数据挖掘导论
Introduction to Data Mining

第五章 文本挖掘

刘 淇
Email: qiliuql@ustc.edu.cn
课程主页：
http://staff.ustc.edu.cn/~qiliuql/DM2017YZ.html

目录

独热表示：One-Hot Representation

- NLP问题转化为机器学习问题需要将语言数学化
  - 最直观的方法：one-hot representation
  - 将词表示为与词库大小等长的向量，仅一个维度的值为1，其余为0
  - 例子：“I want to eat Chinese food”
    - 词库大小=6
    - “I” 的表示为 [1 0 0 0 0 0]
    - “want” 的表示为 [0 1 0 0 0 0]
    - ...
    - “food” 的表示为 [0 0 0 0 0 1]
  - 问题：
    - 一般语料库规模特别大，那么one-hot向量的维度非常大
    - 语义鸿沟：词之间的语义关系被忽略，如“腾讯”和“小马哥”没有关联

分布式表示：Distributed Representation

- 解决one-hot representation缺点：分布式表示
  - 利用低维向量表示每个词（如50维，100维）
  - 每个向量是稠密的
  - 例子：
    - 如某个词的表示为 [0.365 -0.297 -0.139 ..... 0.267 -0.185]
  - 作用：
    - 可以计算词之间的相似度（如使用余弦相似度）
    - 可以将词的分布式表示直接用于其他任务（如分类，预测）
  - 如何得到？
    - 通常这种词的分布式表示是语言模型的副产品
      - 即训练好语言模型后，部分学到的参数就是词的分布式表示

语言模型

- 给定字符串 $w_1, w_2, ..., w_t$,计算它是自然语言的概率 $P(w_1, w_2, ..., w_t)$
- N-gram模型（前面已经介绍）
  - 利用 $P(w_t \mid w_{t-n+1}, ..., w_{t-1})$ 近似 $P(w_t \mid w_1, w_2, ..., w_{t-1})$ ← 计算开销很大
- 神经网络语言模型
  - 利用前n-1个词，预测下一个词

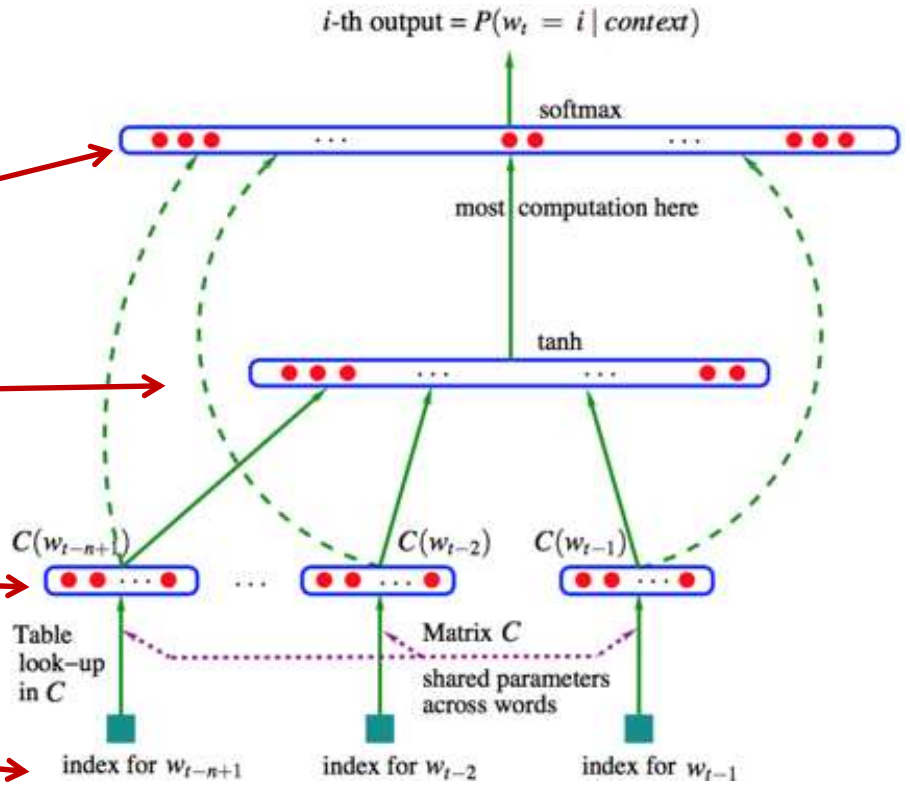
输出层：使目标词概率最大

隐含层

前n-1个词的分布式表示！

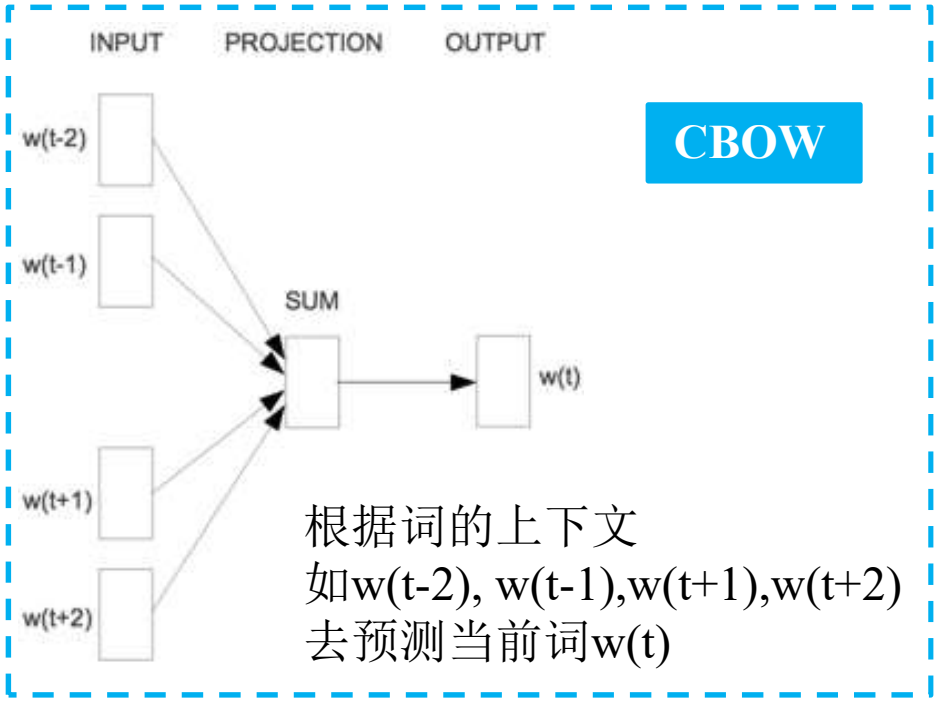
输入层：拼接n-1个向量

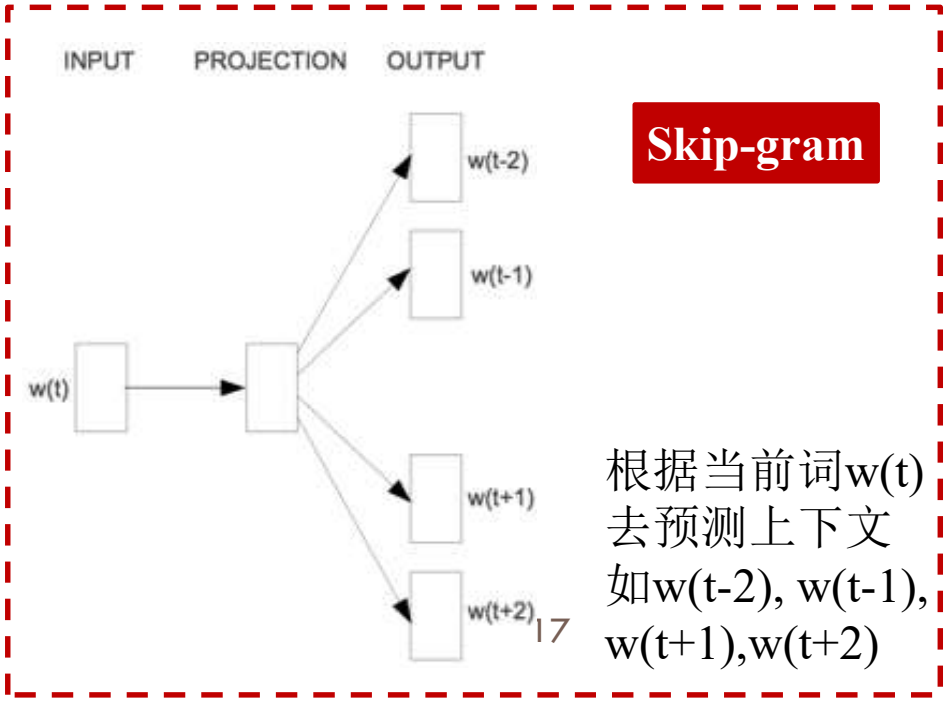
前n-1个词的索引



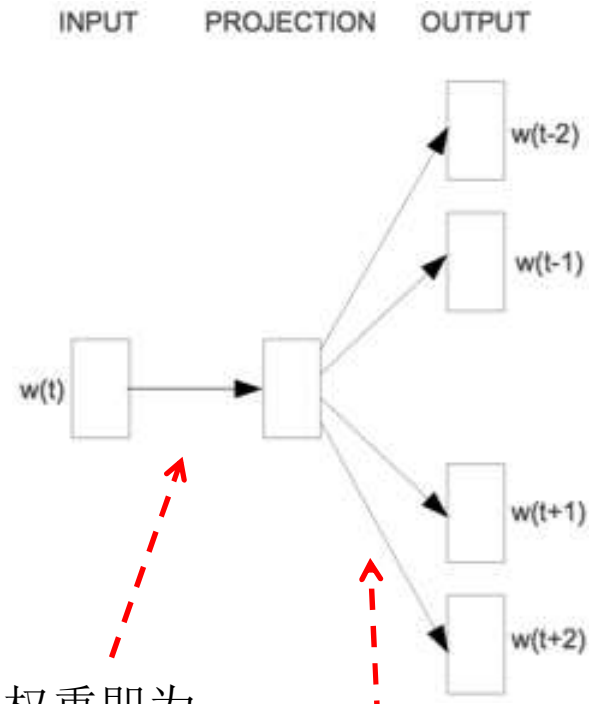
word2vec

- 神经网络语言模型的缺陷
  - 只能处理定长序列；后续有用RNN替代前向反馈神经网络进行解决
  - 主要问题：训练太慢！
- 改进：word2vec模型（及优化方法：层次softmax、负采样）





