自然语言处理中的语义表示学习

张家俊
模式识别国家重点实验室
中国科学院自动化研究所
2015.10.31

自然语言处理-搜索引擎

Bai du 百度 中国海洋大学 百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约591,000个 搜索工具

中国海洋大学
Ocean University of China
www.ouc.edu.cn/ V1 - 百度快照 - 92%好评

中国海洋大学_百度百科

中国海洋大学（Ocean University of China），简称海大或中国海大（OUC），原名私立青岛大学，位于山东省青岛市，始建于1924年，2002年更名为中国海洋大学。学校是教育部直属重点综合性大学，是国家"985工程"和"211工程"重点建设高校之...
历史沿革 办学条件 学术研究 文化传统 学校领导 更多>>
baike.baidu.com/

中国海洋大学高考分数线_招生信息_百度教育

办学类型： 985高校 211高校 教育部直属
院校类型： 综合类
高校地址： 山东省青岛市崂山区松岭路238号 校园全景
相关信息： 官方网站 校园风光 招生章程 高校校花 专业pk

选择生源地： 山东 选择科属： 理科

年份	最高分	平均分	省控线	录取批次
2014	673	652	572	本科一批
2013	662	631	554	本科一批
2012	668	645	582	本科一批

查看更多中国海洋大学信息>>

jiaoyu.baidu.com 2015-09-22 -

山东省高等院校 展开

