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

第七节课：AllenAI的考试机器人

开放领域的QA

检索

推理

本节内容

- Science QA 案例研究：回答自然科学问题的智能系统(An intelligent system that reads, learns, and reasons about science.)
  - AI2 kaggle比赛的解决方案
  - 实验演示
  - AI2的QA工作
- 开放领域的QA
  - DrQA
  - SQuAD 数据

参考文献

- AllenAI Aristo
  - LSTM-based deep learning models for non-factoid answer **selection**
  - Combining retrieval, statistics, and inference to answer elementary science questions
  - Answering complex questions using open information extraction
  - Creating causal embeddings for QA with minimal supervision
- 开放领域的QA
  - Reading Wikipedia to answer open-domain questions
  - DrQA repo
  - SQuAD数据

AllenAI Aristo：信息检索＋推理

SCIENTIFIC QA

Aristo Kaggle 竞赛

- https://aristo-quiz.allenai.org/
- http://aristo-demo.allenai.org/
- Data demo
  - 小规模数据
  - 需要大量外界信息

三个主要解决方案

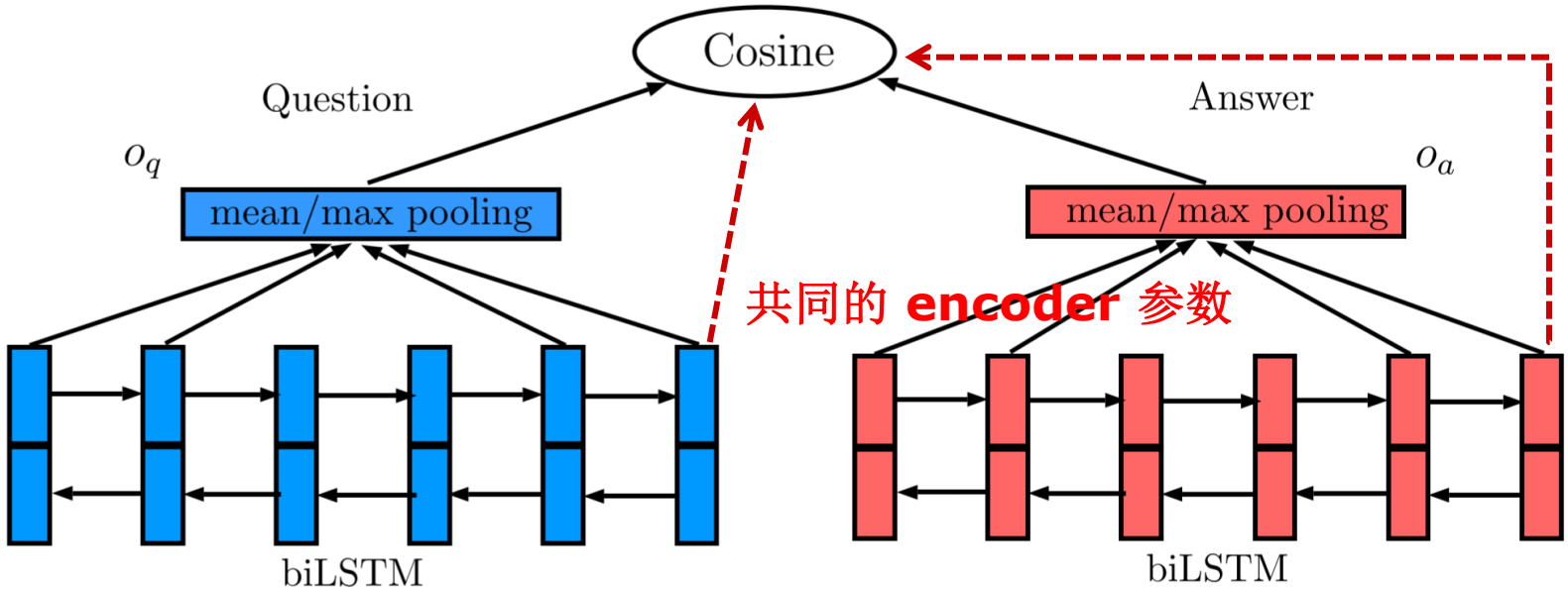
- 深度学习
- 信息检索
  - 代码演示
- 逻辑推理

Deep learning + “shallow” reasoning

- LSTM-based deep learning models for non-factoid answer **selection** (2015)
- 给出问题q，在一个备选答案的集合里面$\{a_1, a_2, \cdots a_s\}$选择最合适的答案a_k
  - 类似于rank-based chatbot任务
  - 将不定句子长度的问题和备选答案编码成feature vectors
  - 量化feature vectors的相关程度

dual-LSTM 升级版

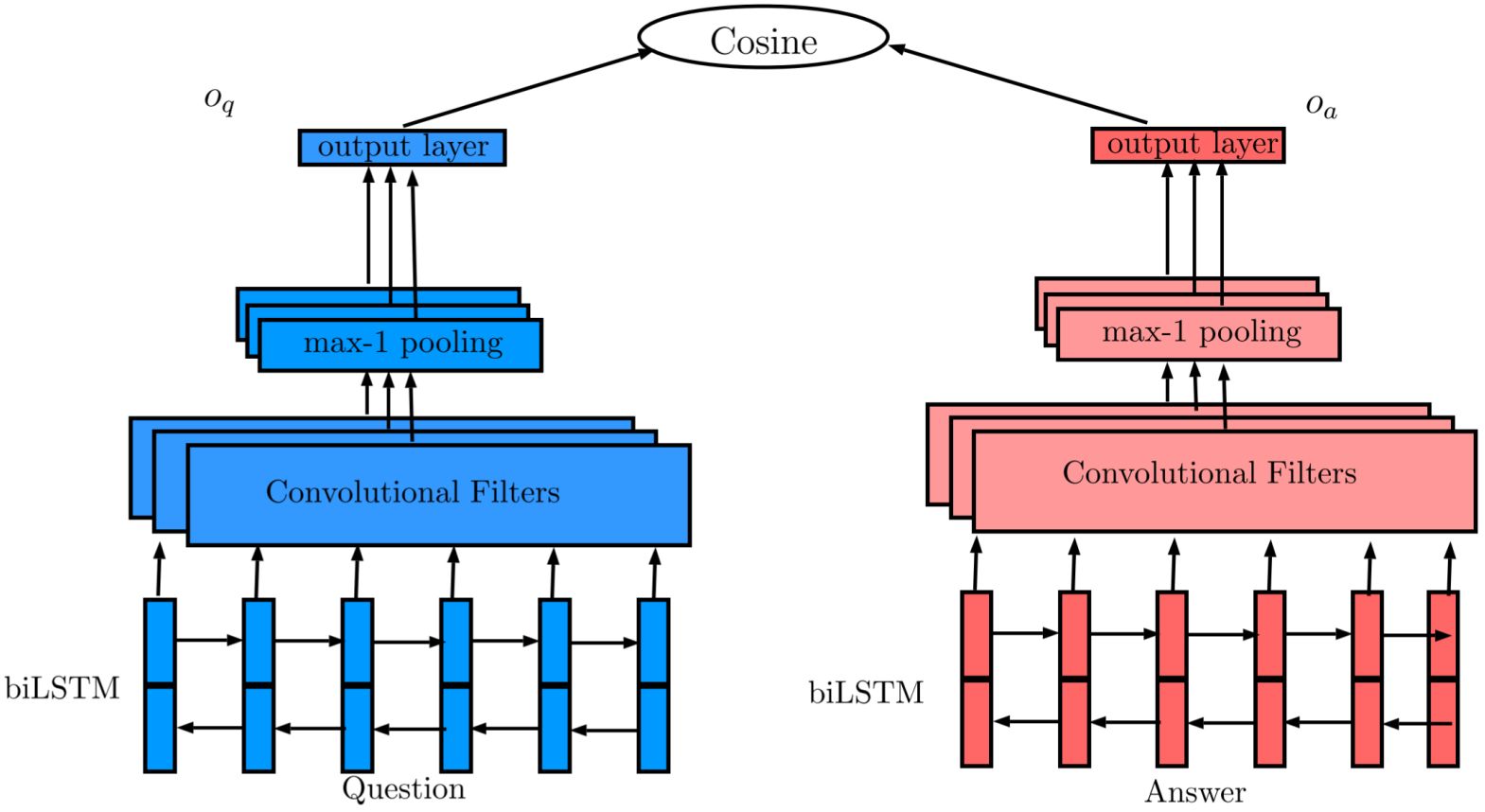
$$L = \max(0, M + cosine(q, a_-) - consine(q, a_+))$$