Skip-gram模型



权重即为词向量 e_t

权重即为辅助量 $\hat{e}_j$

最大化对数概率目标函数

$$L=\sum_{t=1}^{N}\sum_{t-c\le j\le t+c,j\neq t}\log p(w_j\mid w_t)$$

遍历中心词 w_t

遍历周围的词 w_j

$$p(w_j\mid w_t)=\frac{\exp(e_t^T\hat{e}_j)}{\sum_w \exp(e_t^T\hat{e}_w)}$$

softmax函数，用于多分类问题，
可求得属于各个类别的概率。
w_t预测周围词 w_j 时，要使其概率最大

模型求解：负采样

最大化对数概率目标函数

$$L=\sum_{t=1}^{N}\sum_{t-c\le j\le t+c,j\neq t}\log\frac{\exp(e_t^T\hat{e}_j)}{\sum_w \exp(e_t^T\hat{e}_w)}$$

梯度上升方法：求导softmax函数计算量大

解决方案：负采样（对softmax的一种近似）

$$L=\sum_{t=1}^{N}\sum_{t-c\le j\le t+c,j\neq t}\left(\log\sigma(e_t^T\hat{e}_j)+\sum_k E_{k\square P}\log\sigma(-e_t^T\hat{e}_k)\right)$$

仅选取少量负样本（如5、10）
而原式softmax中分母要遍历N次

word2vec结果

- 计算相似性(相似的词距离近)

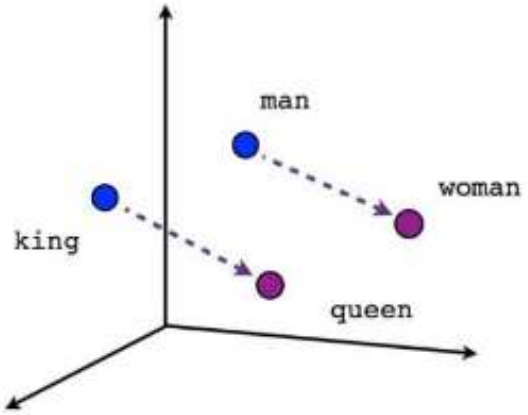
```
请输入词语<exit退出>:中国
大陆      0.66763467
中共      0.57856727
共产党    0.56305367
解放军    0.55761635
台湾      0.5368497
反攻      0.5271177
日本      0.5103535
王文莹    0.49295437
内地      0.48557448
对岸      0.48428434
```

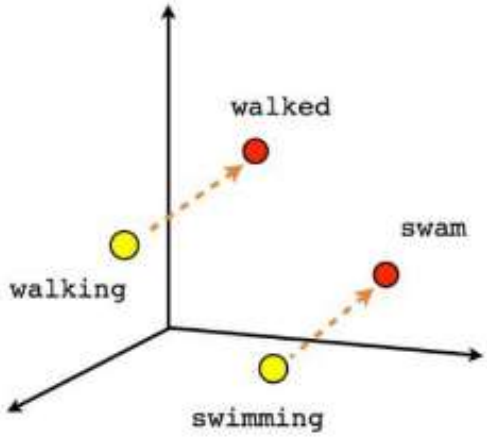
```
请输入词语<exit退出>:钓鱼岛
钓鱼台    0.6219264
钓岛      0.6123347
南海      0.6018163
领土      0.51753837
领海      0.4928774
岛屿      0.4853142
舰队      0.47854927
渔权      0.47229362
主权      0.46729872
东海      0.4613399
```

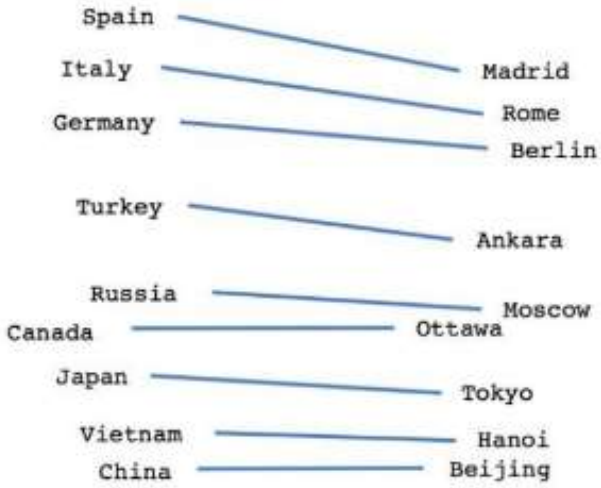
```
请输入词语<exit退出>:旅游
观光      0.65619475
景点      0.60212
陆客      0.59477097
旅行      0.5677106
游憩      0.557839
赏樱      0.5571045
游玩      0.52199984
观光客    0.51974636
行程      0.51943743
参观      0.5077874
```

word2vec结果

- 数学上的加减操作有意义



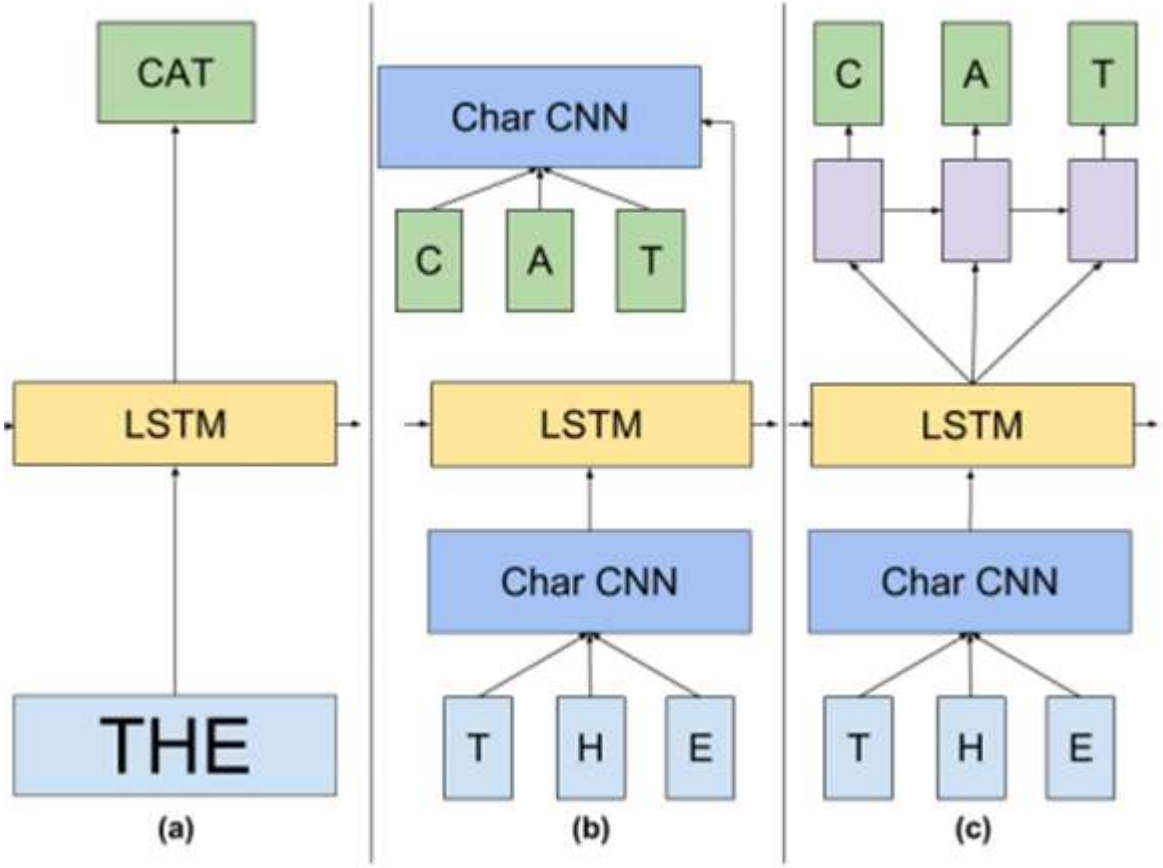




v("king") - v("queen") = v("man") - v("woman")

v("walking") - v("walked") = v("swimming") - v("swam")