中国石油大学 一所石油高等学府
青岛大学 重点综合性大学
山东科技大学 全国大学生满意度50强
济南大学 四所部属本科高校

知名校友 展开


姚劲波 创建易域网论坛


宋祖德 娱乐圈的纪委书记

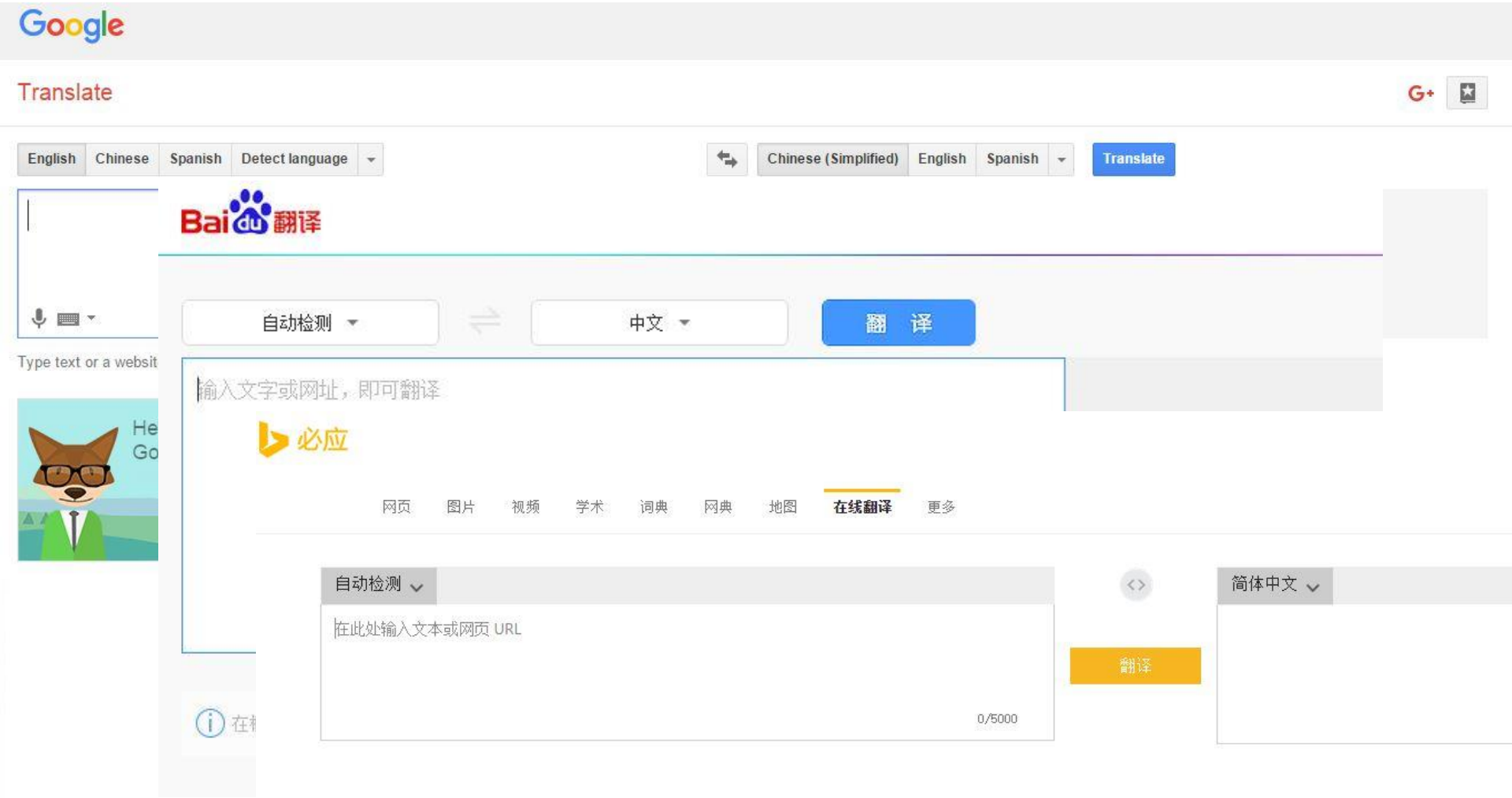

王宏 国家海洋局局长


张志峰 中海硕士生导师

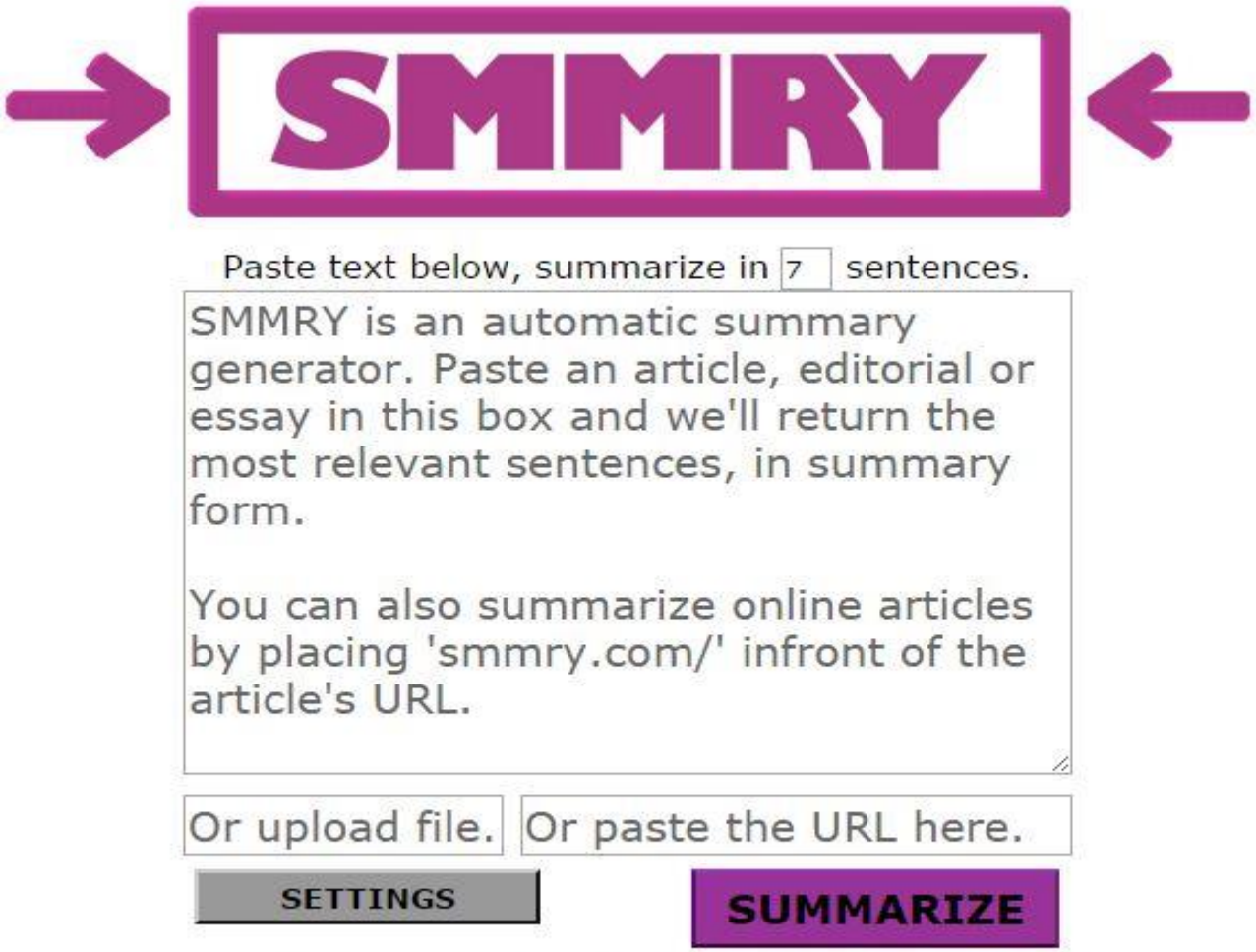
相关重点大学 展开

哈尔滨工业大学（威... 哈工业三个跨省校区
中国地质大学 中国地质大学前身
烟台大学 守信求实 好学力行
青岛科技大学 明德笃学 弘毅拓新

自然语言处理-机器翻译


Google
Translate
English
Chinese
Spanish
Detect language
Chinese (Simplified)
English
Spanish
Translate
Type text or a websit
Baidu 翻译
自动检测
中文
翻 译
输入文字或网址，即可翻译
必应
网页
图片
视频
学术
词典
网典
地图
在线翻译
更多
自动检测
在此处输入文本或网页 URL
0/5000
翻译
简体中文

自然语言处理-自动摘要


SMMRY
Paste text below, summarize in 7 sentences.
SMMRY is an automatic summary generator. Paste an article, editorial or essay in this box and we'll return the most relevant sentences, in summary form.