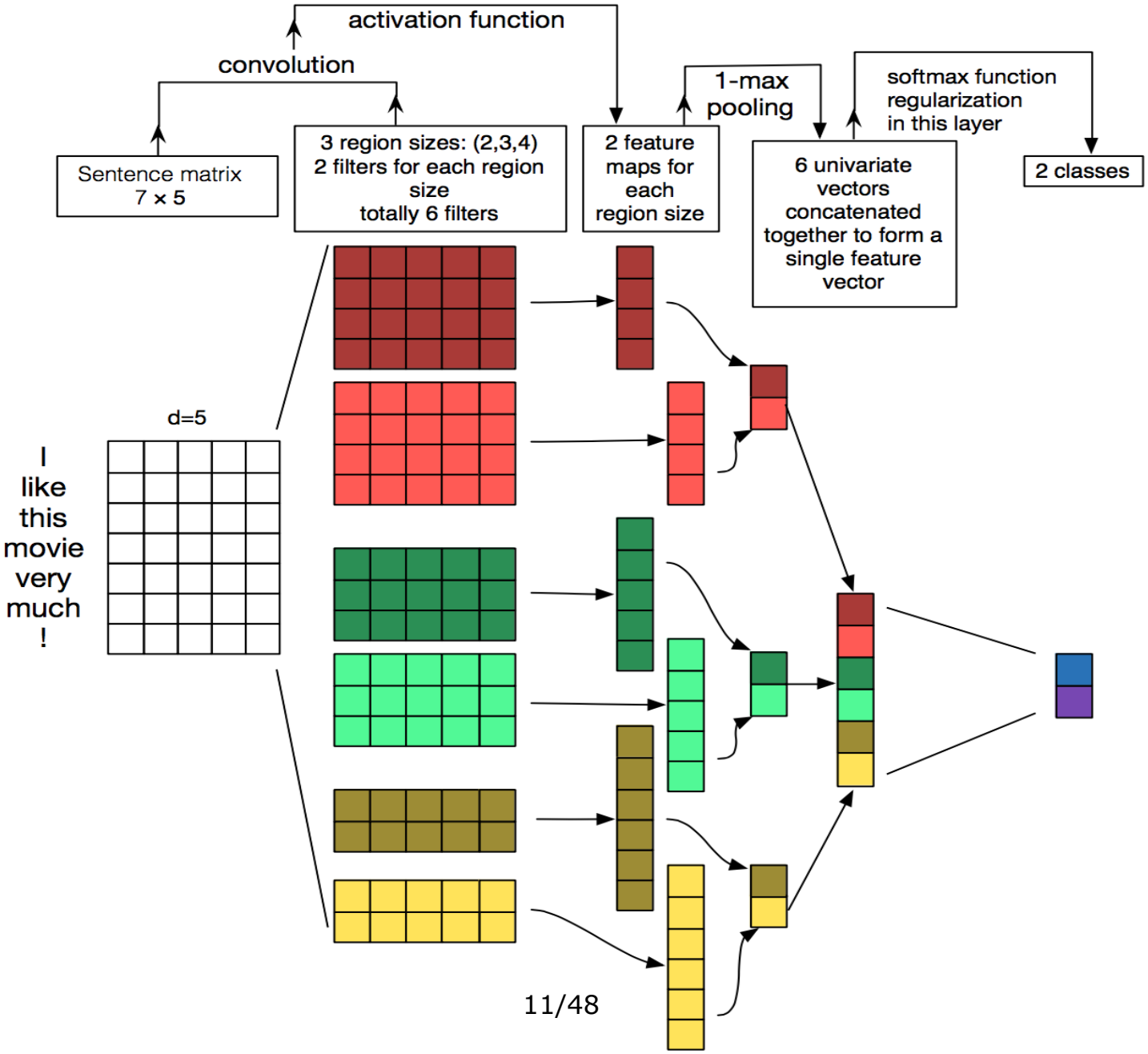


对比第三节课：不再依赖最后一个时间的state，而是所有时间的state的综合

	Model	Validation	Test1	Test2
A	QA-LSTM basic-model(head/tail)	54.0	53.1	51.2
B	QA-LSTM basic-model(avg pooling)	58.5	58.2	54.0
C	QA-LSTM basic-model(max pooling)	64.3	63.1	58.0

dual-LSTM + CNN




activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

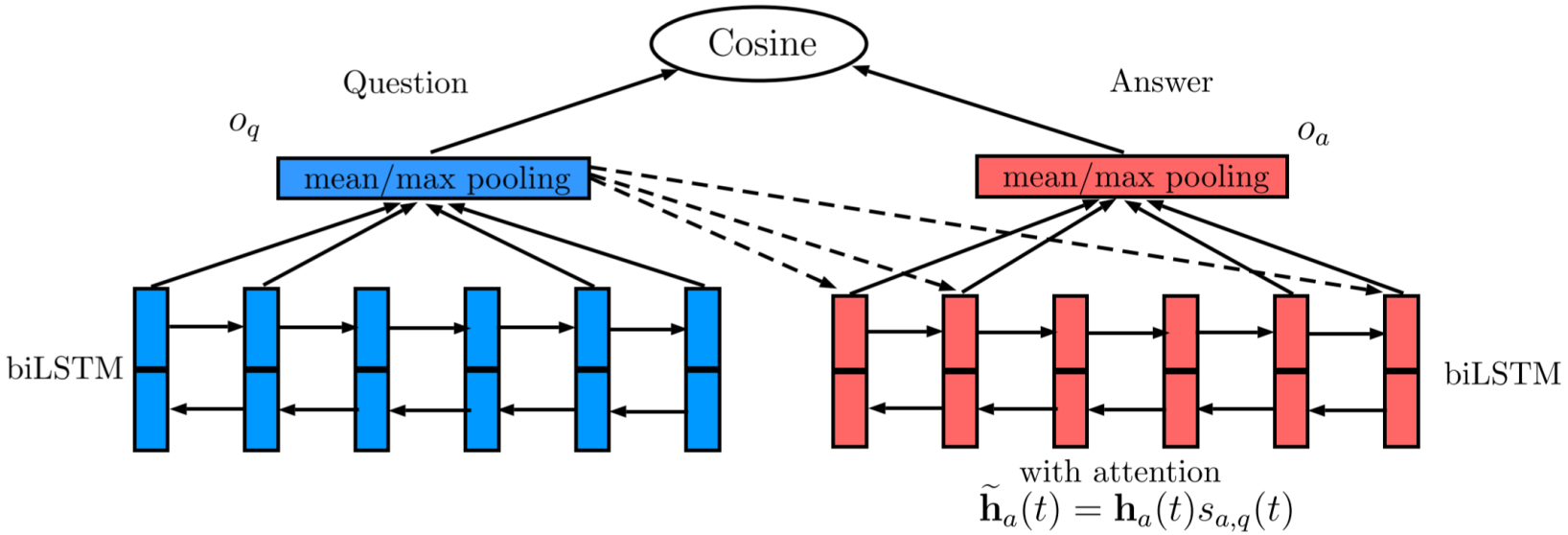
```
pooled_outputs = []
for i, filter_size in enumerate(filter_sizes):
    with tf.name_scope("conv-maxpool-%s" % filter_size):
        # Convolution Layer
        # num_filters 个维度为[filter_size, embedding_size, 1]的filter用在
        # 维度为[batch_size, max_length, embedding_size, 1]的 embedded_chars_expanded
        filter_shape = [filter_size, embedding_size, 1, num_filters]
        W = tf.Variable(tf.truncated_normal(filter_shape, stddev=0.1), name="W")
        b = tf.Variable(tf.constant(0.1, shape=[num_filters]), name="b")
        conv = tf.nn.conv2d(
            self.embedded_chars_expanded,
            W, strides=[1, 1, 1, 1], padding="VALID",
            name="conv")
        # conv.shape = [batch_size, max_length - filter_size + 1, 1, num_filters]

        # Apply nonlinearity
        h = tf.nn.relu(tf.nn.bias_add(conv, b), name="relu")

        # Maxpooling over the outputs
        # 将维度为[batch_size, max_length - filter_size + 1, 1, num_filters]
        # 的tensor变成维度为[batch_size, 1, 1, num_filters]的tensor (其实是matrix)
        pooled = tf.nn.max_pool(
            h,
            ksize=[1, sequence_length - filter_size + 1, 1, 1],
            strides=[1, 1, 1, 1], padding='VALID', name="pool")

        # 将num_filters个宽度为filter_size的filter产生的num_filter个结果和
        # 其他filter_size的convolution结果拼接起来
        pooled_outputs.append(pooled)
```