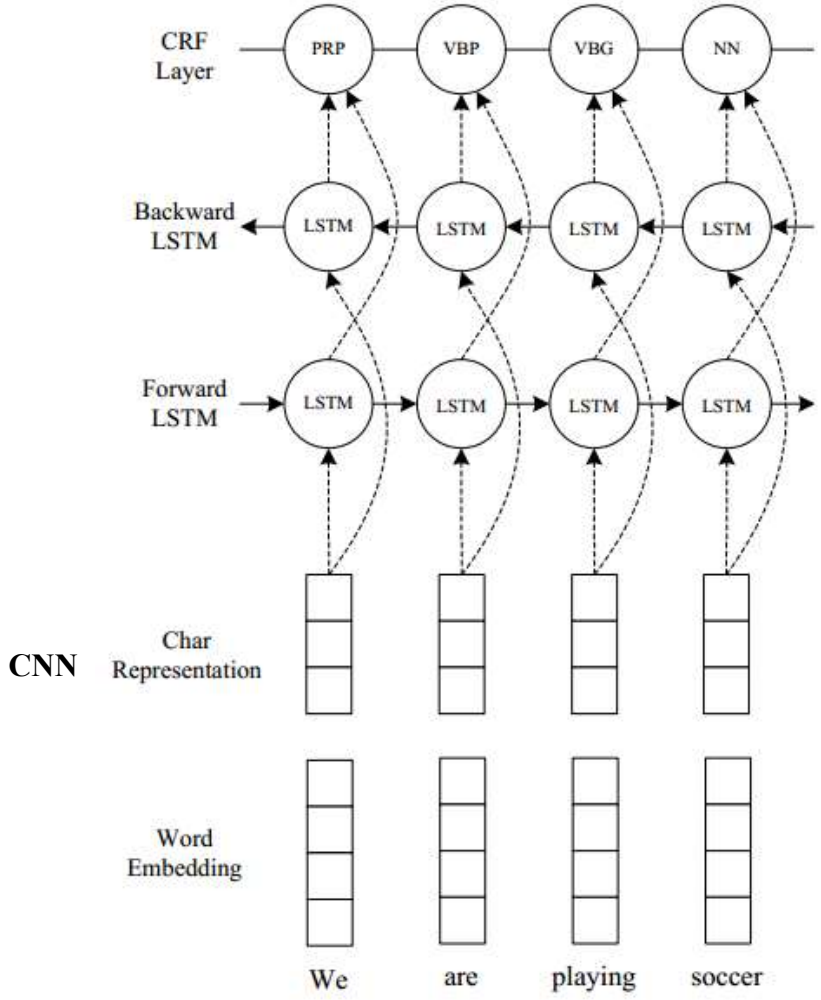
应用：字符级别的语言模型



字符级别可以有效应对低频词、新词的问题

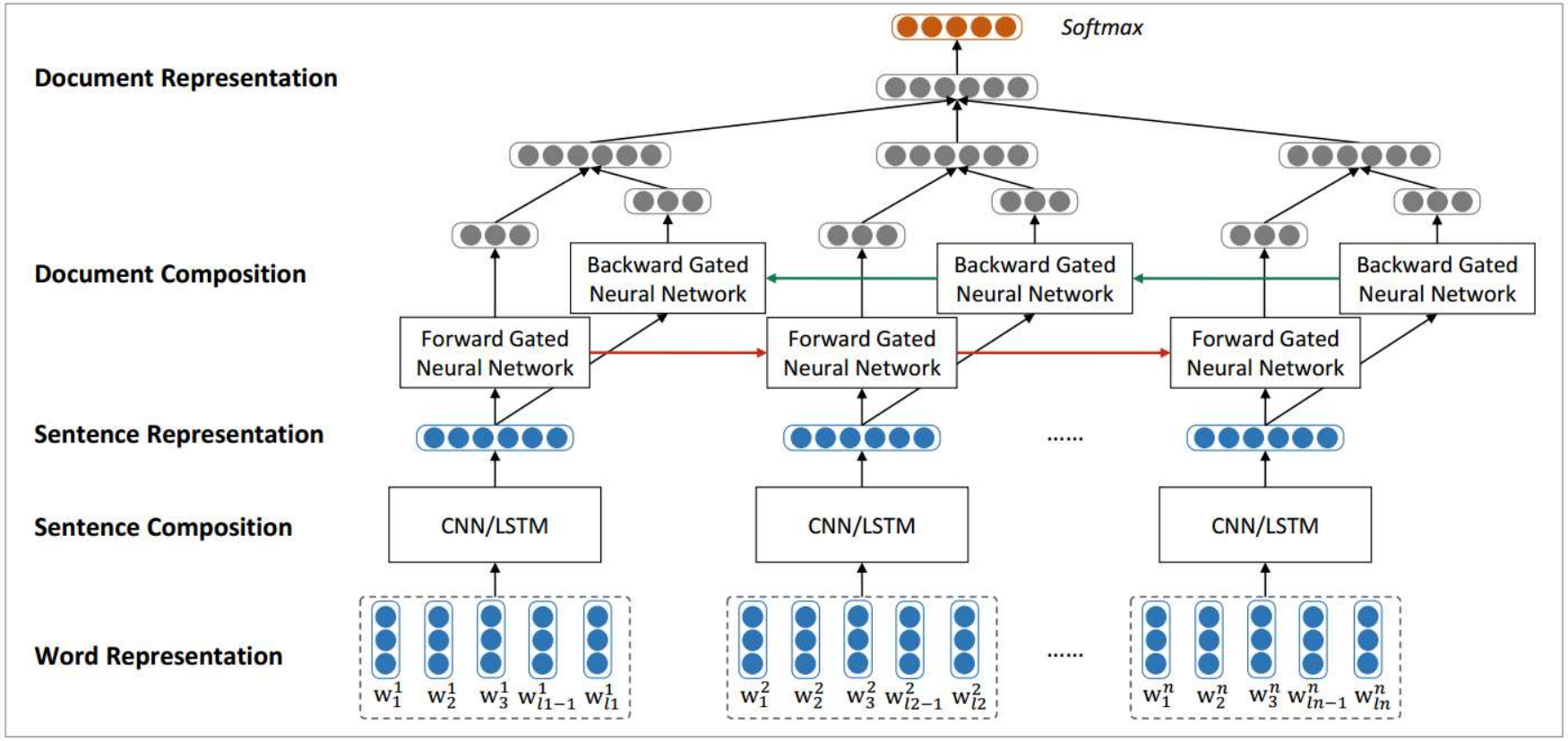
句子：The cat jumped over the lazy dog.

应用：序列标注


CRF Layer
PRP
VBP
VBG
NN
Backward LSTM
LSTM
LSTM
LSTM
LSTM
Forward LSTM
LSTM
LSTM
LSTM
LSTM
CNN
Char Representation
Word Embedding
We
are
playing
soccer

