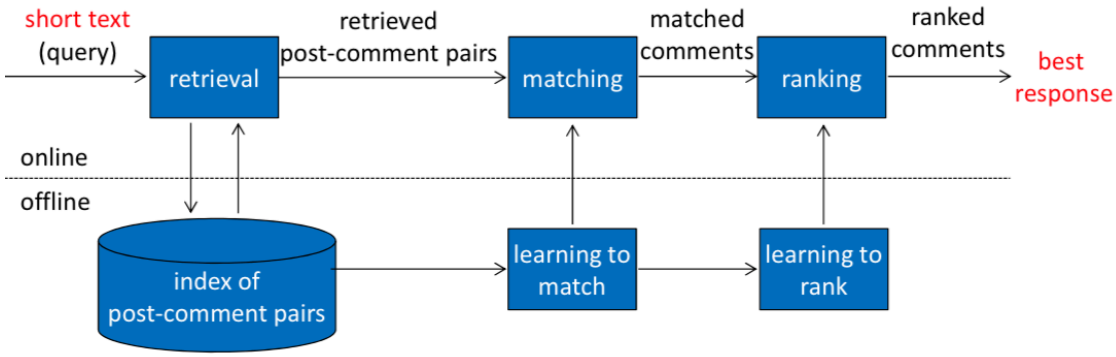
表。

是retrieval based论文中，时间上比较新的，系统描述详细的论文。

只关注单轮会话。

2、系统介绍

系统分为三步



retrieval：先用最基本的文本相似度特征做召回（在三、4、特征的Basic Matching Models部分详细介绍）

matching：离线生成资源的特征值（包括训练那些需要训练的高级特征，Match等）

ranking：基于高级特征，用rankSVM做特征融合做排序任务，做的是一个的ranking，把正例排在负例前面，

3、数据

1）从微博爬的对话数据

2）使用方式

A 用于召回的数据

原始的微博对话，直接挖掘得到的

B 用于训练/测试模型的数据

在原始的微博对话中，人工标注出442组会话的12402个pair的对

注意，这里负例不是随机生成的也是标注的，所以效果会更好

注意！这里数据是特定领域（special-domain）的，而不是开

局限于中国NLP和ML学术圈学者的对话。

Linear

，如Deep

个pairwise

对话。

开放领域的，

4、特征

1）Basic Linear Matching Models

A Query-Response Similarity

$$sim_{Q2R}(\mathbf{q}, \mathbf{r}) = \frac{\mathbf{q}^\top \mathbf{r}}{||\mathbf{q}||||\mathbf{r}||}$$

TF-IDF vectors of q and r

B Query-Post Similarity

$$sim_{Q2P}(\mathbf{q}, \mathbf{p}) = \frac{\mathbf{q}^\top \mathbf{p}}{||\mathbf{q}||||\mathbf{p}||}$$

TF-IDF vectors of q and p

C Query-Response Matching in Latent Space

Wu et al. (2013) 的方法（李航他们自己的方法）

L是学出来的，q和r的表示不是TF-IDF了，可能是embedding

$$LatentMatch(\mathbf{q}, \mathbf{r}) = \mathbf{q}^\top L_\mathbf{q} L_\mathbf{r}^\top \mathbf{r}$$

$$\arg\min_{L_\mathbf{q}, L_\mathbf{r}} \sum_i \max(1 - \sum_i \mathbf{q}_i^\top L_\mathbf{q} L_\mathbf{r}^\top \mathbf{r}_i, 0)$$

2）transLM

为了解决lexical gap 问题，使用Xue et al., 2008的模型，是一种un

language model和translate model结合的方式

$$P_{TransLM}(q|(p,r)) = \prod_{w \in q} P_{TransLM}(w|(p,r))$$

$$P_{TransLM}(w|(p,r)) = (1-\alpha)P_{mx}(w|(p,r)) + \alpha P_{ml}(w|C)$$

$$P_{mx}(w|(p,r)) = (1-\beta)\left[(1-\gamma)P_{ml}(w|p) + \gamma \sum_{t \in p} T(w|t)P_{ml}(t|p)\right] + \beta\left[(1-\gamma)P_{ml}(w|r) + \gamma \sum_{t \in r} T(w|t)P_{ml}(t|r)\right]$$