dual-LSTM + attention

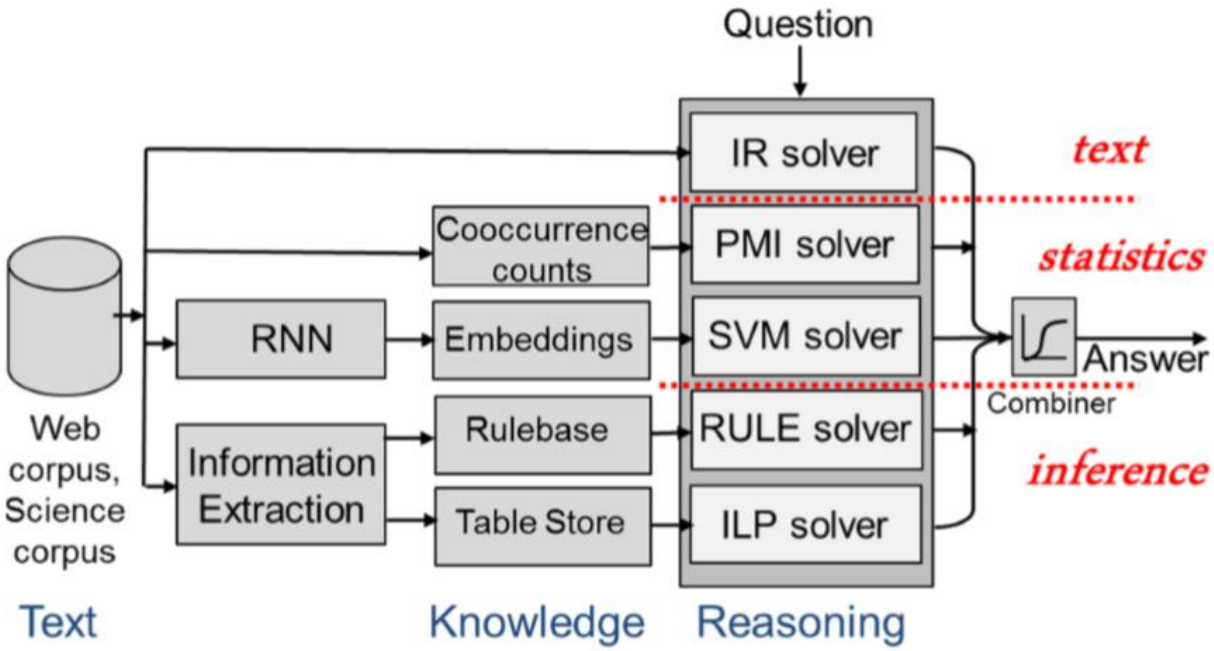


根据和query相关的程度计算response里面每个词语的state的权重

细节见DrQA部分

Aristo: AllenAI的解决方案

- Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions (2016)
- 结合多种模型的解决方案



Information Retrieval (IR) Solver

- 使用关键词语，从text格式的知识库里面
  - 检索（问题q，备选答案a_i）最相关的文本
  - 使用检索结果给备选答案a_i打分
- 最有效的方法之一

Co-ocurrance Statistics

- PMI solver
  - 使用mutual information衡量一对词语的相关程度
  - 来自问题q和来自备选答案a_i的n-gram的组合的最大PMI：$\max PMI(x, y), x \in q, y \in a_i$
- SVM solver
  - 使用embedding向量的cosine similarity衡量一对词语的相关程度
  - （词语的平均cosine similarity，句子的cosine similarity) 来产生衡量句子的相关程度的特征
  - 使用SVM ranker挑选答案

结构化知识

- 规则与表格化信息
  - 前面的textual和statistical并不能理解问题
  - 容易被检索的信息或者词汇误导
  - 例：

PMI无法解决：

- 如果问题里面的词语和错误答案里面的词语有强的相关度
- 例：Which characteristic can a human offspring inherit?
  - (A) facial scar (B) blue eyes **(C) long hair** (D) broken leg
  - “human” 和 “hair” 经常共同出现，干扰“blue eyes” 和 “inherit”
- 如果有极性转换词(negation, good/bad, shortest/longest等)
- 例：Which activity is an example of a good health habit?
  - (A) watching television **(B) smoking cigarettes** (C) eating candy (D) exercising every day
  - “health” 和 “cigarettes” 相对于“good health” 和 “exercising”更加经常共同出现

RULE

- 从自然科学类文献中提取有(cause, enables, purpose, requirement, condition, part, example)关系的记录
- 建立(tupleimplicationtuple)关系库
  - 例： “Some animals grow thick fur to stay warm.”
  - (("Some animals" "grow" "thick fur") EFFECT ("Some animals" "stay" "warm"))
- 例（从抽象规则解决具体问题）：A turtle eating worms is an example of
  - (A) breathing (B) reproducing(C) eliminating waste (D) taking innutrients
  - 根据规则 “IF animals eat THEN animals get nutrients”选择 (D)

ILP solver 和 proof graph

- 美国夏至在几月？
  - 北半球夏至在六月；美国位于北半球；美国夏至在六月

Country	*Location*
France	north hemisphere
USA	north hemisphere
...	
Brazil	south hemisphere
Zambia	south hemisphere
...	

Hemisphere	*Orbital Event*	*Month*
northern	summer solstice	Jun
northern	winter solstice	Dec
northern	autumn equinox	Sep
...		
southern	summer solstice	Dec
southern	autumn equinox	Mar
...		

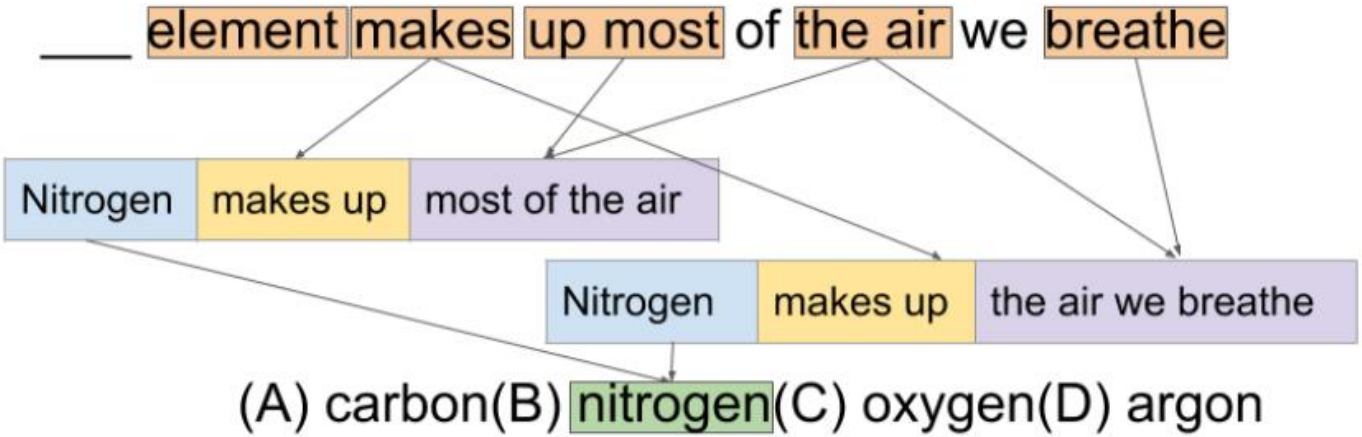
- ILP solver：使用整数优化从一些表中提取出和问题相关的一系列信息
  - Answering complex questions using open information extraction (2017)