词性标注

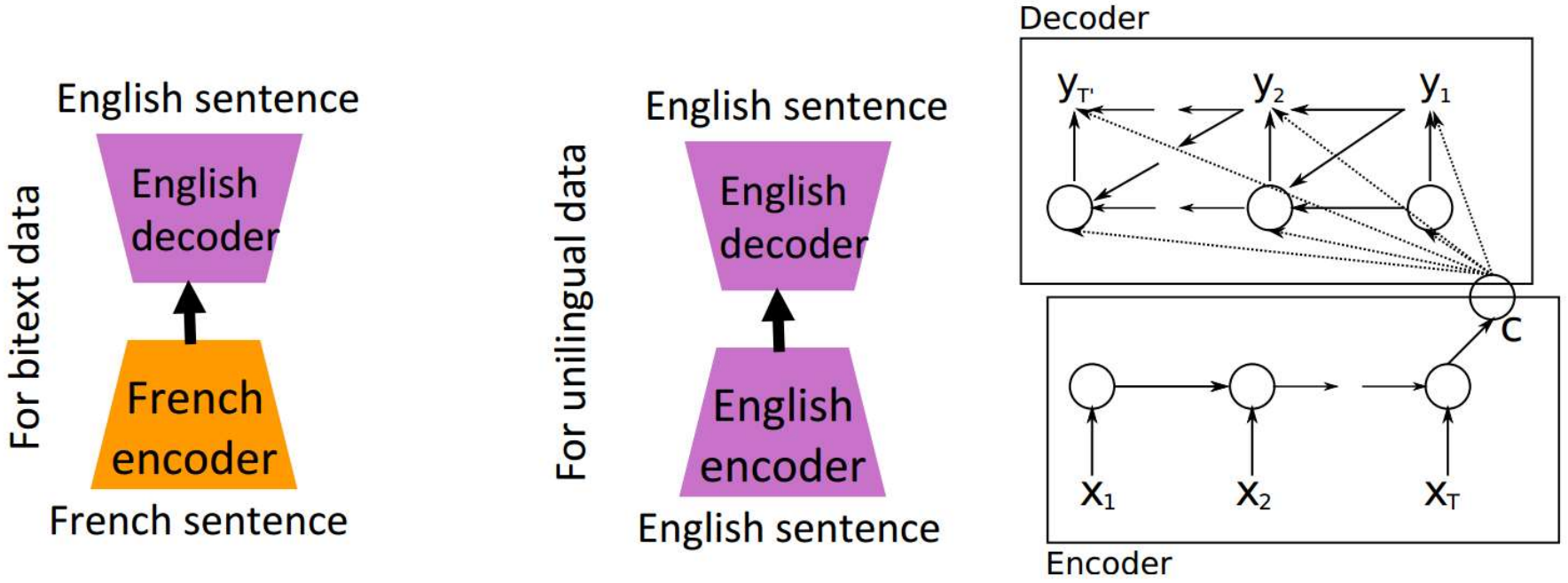
应用：文档分类



Tang et al. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. EMNLP 2015

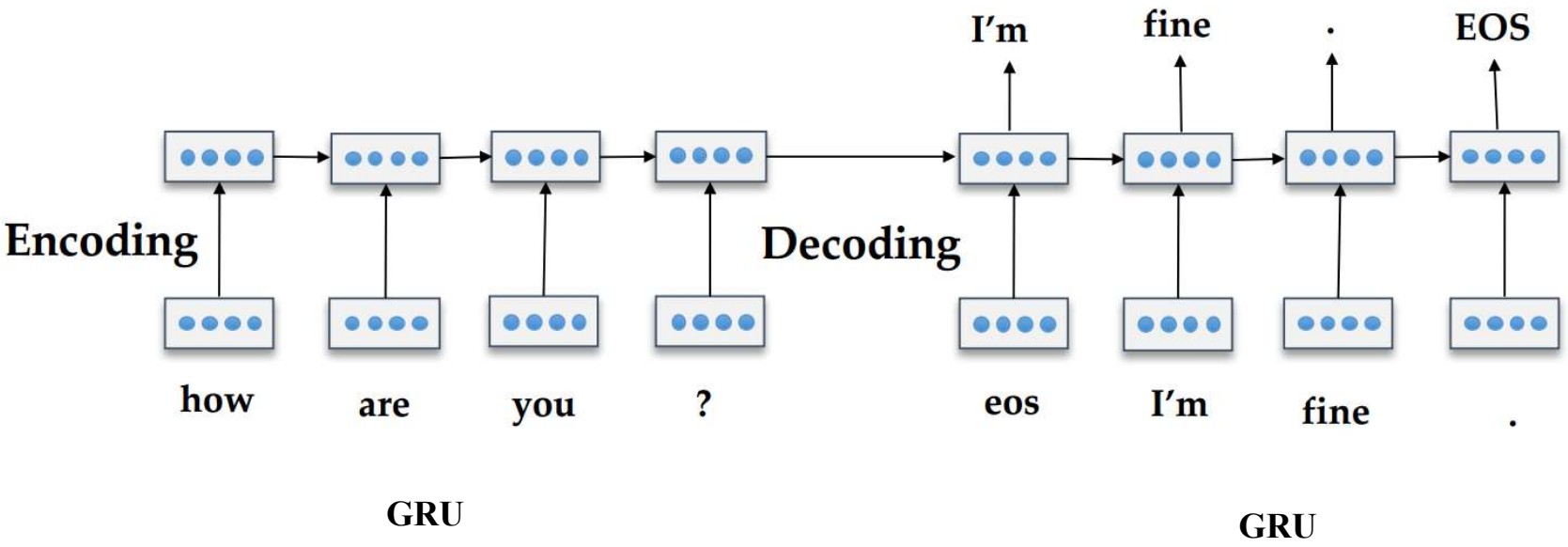
Encoder-Decoder框架

- Encoder: 将输入序列编码成向量表示
- Decoder: 将向量表示解码成输出序列的概率分布



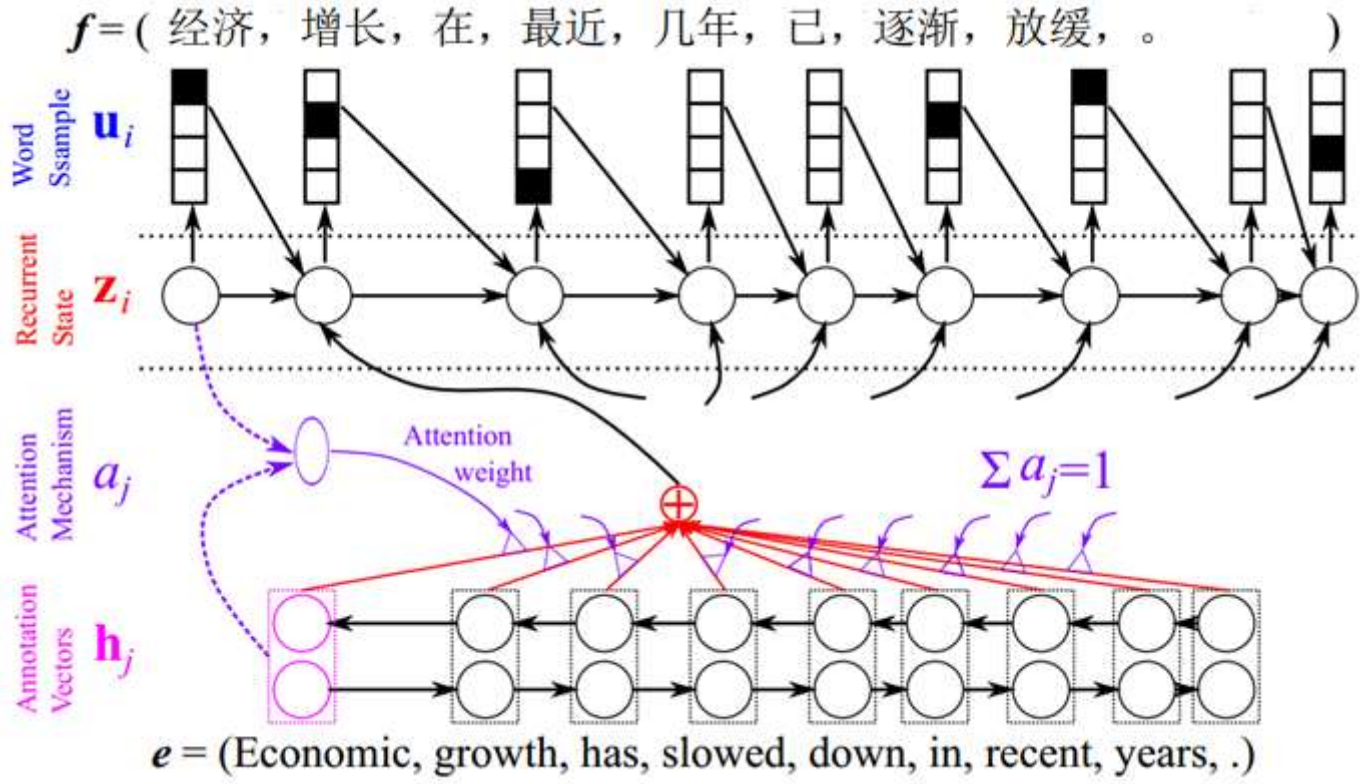
应用：对话系统

- 基于encoder-decoder框架的对话系统



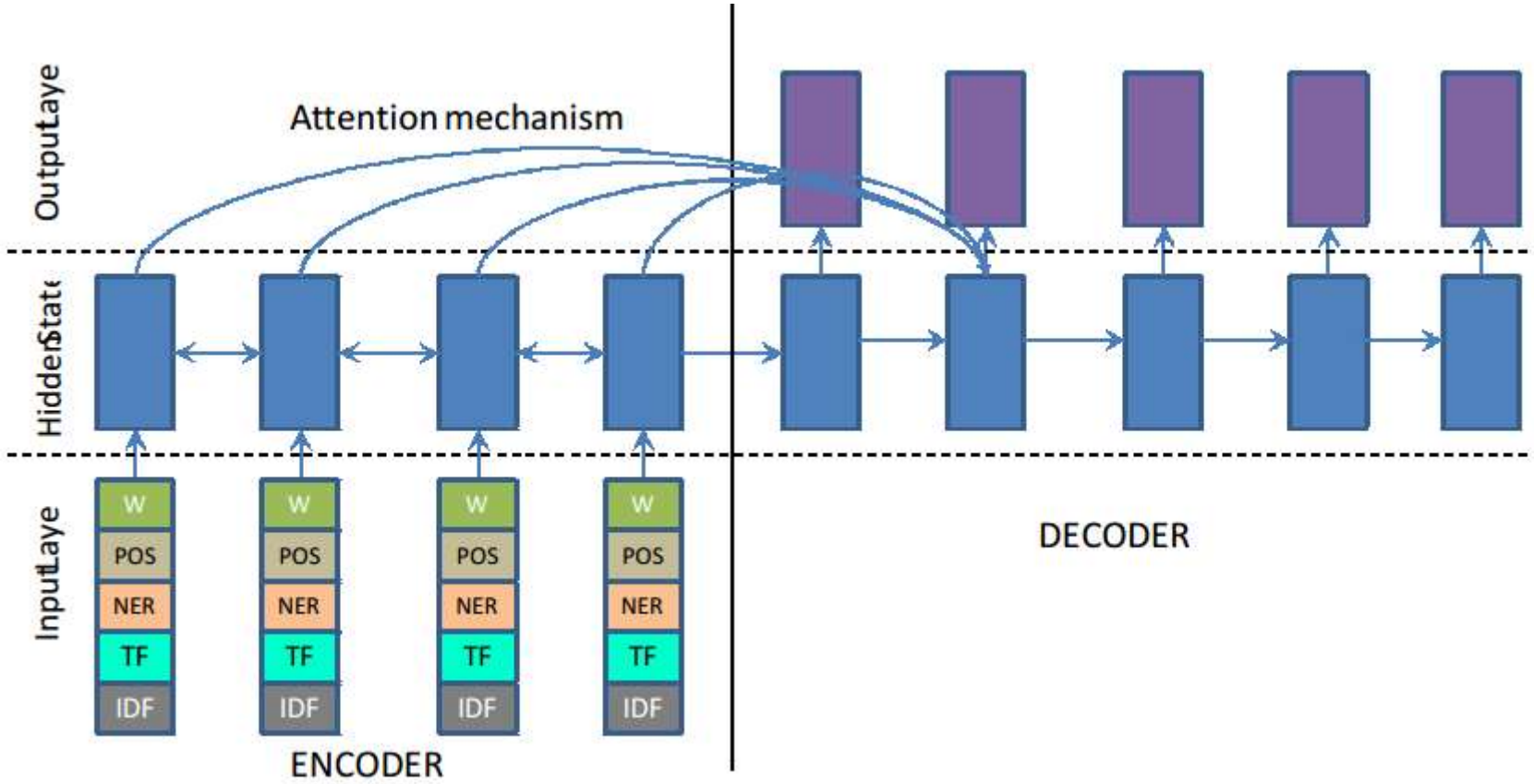
应用：机器翻译

- 基于Attention机制的机器翻译



应用：自动文本摘要

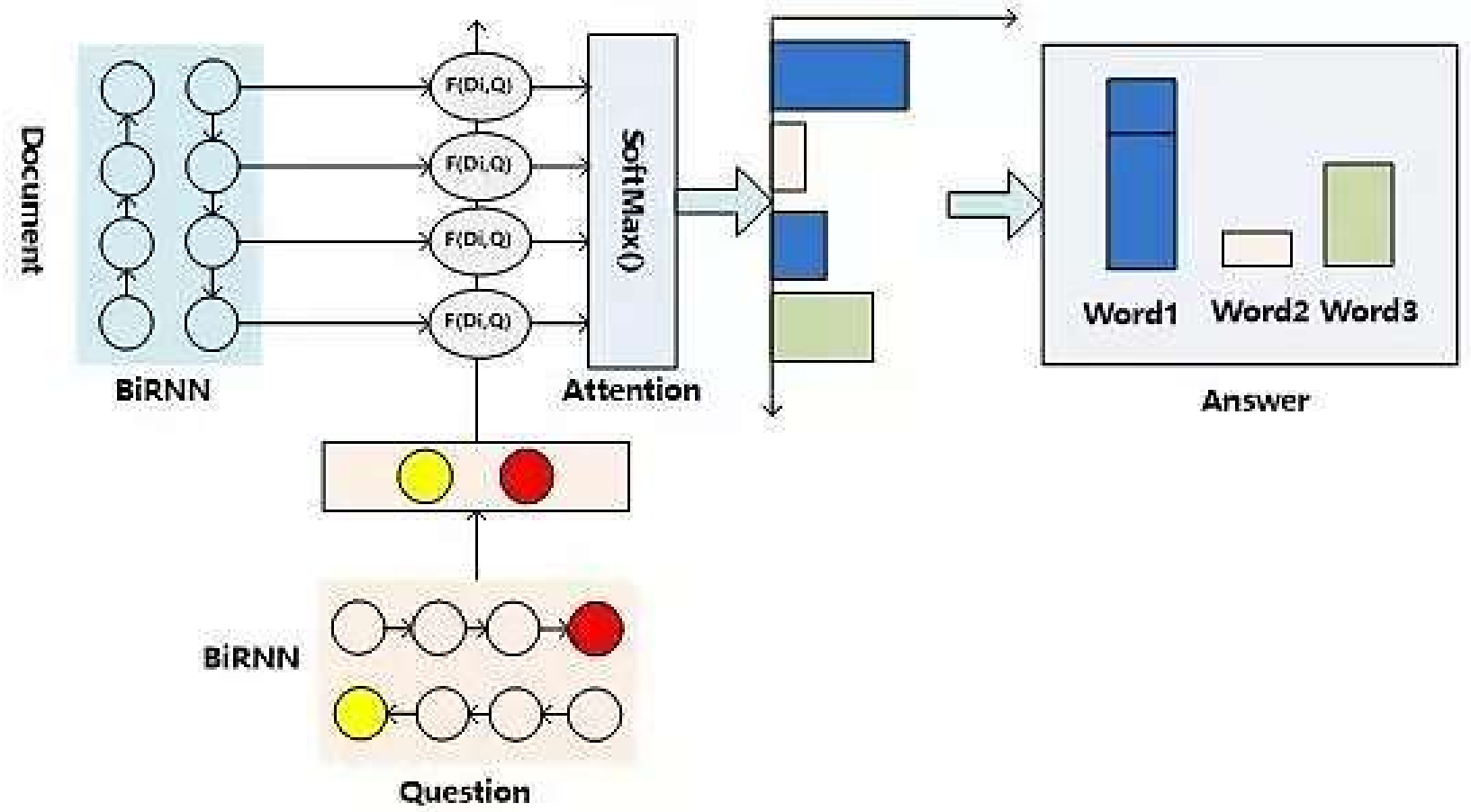
- 基于Attention机制的自动文本摘要



Rush et al. A Neural Attention Model for Abstractive Sentence Summarization. EMNLP 2015

应用：阅读理解

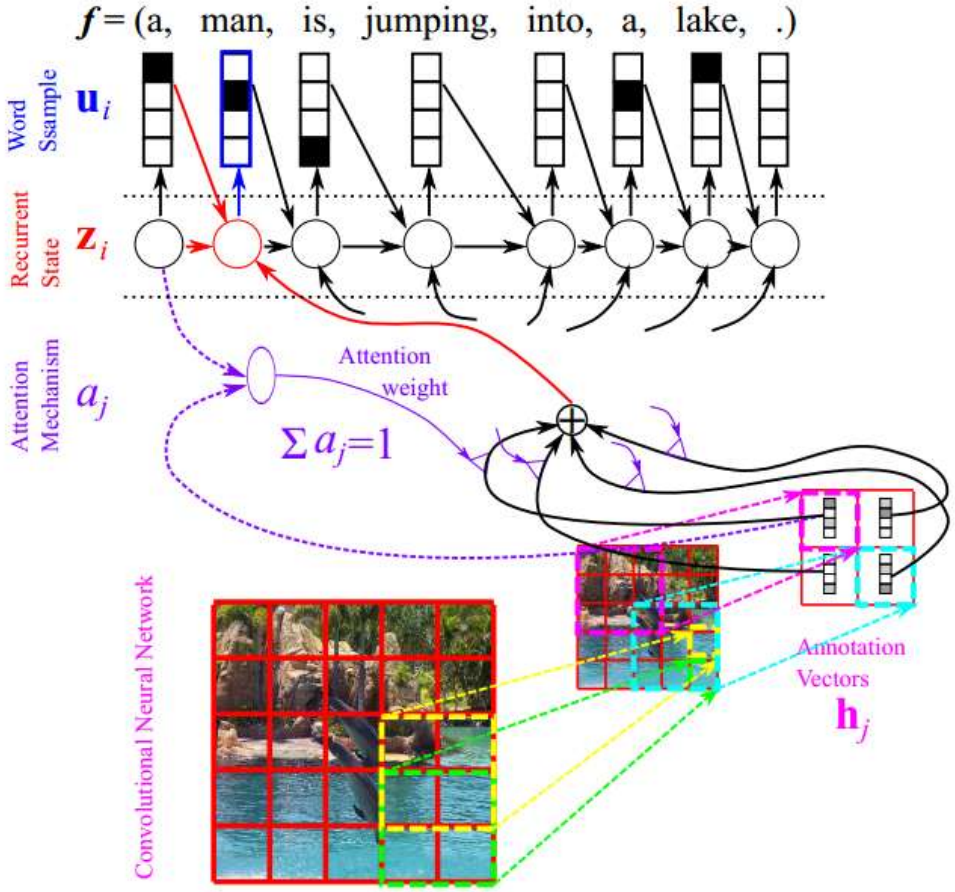
- 基于Attention机制的阅读理解



Kadlec et al. Text Understanding with the Attention Sum Reader Network. ACL 2016