3）Deep Match

基于DNN，使用Lu and Li (2013).的方法（李航他们自己的方法）

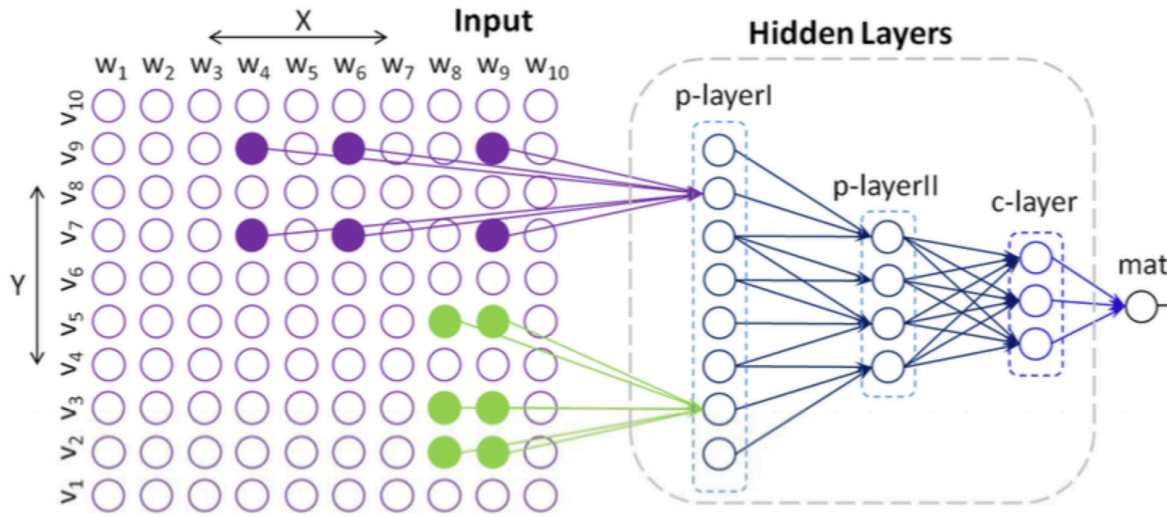
优化方式：loss function hingeloss + pairwise ranking

? ?

nigram

优化方式：loss function hingeloss + pairwise ranking

$$a^{(k)}(\mathbf{x},\mathbf{y}) = f^{(k)}\left((\mathbf{x}^{(k)})^{\top} L_x^{(k)} (L_y^{(k)})^{\top} \mathbf{y}^{(k)} + b^{(k)}\right), \quad k = 1, \cdots$$



hidden layer部分，是MLP计算相似度打分，输入是k个前面的输入
k个输入学成一个score的输出

4）topic word model

基于一些基本语义特征，用logistic regression分类器学出“短文本
topic word “，使用方式是，把是topic word的词进行加权，计算两
相似度，权重是预测topic word的概率。