ILP solver 和 proof graph

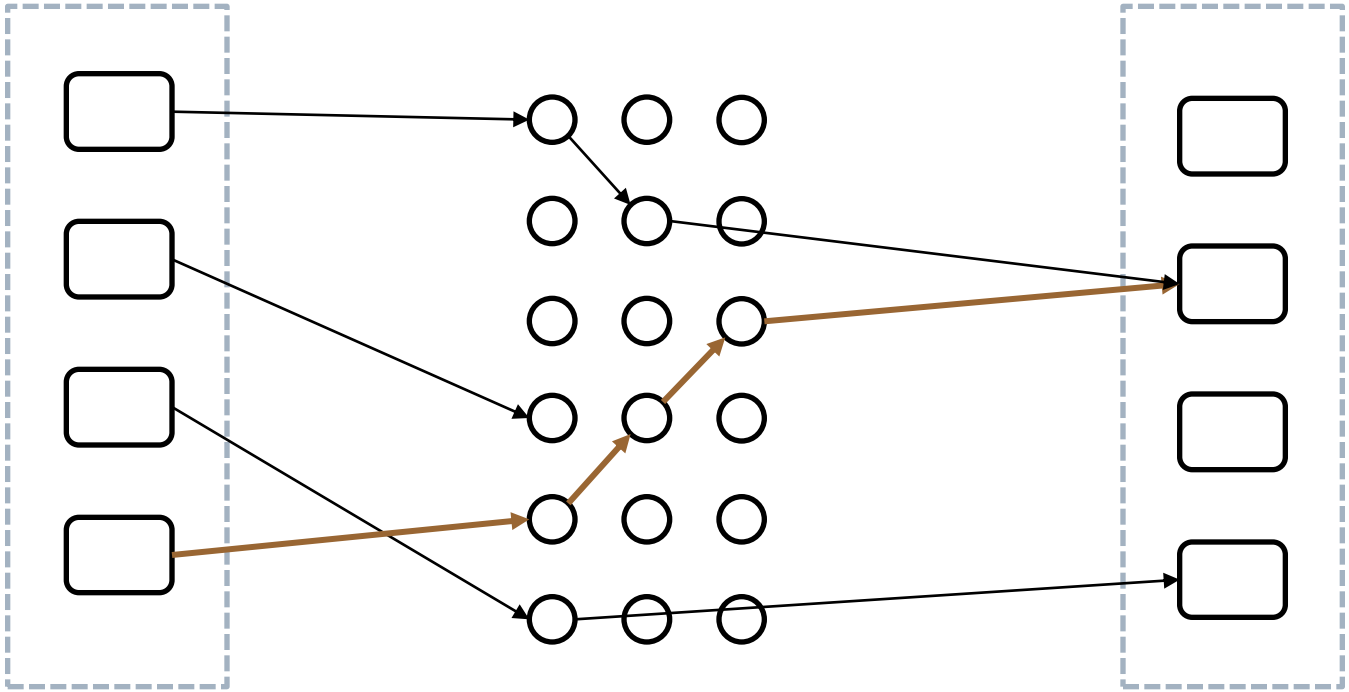
- 下面哪个是太阳系里面能反光的并且绕一个星球旋转的卫星？
  - (A) 地球 (B) 水星 (C) 太阳 (D) 月球
- 需要综合多条信息推理解答
  - (Moon; is; in the solar system),
  - (Moon; reflects; light)
  - (Moon; is; satellite)
  - (Moon; orbits; around one planet).

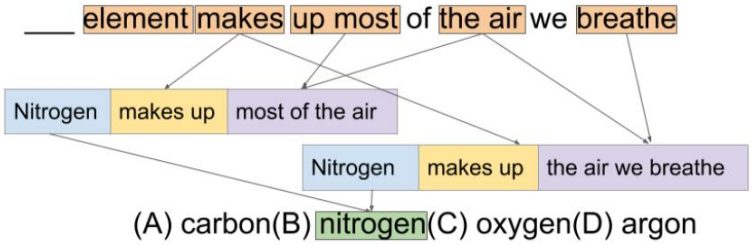
ILP solver 和 proof graph

- {query, candidate answer, KB}：巨大的知识graph的nodes
- 使用整数优化寻找proof graph



ILP solver 和 proof graph





ILP solver 和 proof graph

- 擅长需要综合多项信息来回答问题的情况
- 例：Which gas is given off by plants?
  - (A) Hydrogen (B) Nitrogen (C) Oxygen (D) Helium
  - 基于下面两条信息选择C
  - **has-part(“plant”, “stomata”)**
  - **part-output(“stomata”, “oxygen”)**

ILP solver 和 proof graph

- 尚且不能解决的问题：
- 遗漏关键词，例：Which material will spread out to completely fill a larger container?
  - (A) air (B) ice (C) sand (D) water
  - 知识库中有“water will spread out and fill a larger container”着一条记录，但是并不包含“完全(completely)”这一个关键词.
- Alignment 问题，例： Which of the following gases is necessary for humans to breathe in order to live?
  - (A) Oxygen (B) Carbondioxide (C) Helium (D) Water vapor
  - 知识库中有“human breathe **out** carbondioxide”这样的记录

困难问题

- 1. 比较类问题(Comparison questions)
  - 例：...moves faster or slower than ...?
- 2. 需要数学运算的推理(Simple arithmetic reasoning)
  - 例：7 rotations of Earth equals (A) 1 week (B) 2 weeks
- 3. 复杂的推理(Complex inference)
  - 例：A white rabbit is best protected in (A) a snowy field..
- 4. 结构化问题(Structured questions)
  - 例：Which structure is correctly paired with its function?
- 5. 基于故事的问题(Story Questions)
  - 例：A puddle formed. Then the sun came out ...

Aristo成绩

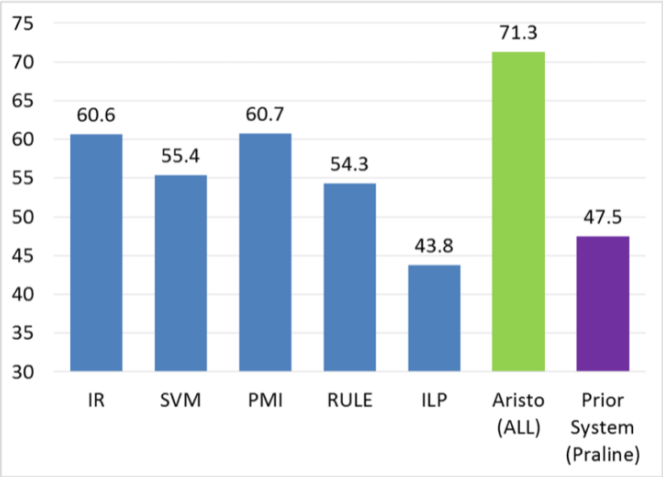


Figure 2: Aristo significantly outperforms all individual solvers and the best previously published system (Praline).