应用：图片标题生成

- 基于Attention机制的图片标题生成



A stop sign is on a road with a mountain in the background.

A group of people sitting on a boat in the water.

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

我们的工作1：诗歌自动生成

- 诗歌是一种用于抒情言志，具有美感，适宜吟咏的文学形式。我国古代的唐诗，宋词等具有较为严格的诗词格律要求。
- 诗歌创作对人类是一个很难的问题，大部分普通人不具备诗歌创作能力

用户写作意图

江、船、秋风

春风、杨柳

松、竹、山、牧童

诗歌自动生成系统

对应诗词

江北江南万顷秋，
船头人去水悠悠。
一帆一棹秋风急，
又有离人万里愁。

杨柳千条拂地垂，
一川春水浸桃花。
游人不识湖中路，
游遍人间野水涯。

乔松古木两三间，
松竹阴中一径斜。
白鸟不知山路远，
牧童踏过野人家。

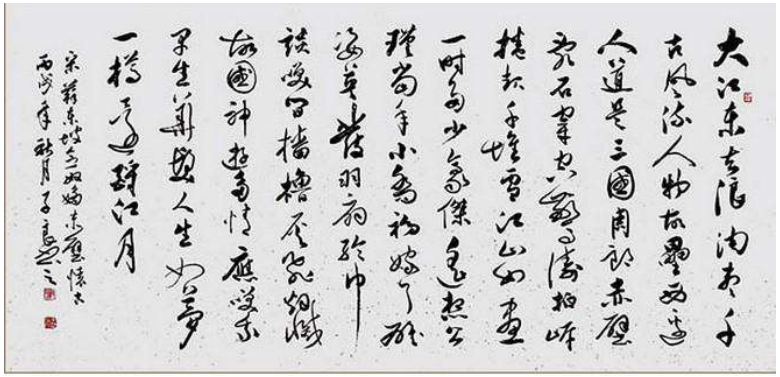
我们的工作1：诗歌自动生成

- 诗歌自动生成的挑战：
  - 诗歌**体裁**（如唐诗，宋词）众多，不同体裁的诗歌有不同的**格律**形式要求（如平仄，押韵），需要大量专业知识。
  - 诗歌对**逻辑性**，**流畅性**，**意境**等有较高的标准要求，而这些标准难以人工量化，给自动生成带来很大困难。