5、特点

1）小数据、窄领域。

2）系统说明详细。

3）为啥没在会议中被录用？可能用的方法没啥创新点，使用的都是已有的

4）本文对对话领域相关工作（related work）的介绍也很详细，适合去了
的发展。

四、generation based

1、Neural Responding Machine

1）简介

Neural Responding Machine 是lifeng shang和 LI Hang发表在2

$\cdot, K$

ch score
→

，MLP是把

中哪个词是
两个文本的

的技术？
解对话领域

015ACL的

论文。用基于RNN的autoencoder的思想做generation-based的对

和三、中Zongcheng Ji，2014的都是华为诺亚方舟搞的，数据集都
据，两项工作比较可以对比着看。本文方法声称效果比Zongcheng
好。

同样只关注单轮会话。

2）retrieval-base的缺点：

A、不能随意的生成想要的回复，如 主题、情感都特别贴切的

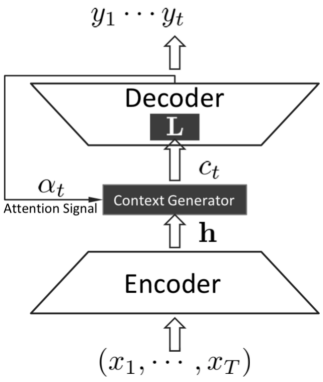
B、如果使用retrieval-based的，一些特征对于检索出合适的回复，
这是本文论述的，不过感觉retrieval-base缺点并不致命，因为这些
generation-based有些牵强。