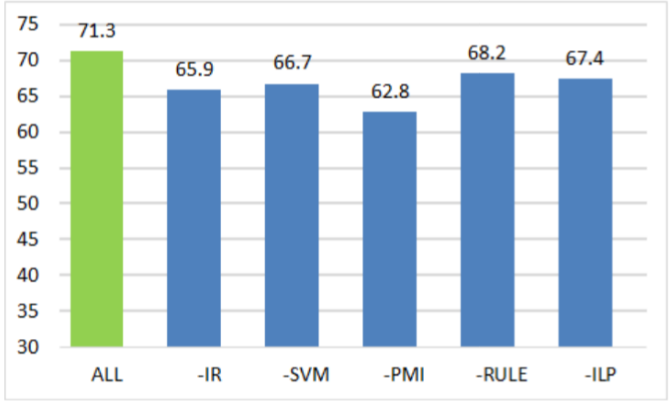


Figure 3: Effect of removing an individual solver from the ensemble. The results suggest each solver contributes to the overall score.

四年级题目的测试成绩

回顾词向量 embedding

基于词向量的分类器

- 使用word2vec，不使用其他信息来源，lstm可以达到不错的结果
- 使用外界信息（studystack）重新训练词向量，可以达到更好的效果

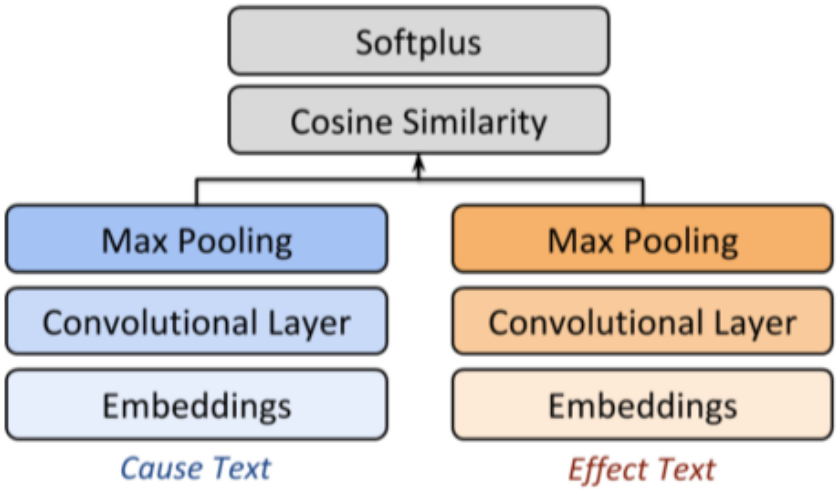
Model	Default Embedding	Story Embedding
QA-LSTM w/Attention	62.93%	76.27%
QA-LSTM bidirectional w/Attention	60.43%	76.27%

AI2的causal embedding

- Creating causal embedding for QA with minimal supervision （2016）
- 第一步：提取有因果关系的样本(cause-effect pair)
  - 例：The collapse of the housing bubble caused stock prices to fall
  - cause (the collapse of the housing bubble)
  - effect (stock prices to fall)

AI2的causal embedding

- 第二步：训练模型预测因果关系

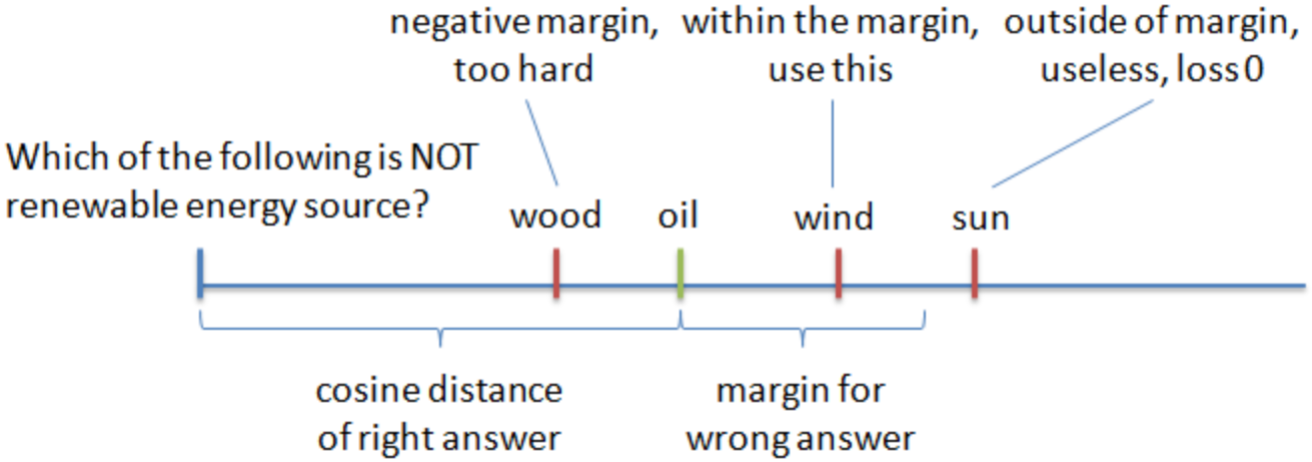


AI2的causal embedding

- 第三步：使用causal embedding做reranking
  - 使用retrieval检索一个备选答案集合
  - 使用causal embedding和”learning to rank”重新排序挑选最佳
- 一个ranking模型中的实际问题：Negative sampling
  - e.g. 在基于检索的chatbot里面，训练或者测试的时候，真实和整个模板库比较1) 低效且 2)易错
  - 选择适当难度的错误回复1)随机样本太弱 2)不选过于挑战的“错误”回复

AI2的causal embedding

- 一个ranking模型中的实际问题：Negative sampling
  - e.g. 在基于检索的chatbot里面，训练或者测试的时候，真实和整个模板库比较1) 低效且 2)易错
  - 选择适当难度的错误回复1)随机样本太弱 2)不选过于挑战的“错误”回复 （选择semi-hard negative样本）