You can also summarize online articles by placing 'smmry.com/' infront of the article's URL.
Or upload file.
Or paste the URL here.
SETTINGS
SUMMARIZE
HOME | ABOUT | AUTO | API | CONTACT | PARTNER | REGISTER | LOGIN
© 2015 Smmry.com

自然语言处理-问答聊天


中国移动 下午9:20 46%
你可以这样问我：
从上海到苏州怎么走?
安排9点钟开会
日经指数如何?
查找关于北极光的视频
这个星期的天气怎么样?
提醒我打电话给妈妈
中国移动 下午9:16
微信
小冰
今天降温
很冷怎么办?
看笑话笑暖和了就好了
这么久才升一级😖说好的一飞冲天呢😂掌握一门技术，「冰PK」我不是一个好斗的人，但是我斗起来不是人！😝
明天会降温吗?
十来度吧 下点小雨
要穿什么衣服?
带个长点的羽绒服来吧
中国移动 下午9:20 46%
微信
小度机器人
明天天气怎样?
你想要查询哪儿的天气?
北京的
北京那边的天气啊，小度知道~北京周一21°C～12°C，晴转多云，微风。建议着薄外套、开衫牛仔衫裤等服装。年老体弱者应适当添加衣物，宜着夹克衫、薄毛衣等。
http://m.baidu.com/s?word=北京天气
中国海洋大学在哪
我知道:山东省青岛市市南区鱼山路5号