3）数据

微博数据集。虽然没有说，但是可能和三、中Zongcheng Ji，2014
也没有说明是否是open-domain的

4）模型

encoder-decoder的框架，其中encoder部分做了着重的设计



A decoder部分

类似一个语言模型，除了输入是一个encoder过程中传进来的c
generator（写做c_t）之外，其实就是一个语言模型，而且是

话系统。

是微博数
Ji，2014的

远远不够
缺点换成

差不多。

context-
基于GRU

的。

是一个简单

公式

$$p(y_t|y_{t-1},\cdots,y_1,\mathbf{x}) = g(y_{t-1},s_t,c_t)$$

$$s_t = f(y_{t-1},s_{t-1},c_t)$$

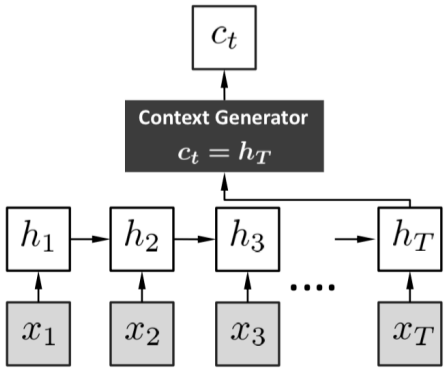
这里s_t是GRU的隐层，g()是softmax

B encoder部分

这里其实是在传入GRU的基础上，在合适的地方加入attention

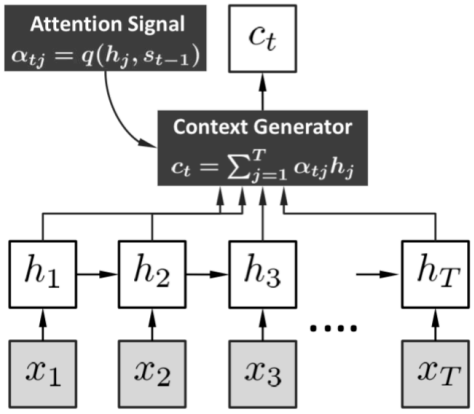
模型的演变分为三个阶段：

1）**Global Scheme**



典型的GRU，没什么特色

2）**Local Scheme**



直接用attention model加权之后产生的向量作为c_t

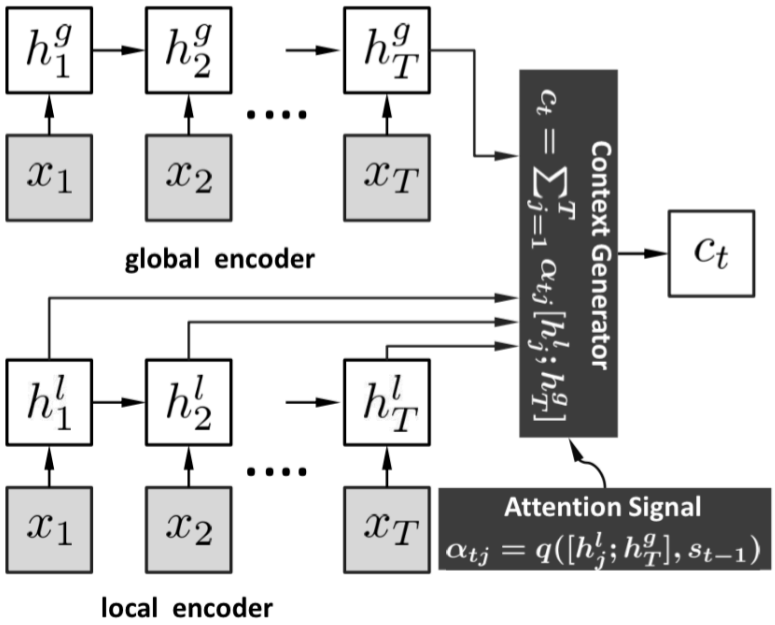
注意这里计算attention权重的时候用的是s_t，应该是encode

果

n model。

r阶段的结

3）**Local and Global Model**



1）global scheme和2）local scheme的融合

这里注意一点：计算attention权重和context generator中，l
encoder部分的隐层用的仍然是local scheme的h_j（这个没问
global用的是整句话的隐层h_T，这原因是h_T是负责对整句话
示，这里希望用整句话的。

训练的过程中，local的和global的单独训练，再combine到一
tune训练。

2、A Neural Conversational Model

1）简介

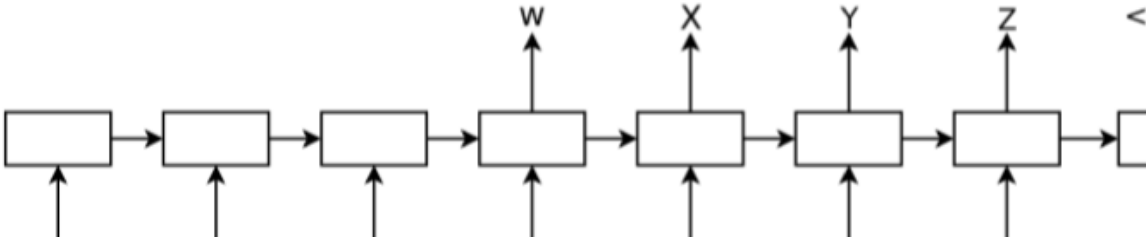
A Neural Conversational Model

google的Oriol Vinyals、Quoc V. Le

2015 arxiv

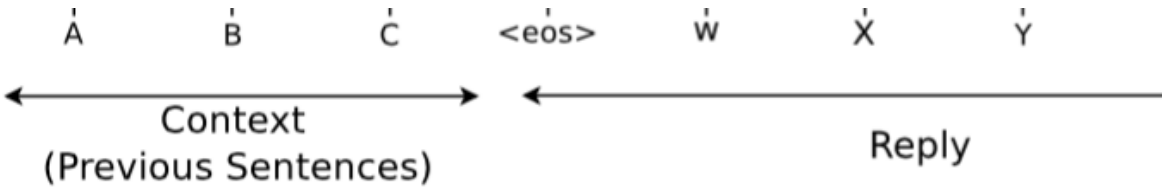
2）思想

没啥特别的想法，没啥特色，就是用一个RNN做生成式会话



ocal
题），而
的一个表

起fine



3）数据集

open-domain 电影字幕 900M+ pairs （这应该是第一个明确说明domain数据集的generation-base的方法）

closed-domain 人工IT服务对话 30M+ pairs

4）实验

没有实验指标，只是贴了些case，讲了一下缺点和未来的方向，应该完成的paper

他提到本方法的缺点：语句不连贯、无法利用知识库资源

五、其他论文

数据集介绍的paper

发现对话任务很多是在特定领域（specific demain）上做的。

发现特征方面TF-IDF是个常用的特征。

李航的talk

在ccir_2015和nlpcc_2015上都有talk介绍对话的整体工作，虽然都是国内不过可以了解华为诺亚方舟的相关工作。

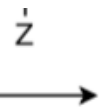
Natural Language Dialogue - Future Way of Accessing Information

Toward Building A Natural Language Dialogue System Using Big Da

Deep Learning

六、参考资料

1503.02364



使用open-

该是一篇没

的会议，

ata and

1408.6988v1

1506.05869v3

nlpcc_2015_keynote

b6d786_b09eff1bdf87…

ccir_2015_hangli