选择当前的模型能够正确分类但是不够confident的错误回复

图片来源

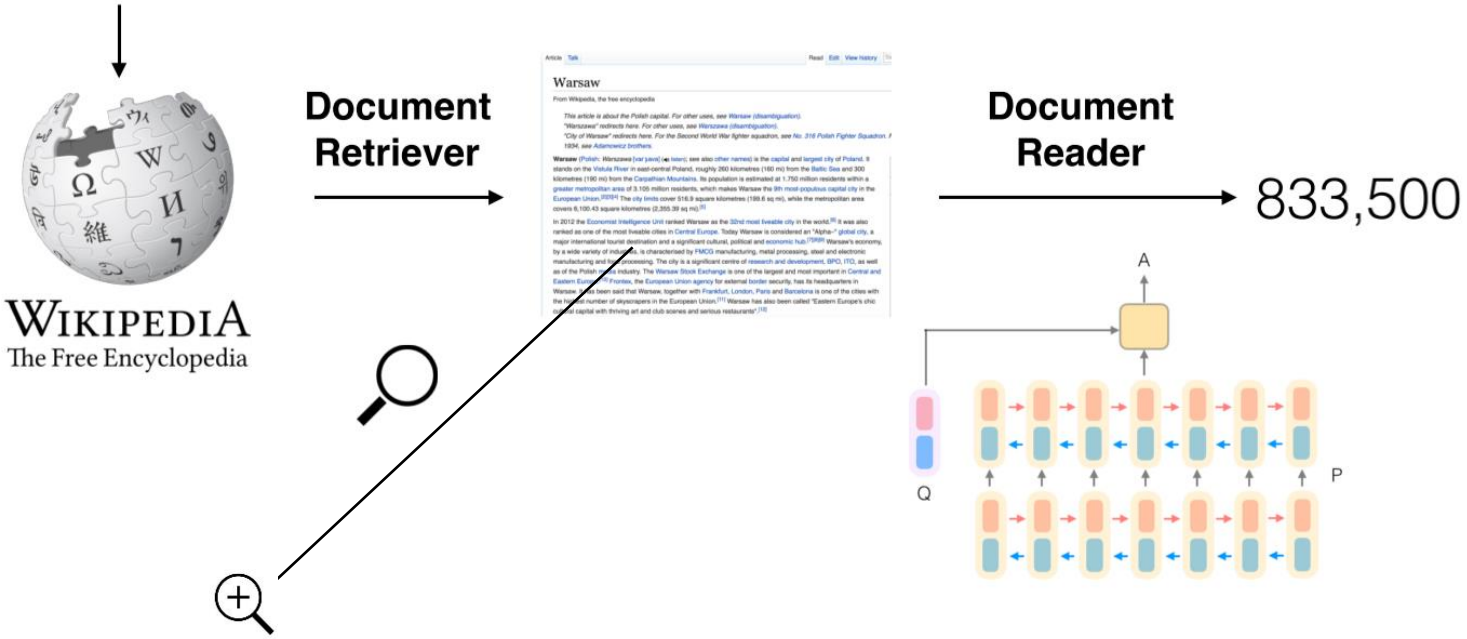
信息检索+信息理解

OPEN-DOMAIN QA

使用Wikipedia 的开放领域 QA

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



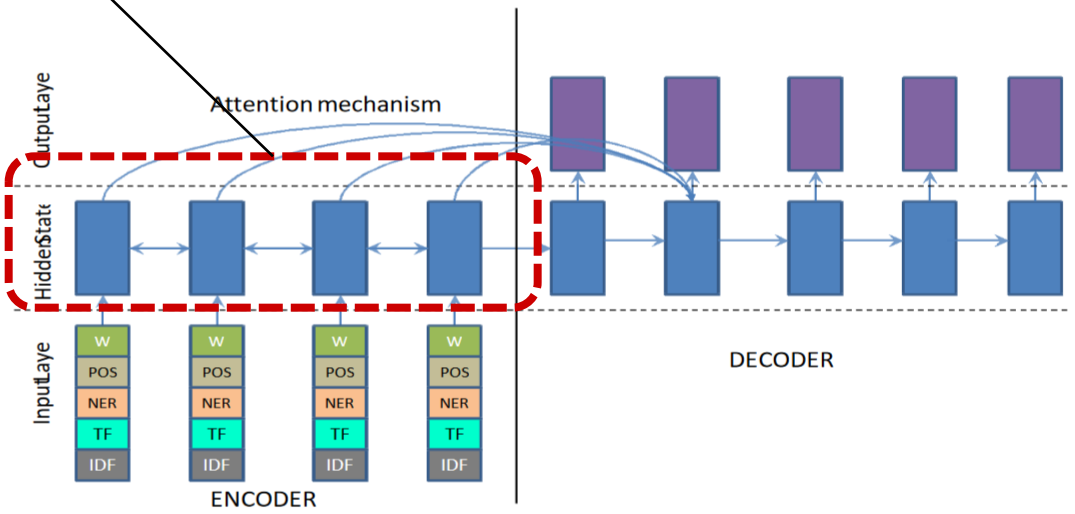
total population.[42] In 1933, out of 1,178,914 inhabitants 833,500 were of Polish mother tongue.[96] World War II changed the demographics of the city, and to this day there is much less ethnic diversity than in the previous 300 years of Warsaw's history.[42] Most of the modern day

Document Retriever: 维基百科上的文本检索

- TfIdf + bigram + hashing
  - 使用TF-IDF weighted bag-of-word 向量衡量问题和wiki文章的相关程度

Document Reader 1:特征工程

- 对检索到的段落使用RNN提取特征
  - $\{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_m\} = RNN(\{\hat{\boldsymbol{p}}_1, \cdots, \hat{\boldsymbol{p}}_m\})$
  - 保留RNN中每个单词对应的hidden state



第五节课：Abstractive text summarization using sequence-to-sequence RNNs and beyond

使用多种输入特征的RNN编码

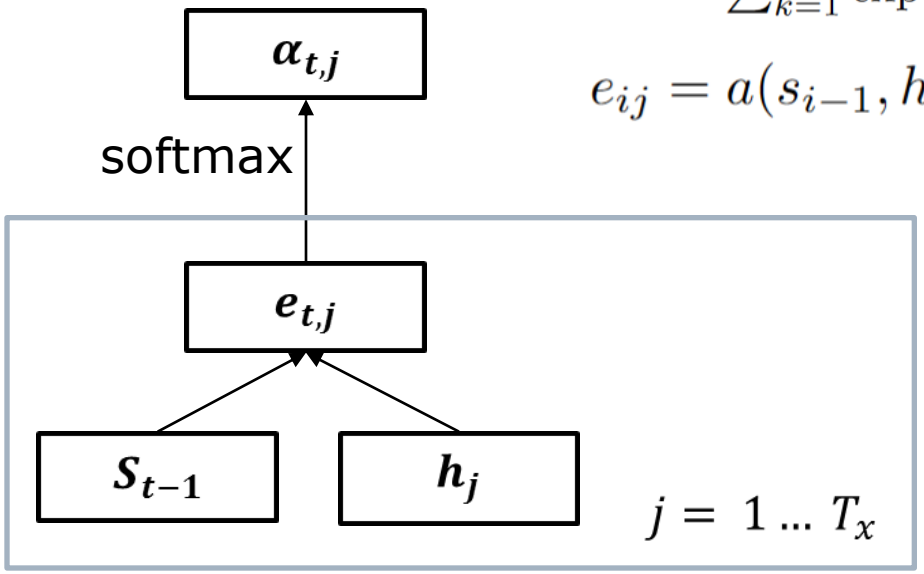
- 以一个段落中的单词p_i为例，计算下列特征并拼接起来作为RNN ENCODER的输入特征
  - 300维的Glove 词向量 $f_{emb}(p_i) = \boldsymbol{E}(p_i)$;
    - 使用glove初始化后，在训练QA模型的过程中fine-tune最常见的1000个单词
  - Token特征：{POS, NER, TF}
  - 信息检索特征(*exact match*): {p_i, lower-case(p_i), p_i的单词原型}是否出现在问题里面
  - “问题”对p_i的看法：$f_{align}(p_i) = \sum_j a_{ij}\boldsymbol{E}(q_j)$
    - 见下页