研究任务-词法分析

从青岛站到海大怎么走？

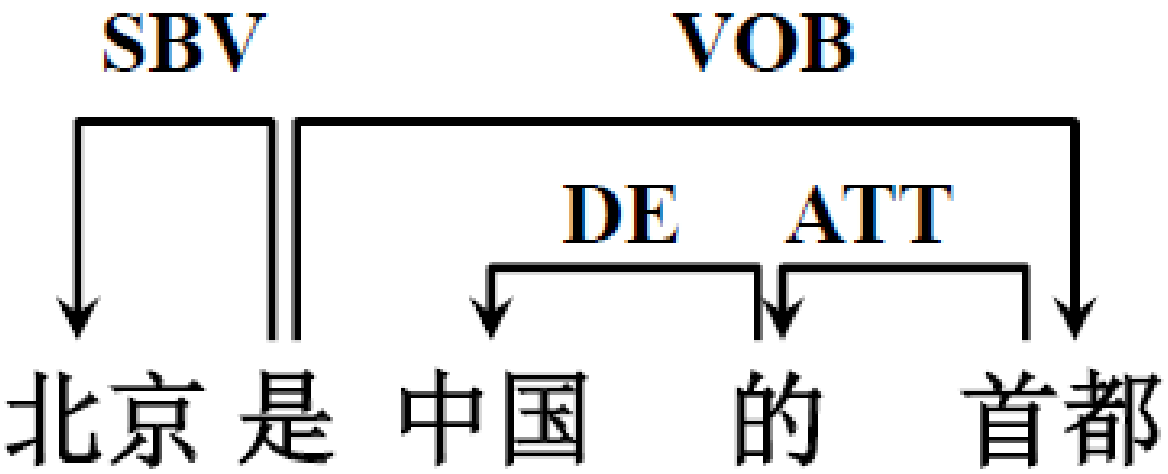
从 青岛站 到 海大 怎么 走 ？

从 **青岛站** 到 **海大** 怎么 走 ？

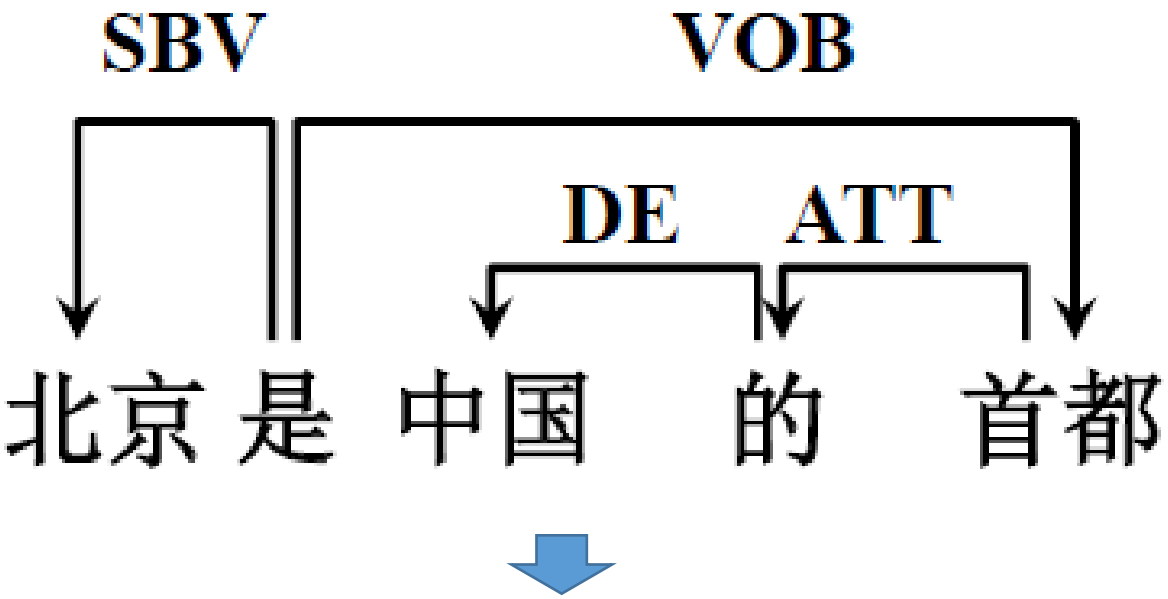
地名　　机构名

研究任务

研究任务-句法分析



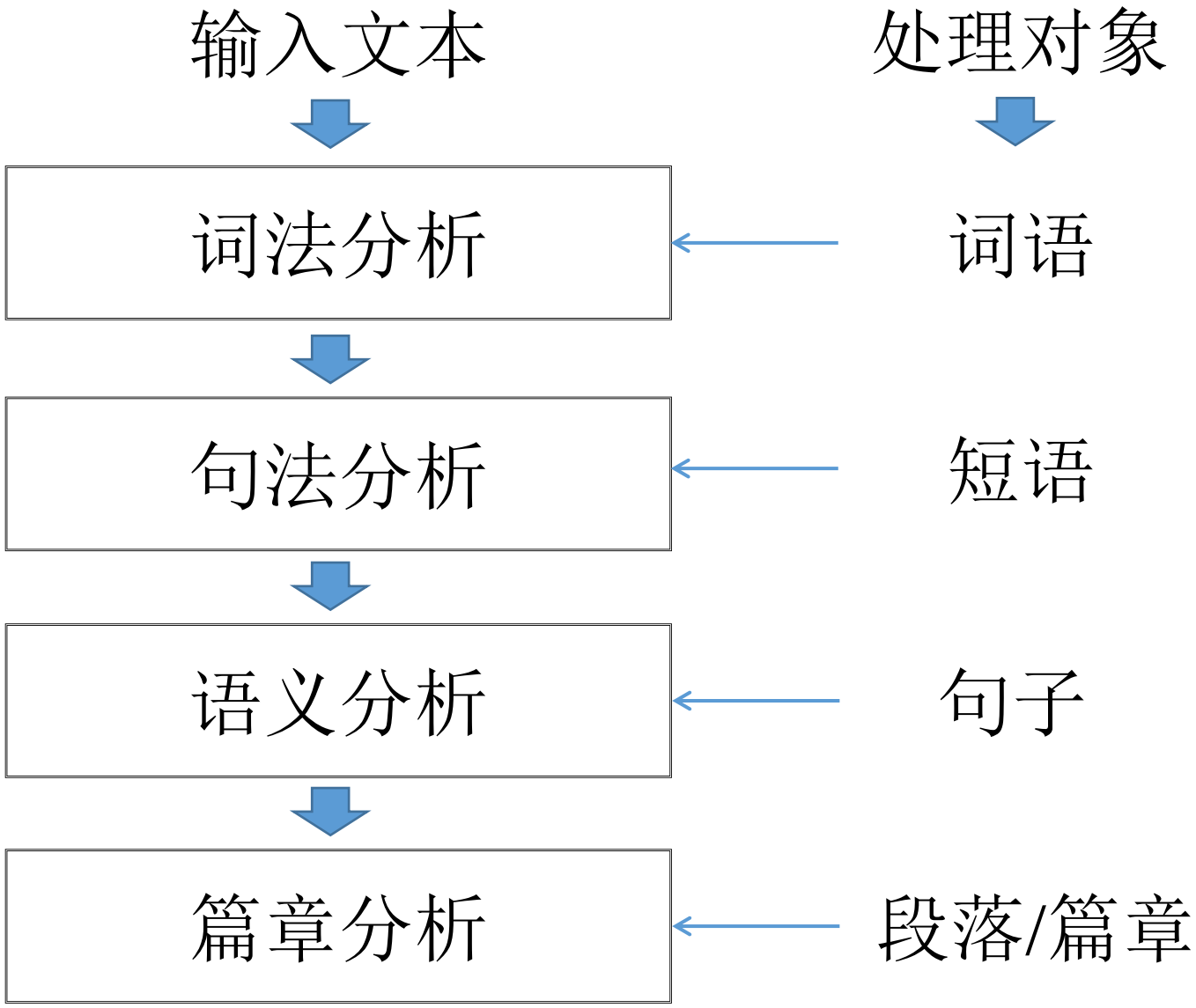