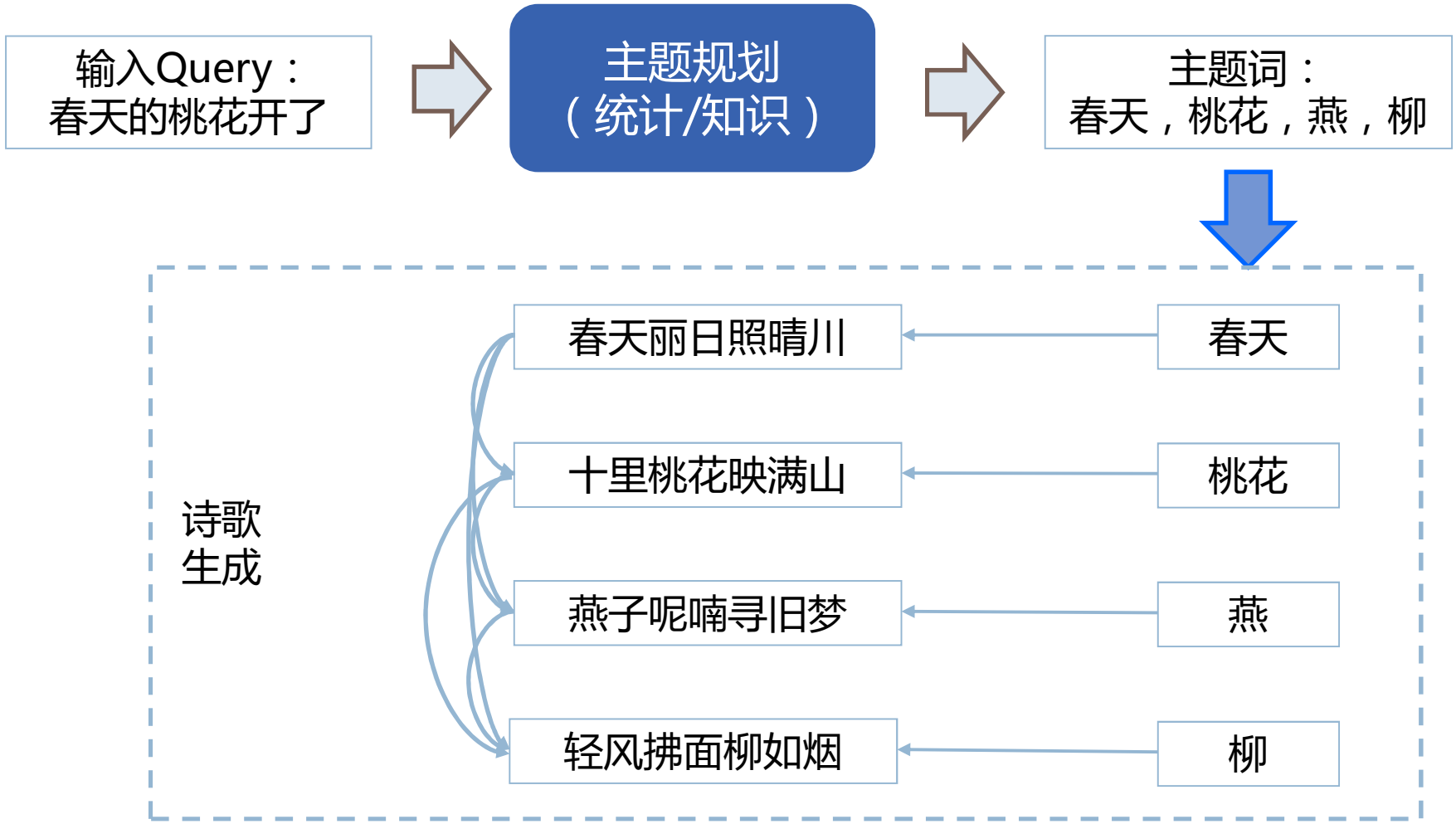
体裁格式众多

兩個黃鸝鳴翠柳，
一行白鷺上青天。
窗含西嶺千秋雪，
門泊東吳萬裏船

格律形式规整，专业性很强

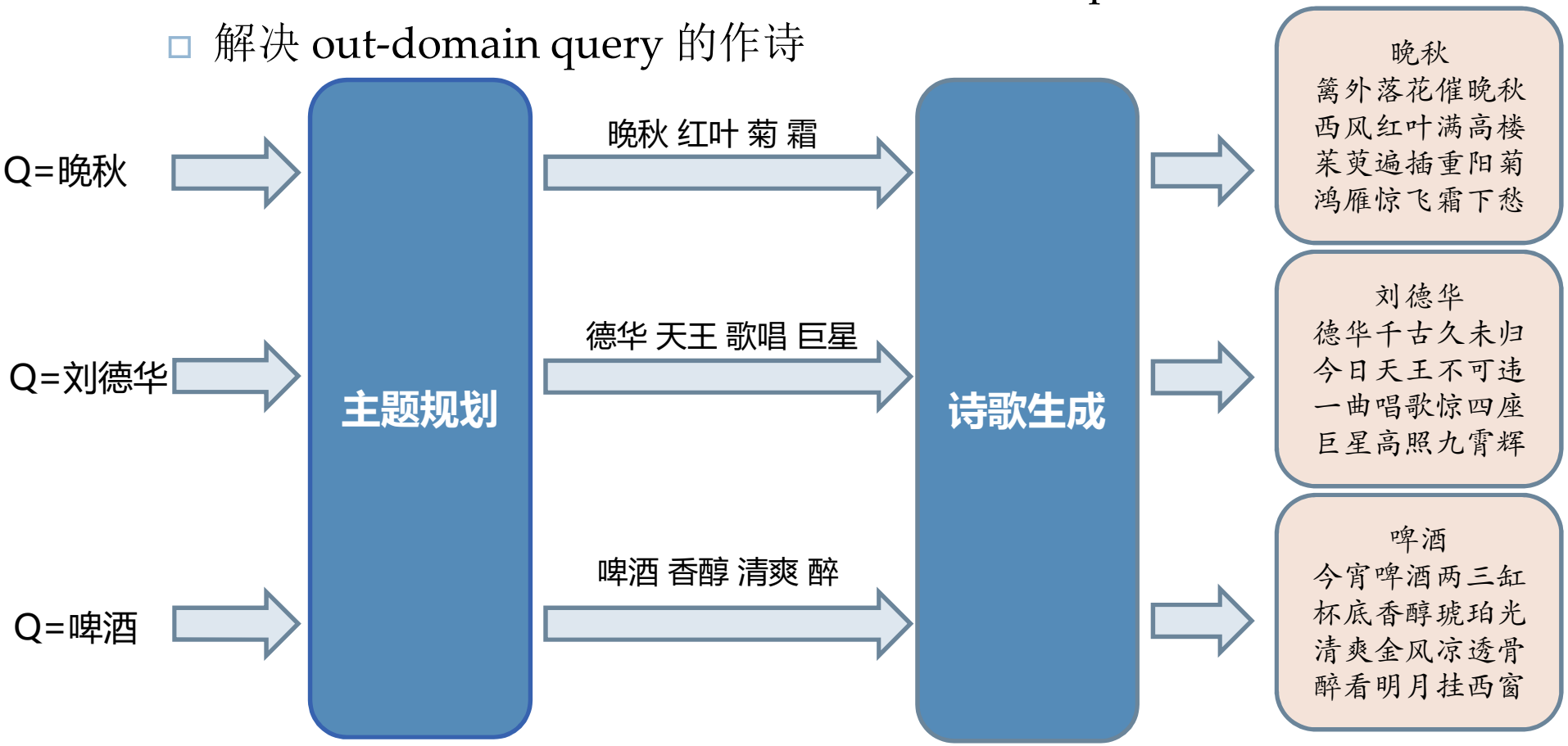


基于规划的诗歌生成


输入Query：
春天的桃花开了
主题规划
（统计/知识）
主题词：
春天，桃花，燕，柳
诗歌
生成
春天丽日照晴川
十里桃花映满山
燕子呢喃寻旧梦
轻风拂面柳如烟
春天
桃花
燕
柳

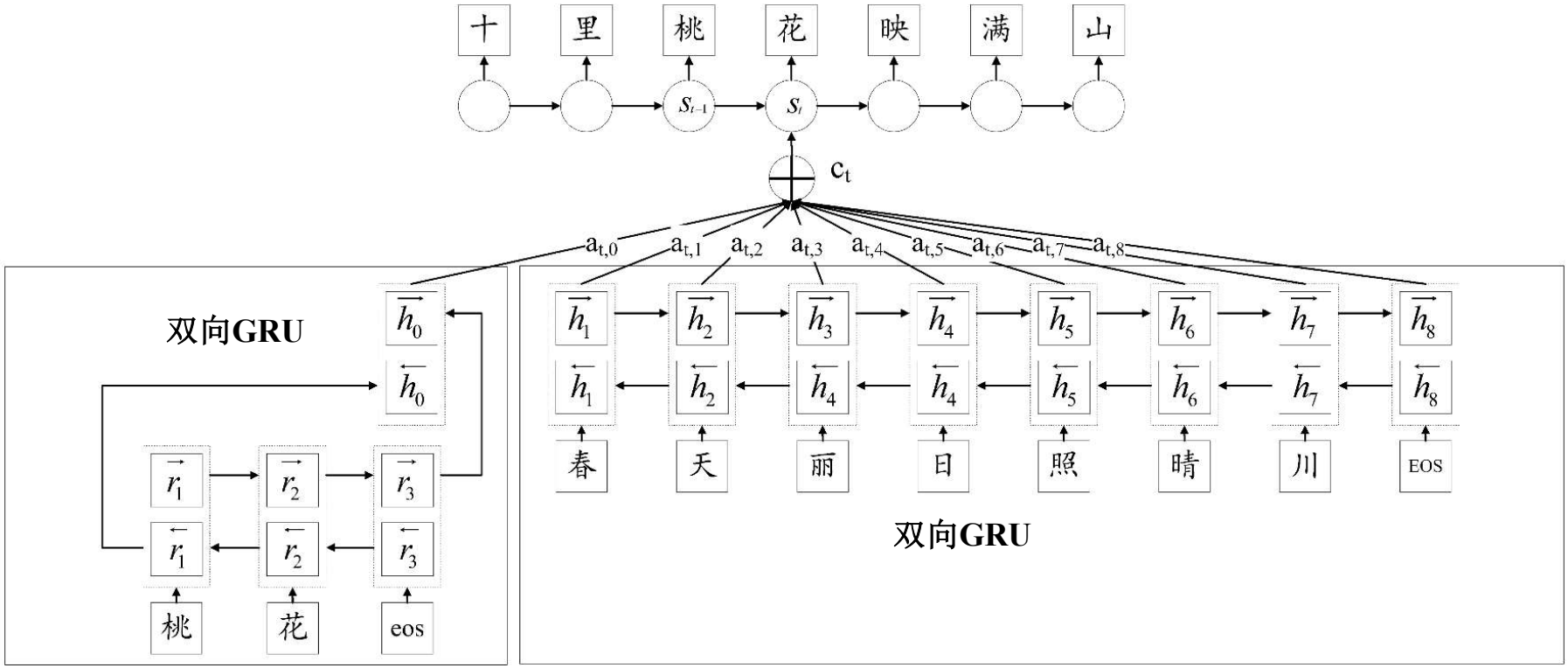
主题规划模型

- 在主题规划阶段引入不同来源的知识
  - 诗词语料，知识图谱，百科，搜索推荐，人工pattern
  - 解决 out-domain query 的作诗



诗歌生成模型

- 基于Attention的句子生成模型



人工评估结果

- 评估标准

Poeticness	Does the poem follow the rhyme and tone requirements ?
Fluency	Does the poem read smoothly and fluently?
Coherence	Is the poem coherent across lines?
Meaning	Does the poem have a certain meaning and artistic conception?