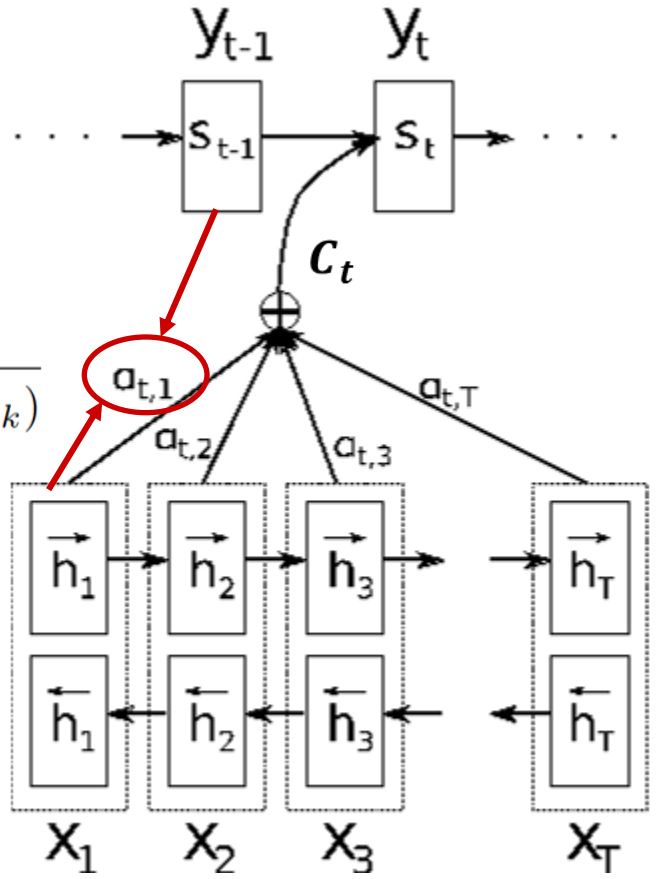
attentive特征$f_{align}(p_i)$

$$c_i = \sum_{i=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$





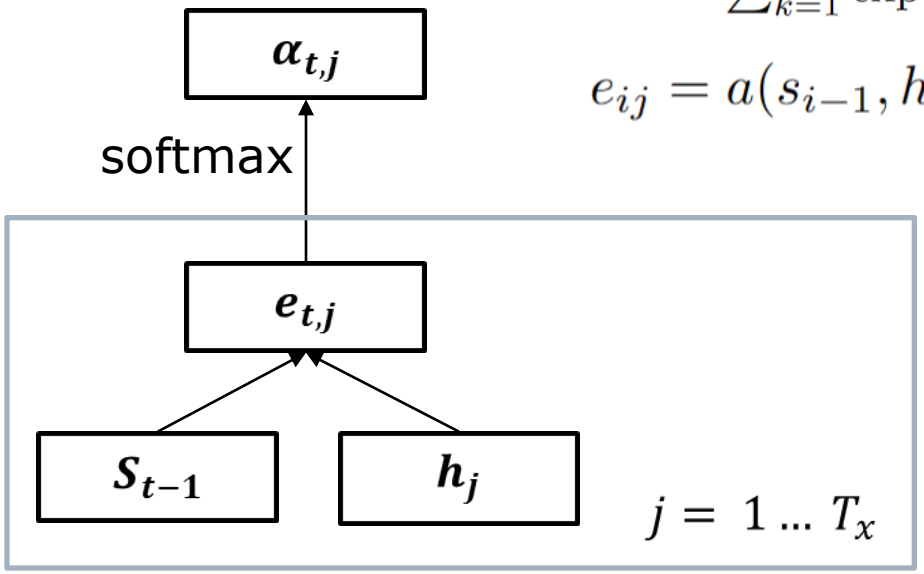
RNN Encoder–Decoder
随时调整的 context vector

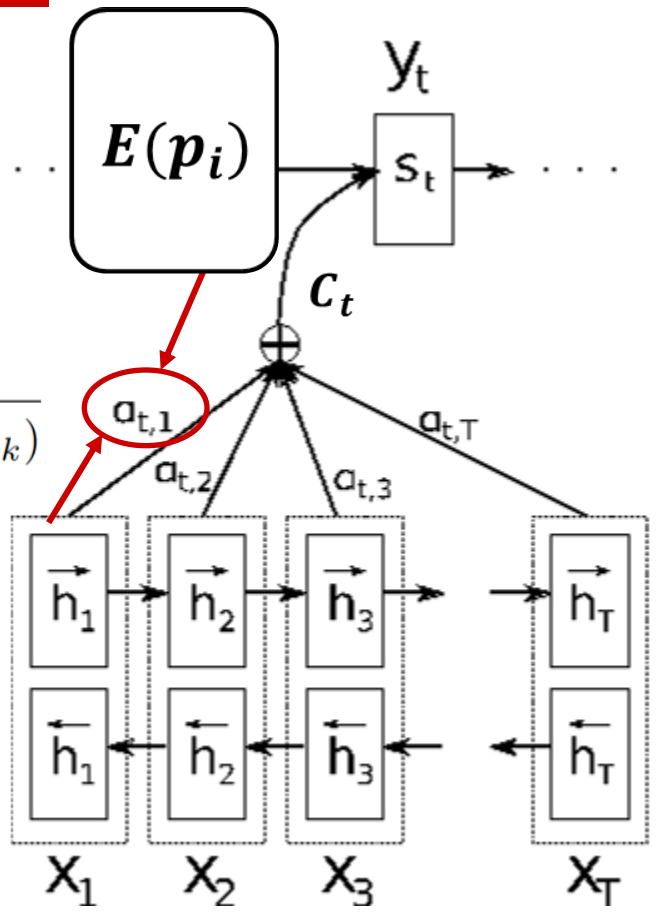
attentive特征$f_{align}(p_i)$

$$c_i = \sum_{i=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$





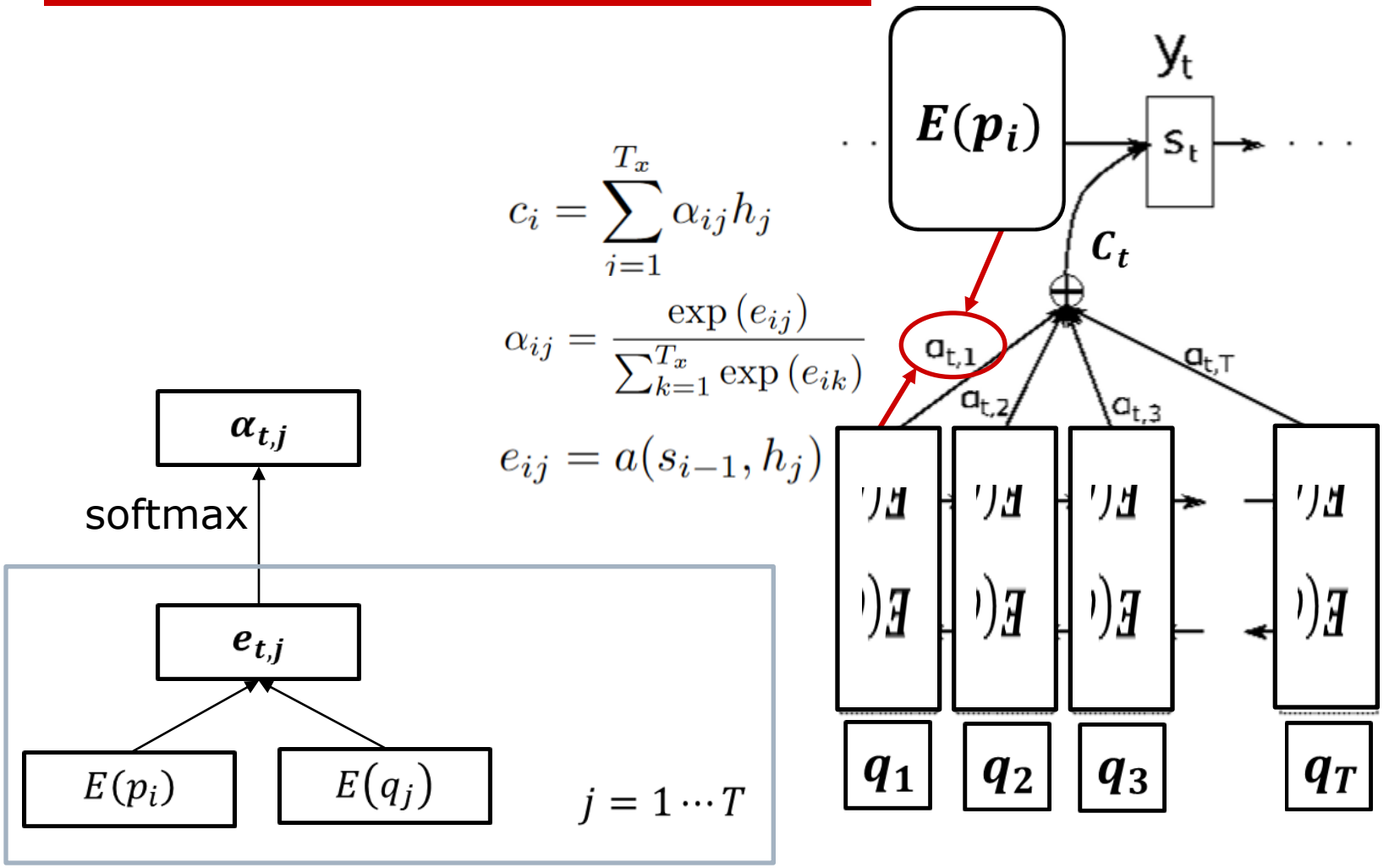
RNN Encoder–Decoder
随时调整的 context vector