研究任务-语义分析



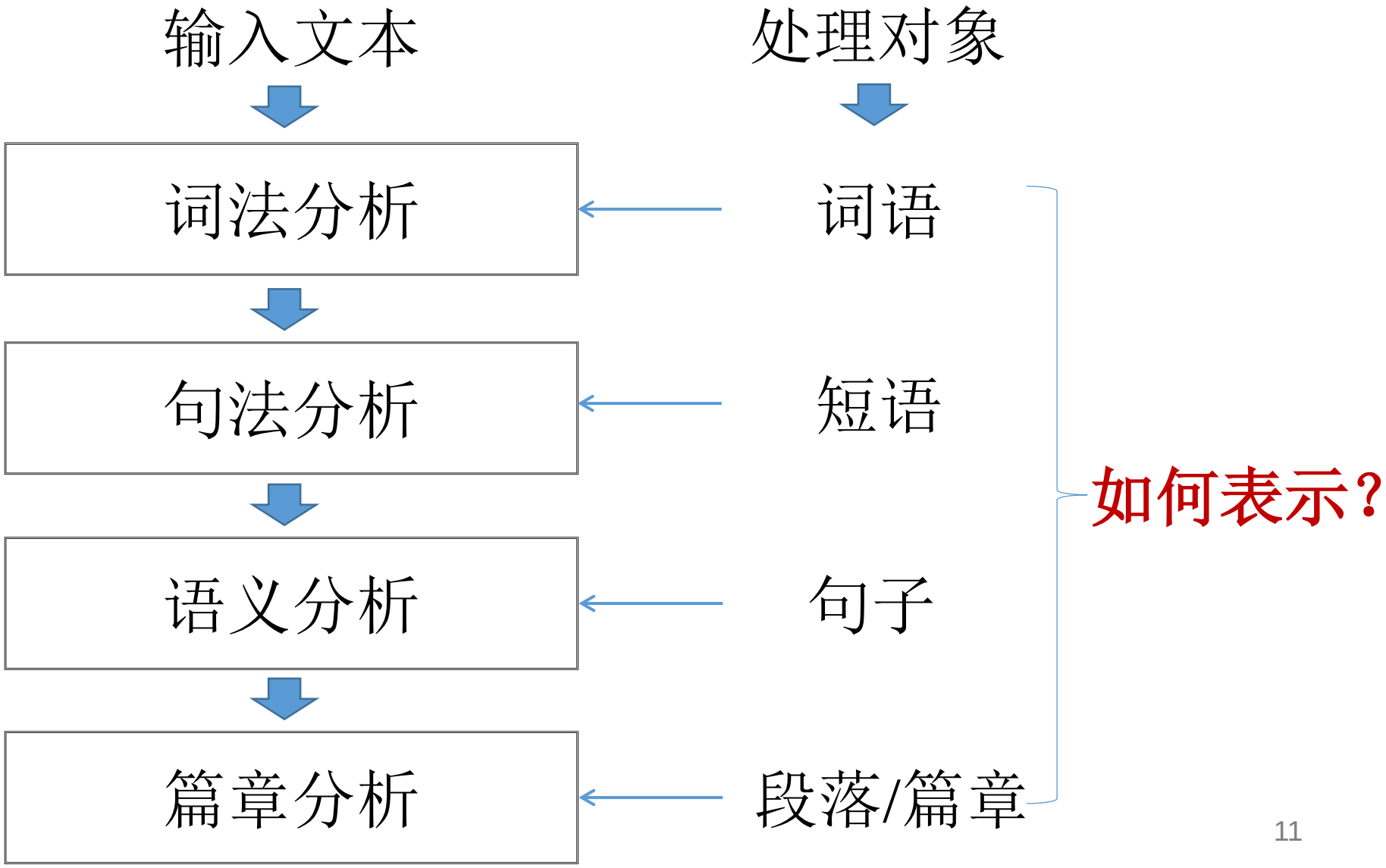
（北京，中国，首都）

实体　实体　关系

自然语言处理



自然语言处理-最基础问题



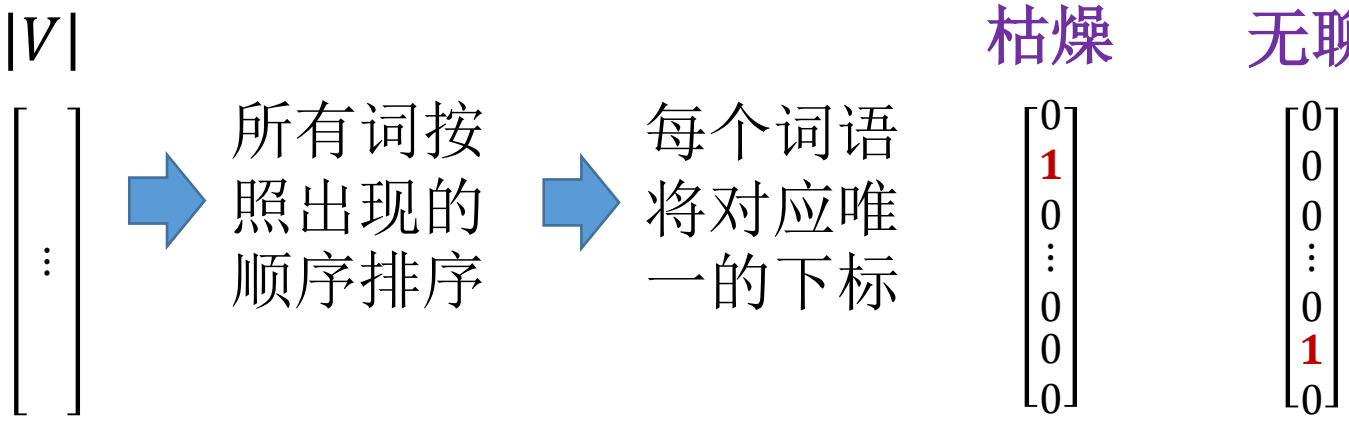
词语表示

- 典型方法：抽象符号（字符串）

  该 报告 很 枯燥 ， 大家 都 觉得 无聊 。