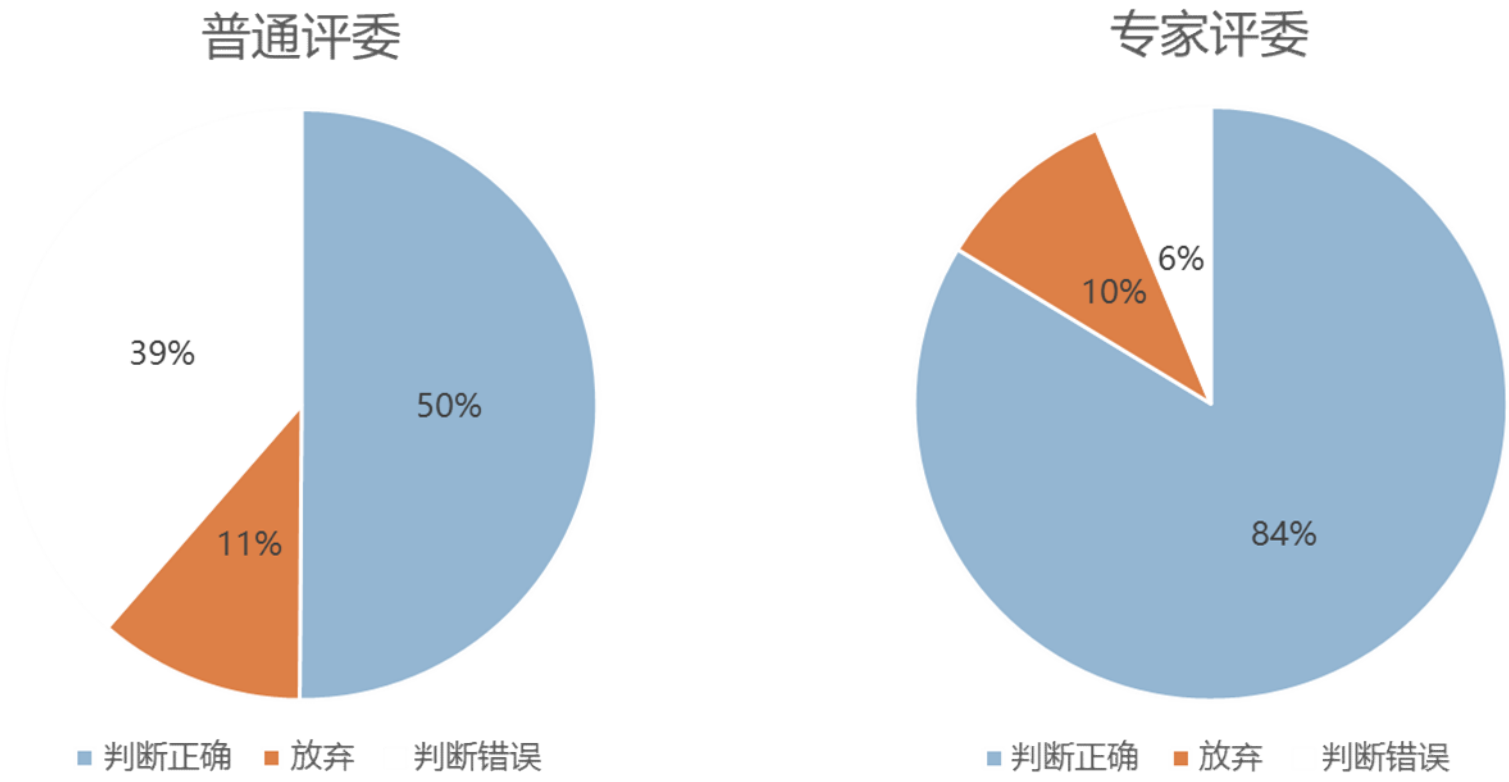
- 实验结果

Models	Poeticness		Fluency		Coherence		Meaning		Average	
	5-char	7-char	5-char	7-char	5-char	7-char	5-char	7-char	5-char	7-char
SMT	3.25	3.22	2.81	2.48	3.01	3.16	2.78	2.45	2.96	2.83
RNNLM	2.67	2.55	3.13	3.42	3.21	3.44	2.90	3.08	2.98	3.12
RNNPG	3.85	3.52	3.61	3.02	3.43	3.25	3.22	2.68	3.53	3.12
ANMT	**4.34**	4.04	**4.61**	4.45	4.05	4.01	4.09	4.04	4.27	4.14
PPG	4.11	**4.15**	4.58	**4.56***	**4.29***	**4.49****	**4.46****	**4.51****	**4.36****	**4.43****

由5个专家分别对每个模型生成的20首诗5言诗和7言诗进行了评估，可以看到我们的PPG模型在绝大部分指标上都超过了其他模型。

机器作诗 PK 古代诗人

- 写诗图灵测试
  - 随机选20首古诗，让机器用相同题目作诗，组成20个 pair
  - 让用户判断哪一首是真正的古诗



机器诗人 PK 古代诗人

哪首诗是人写的?

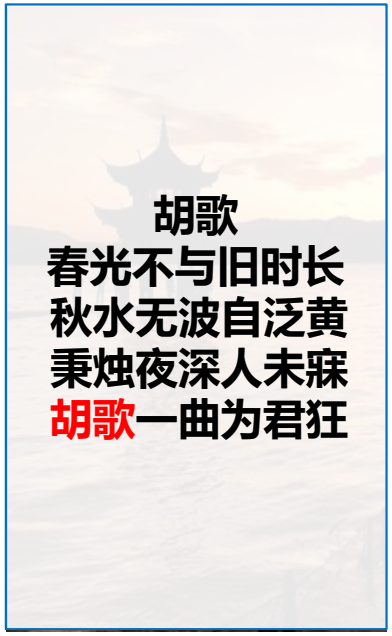
秋夕湖上	秋夕湖上
一夜秋凉雨湿衣， 西窗独坐对夕晖。 湖波荡漾千山色， 山鸟徘徊万籁微。	荻花风里桂花浮， 恨竹生云翠欲流。 谁拂半湖新镜面， 飞来烟雨暮天愁。