attentive特征$f_{align}(p_i)$

$$c_i = \sum_{i=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

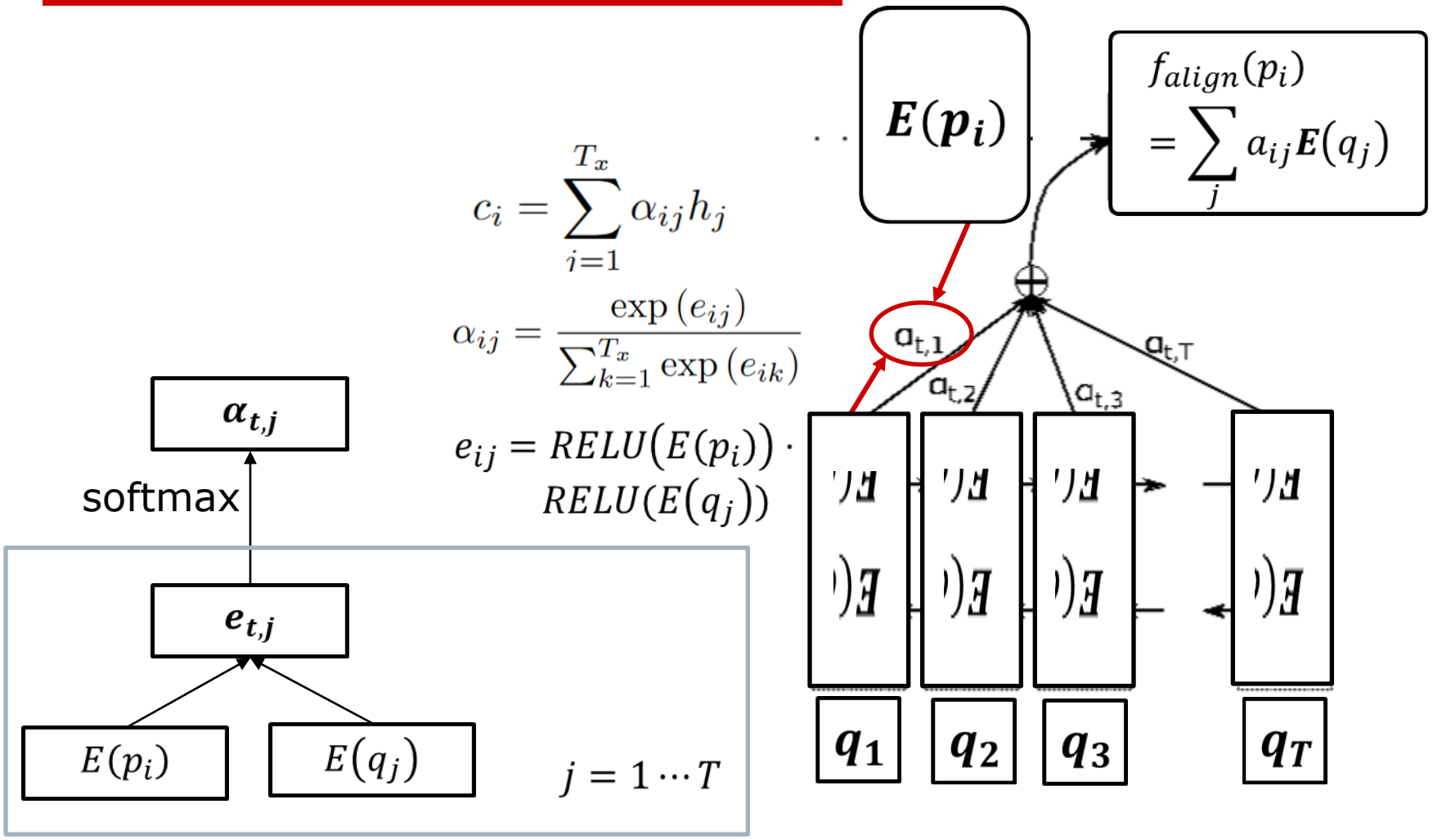


attentive特征$f_{align}(p_i)$

$$c_i = \sum_{i=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = RELU\big(E(p_i)\big) \cdot RELU\big(E(q_j)\big)$$



Document Reader 1:特征工程

- 对问题使用RNN提取特征
  - $\{\boldsymbol{q}_1, \cdots, \boldsymbol{q}_m\} = RNN(\{E(q_1), \cdots E(q_l)\})$
  - 每个问题使用一个向量总结：
    - $\boldsymbol{q} = \sum_j b_j \boldsymbol{q}_j$
    - $\mathrm{b_j} = \frac{\exp(\boldsymbol{w} \cdot \boldsymbol{q}_j)}{\sum_k \exp(\boldsymbol{w} \cdot \boldsymbol{q}_k)}$

Notation

$\{q_j, \boldsymbol{E}(q_j), \boldsymbol{q}_j\}$: 问题中的{第j个词语, 第j个词语的词向量, 第j个RNN cell的state}

$\boldsymbol{w}$:一个待学习的参数

Document Reader 2 :预测

1. 在备选段落中提取最相关的片段

Demographically, it was the most diverse city in Poland, with significant numbers of foreign-born inhabitants.[94] In addition to the Polish majority, there was a significant Jewish minority in Warsaw. According to the Russian census of 1897, out of the total population of 638,000, Jews constituted 219,000 (around 34% percent).[95] Warsaw's prewar Jewish population of more than 350,000 constituted about 30 percent of the city's total population.[42] In 1933, out of 1,178,914 inhabitants 833,500 were of Polish mother tongue.[96] World War II changed the demographics of the city, and to this day there is much less ethnic diversity than in the previous 300 years of Warsaw's history.[42] Most of the modern day population growth is based on internal migration and urbanisation.

2. 在所有摘录片段中寻找最合适的一个

在备选段落中提取最相关的片段

- SQuAD数据
  - Wikipedia的段落
  - Crowdworker用自己的语言提问，挑选段落中的一部分内容作为答案
  - EM和F1 metric

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

在备选段落中提取最相关的片段

$$P_{start}(i) \propto \exp(\boldsymbol{p}_i \cdot W_s \cdot \boldsymbol{q})$$

Demographically, it was the most diverse city in Poland, with significant numbers of foreign-born inhabitants.[94] In addition to the Polish majority, there was a significant Jewish minority in Warsaw. According to the Russian census of 1897, out of the total population of 638,000, Jews constituted 219,000 (around 34% percent).[95] Warsaw's prewar Jewish population of more than 350,000 constituted about 30 percent of the city's total population.[42] In 1933, out of 1,178,914 inhabitants 833,500 were of Polish mother tongue.[96] World War II changed the demographics of the city, and to this day there is much less ethnic diversity than in the previous 300 years of Warsaw's history.[42] Most of the modern day population growth is based on internal migration and urbanisation.

$$P_{end}(i) \propto \exp(\boldsymbol{p}_i \cdot W_s \cdot \boldsymbol{q})$$

$$\max P_{start}(i) \times P_{end}(i') \text{ s.t. } i \leq i' \leq i + 15$$

DrQA的综合效果

Method	Dev		Test	
	EM	F1	EM	F1
Dynamic Coattention Networks (Xiong et al., 2016)	65.4	75.6	66.2	75.9
Multi-Perspective Matching (Wang et al., 2016)†	66.1	75.8	65.5	75.1
BiDAF (Seo et al., 2016)	67.7	77.3	68.0	77.3
R-net†	n/a	n/a	71.3	79.7
DrQA (Our model, Document Reader Only)	**69.5**	**78.8**	70.0	79.0

上：Documents reader效果
下：Documents retreiver和reader最终效果

Dataset	YodaQA	DrQA		
		SQuAD	+Fine-tune (DS)	+Multitask (DS)
SQuAD *(All Wikipedia)*	n/a	27.1	28.4	29.8
CuratedTREC	31.3	19.7	25.7	25.4
WebQuestions	39.8	11.8	19.5	20.7
WikiMovies	n/a	24.5	34.3	36.5

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