基于主题规划衍生出的多种产品形式

语音作诗

Q=西湖烟雨

藏名诗



Q=胡歌

藏头诗

Q=为你写诗

看图作诗

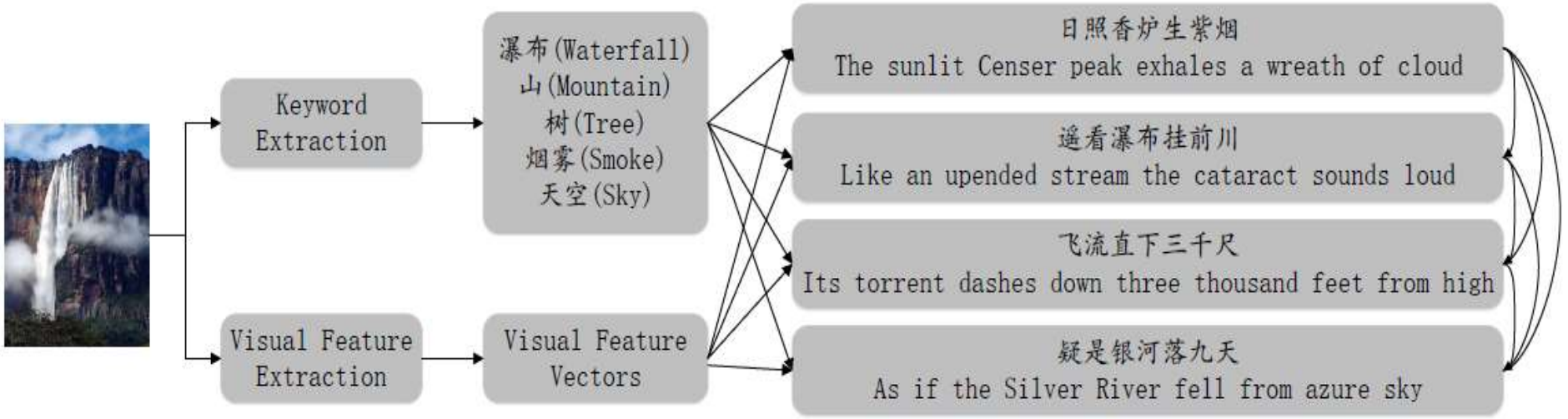
输入：图片

风格作诗

风格：爱情诗

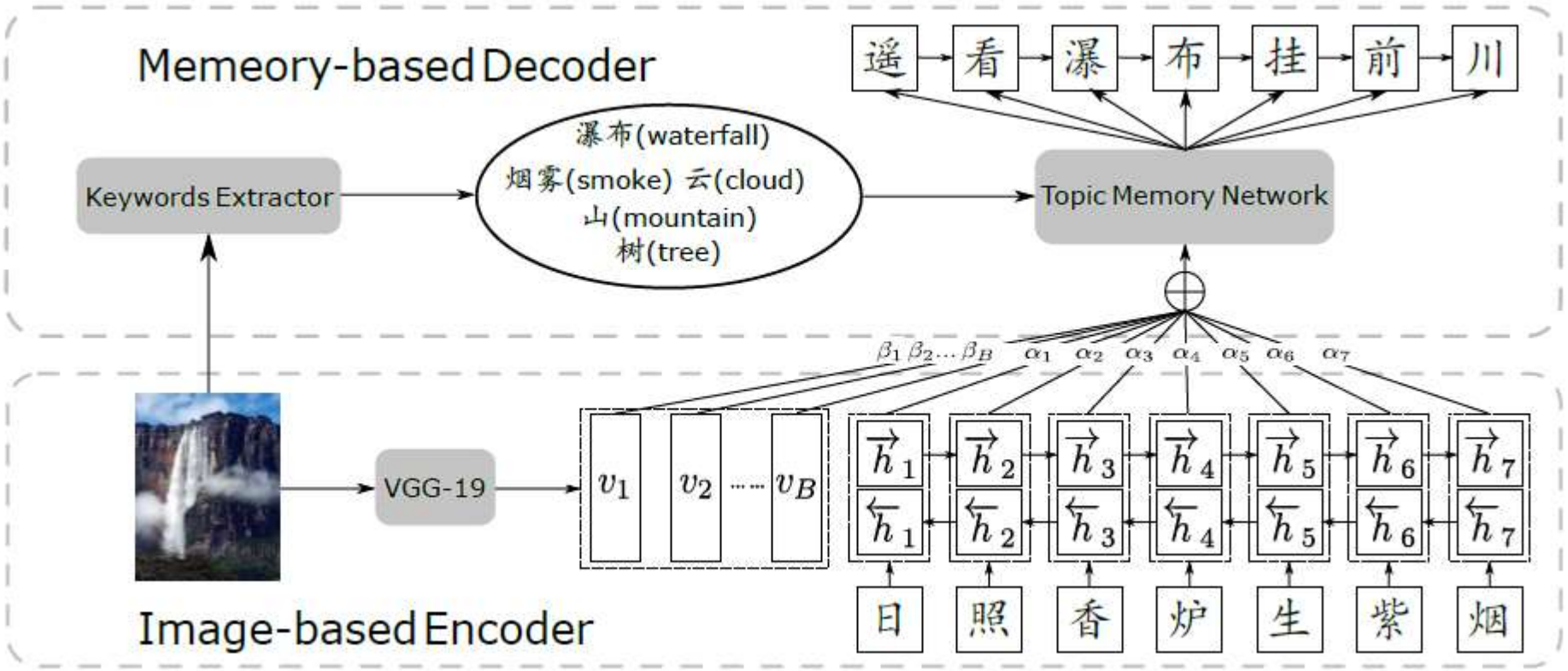
改进的工作：看图写诗

- 看图写诗流程
  - 从图片中提取关键词和视觉特征
  - 通过深度学习方法生成诗



改进的工作：看图写诗

- 基于Encoder-Decoder的看图写诗框架



看图写诗效果

扁舟一曲水平堤，
I sing a fishing song on a boat in the lake overflowing its bank,
一棹渔舟日向西。
rowing oars with the sun setting in the west.
长忆西湖水中月，
I often miss the moon reflected in the West Lake,
东风吹过武陵溪。
the east breeze blew across the WuLing river.

春风庭院养花姿，
Breeze blows beautiful flowers in the courtyard,
春入帘栊叶满枝。
Spring comes into my window, and leaves cover the branches.
堪笑门前青草树，
Glad to see green grass and trees in front of my door,
谁家芳节几多时。
However spring will not last very long.

我们的工作2：词汇蕴含识别

- 词汇蕴含识别是文本蕴含识别的重要组成部分
  - 前提句：小明被一只狗咬了
  - 假设句：小明被一只动物攻击了
  - 如果我们可以知道：
    - “狗”蕴涵“动物”
    - “咬”蕴涵“攻击”
  - 就可以从前提句推断出假设句

- 实际应用中的例子：
  - You should take umbrella when you go out because it’s __ outside

    A: rainy　B: sunny　C : cloudy　D : overcast

我们的工作2：词汇蕴含识别

词汇蕴含的定义：

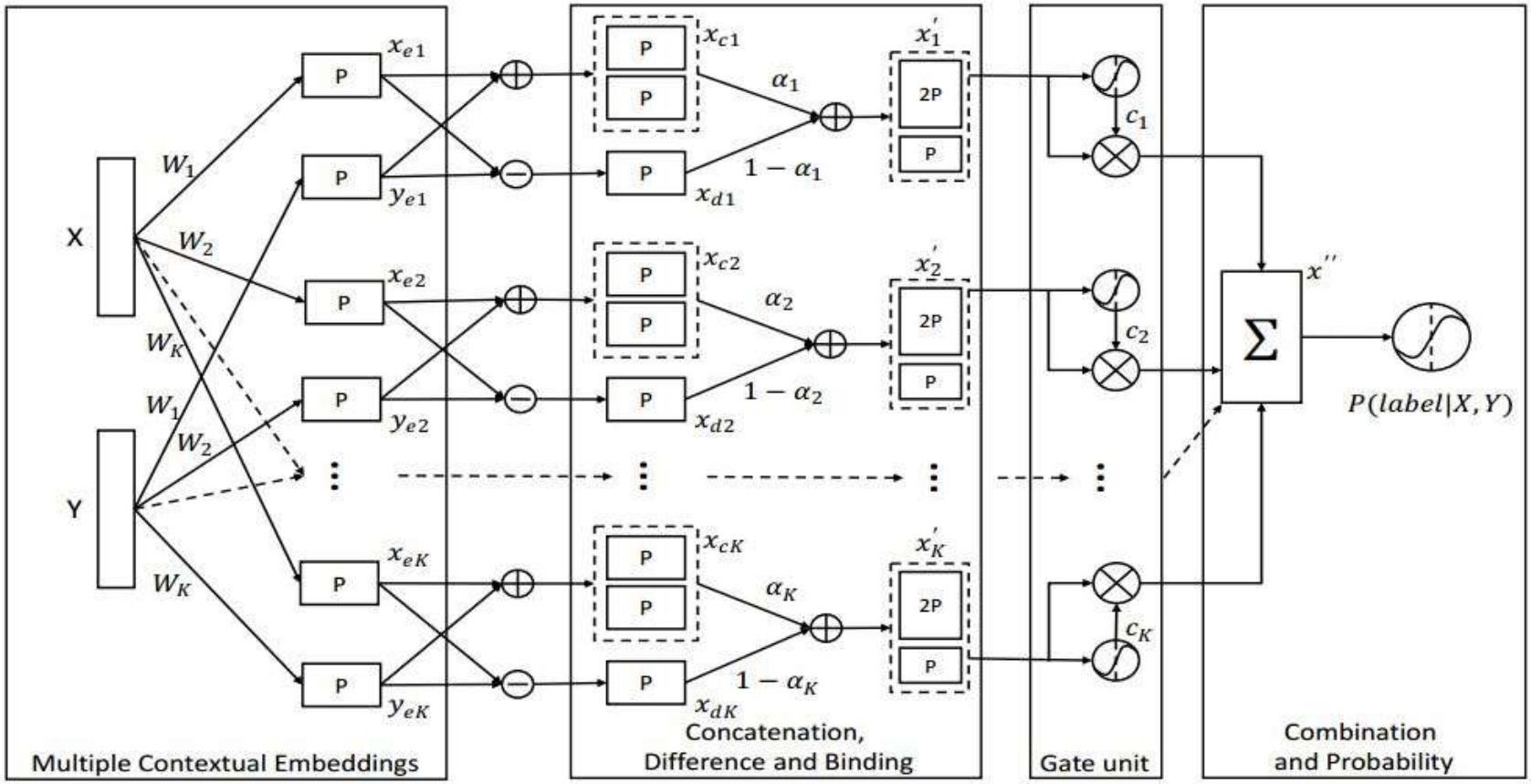
- 词义的包含性
  - （狗，动物）、（果实，成果）
- 在特定条件下词对之间的可替换性
  - 要有辛勤耕耘,才会收获丰硕的果实。
  - 要有辛勤耕耘,才会有丰硕的成果。

  - 秋天，山野里散发出果实的芳香
  - 秋天，山野里散发出成果的芳香

我们的工作2：词汇蕴含识别

- 当前的挑战
  - 词义的多样性
    - 门槛、算账、包袱、果实
    - book、cook、change、bear
  - 蕴含关系的多样性
    - 因果关系：流感 ⟹ 发烧
    - 上下位关系：狗 ⟹ 动物
  - 组合方式的多样性
    - 不同的组合方式揭示词的不同的信息

Context-Enriched Neural Network

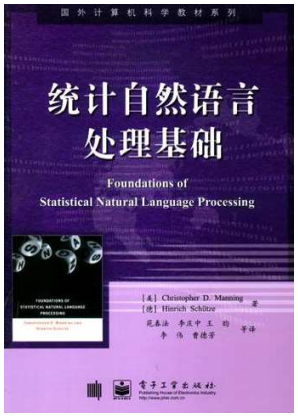


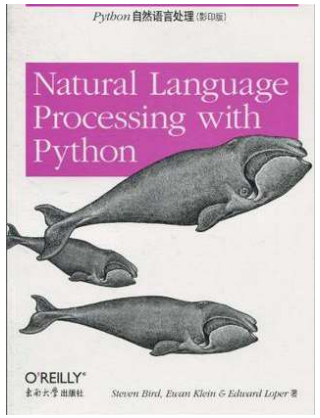
CENN的模型框架图

资料推荐

- 论文
  - Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.
  - Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.
  - Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]// Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2004:487-494.
  - LDA数学八卦
- 会议&书籍
  - ACL、COLING、EMNLP

ACL COLING emnlp





目录

- 传统的文本处理
  - 自然语言处理
  - N-Gram语言模型
- 文本挖掘
  - 文本挖掘简介
  - 特征抽取
  - 特征选择
  - 文本分类
  - 文本聚类
  - 文本挖掘的常用方法
    - 主题模型
    - word2vec