  w_0=该 w_1=报告 w_2=很 w_3=枯燥 w_4=，
  w_5=大家 w_6=都 w_7=觉得 w_8=无聊 w_9=。

- 等价表示方法：one-hot表示法



词语表示

- 问题

枯燥

$$\begin{bmatrix}0\\1\\0\\\vdots\\0\\0\\0\end{bmatrix}$$

无聊

$$\begin{bmatrix}0\\0\\0\\\vdots\\0\\1\\0\end{bmatrix}$$

1，数据稀疏

2，无法捕捉词语间的相似性

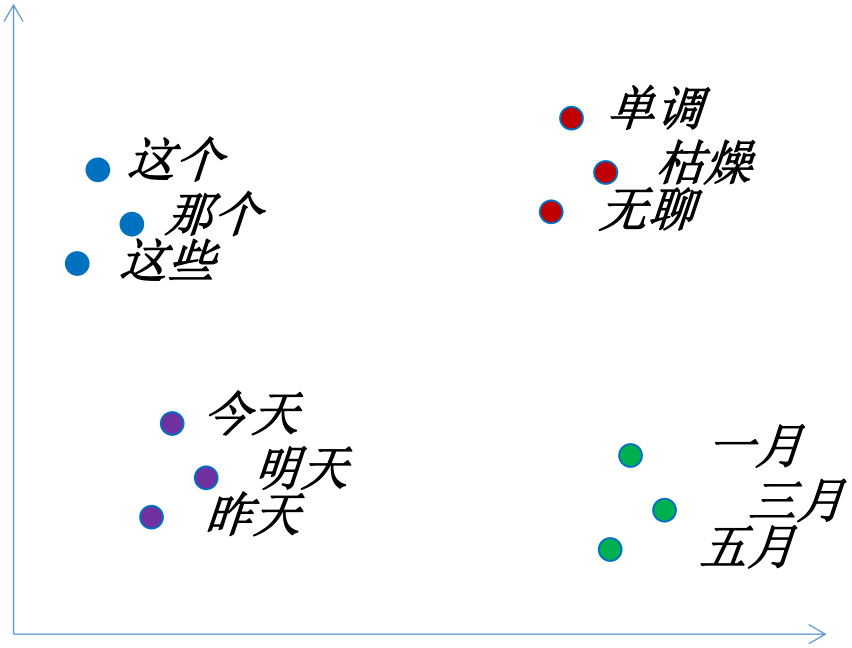
枯燥 $\otimes$ 无聊

$$\begin{bmatrix}0\\1\\0\\\vdots\\0\\0\\0\end{bmatrix} \times \begin{bmatrix}0\\0\\0\\\vdots\\0\\1\\0\end{bmatrix} = 0$$

任意两个词之间的相似度都为0！

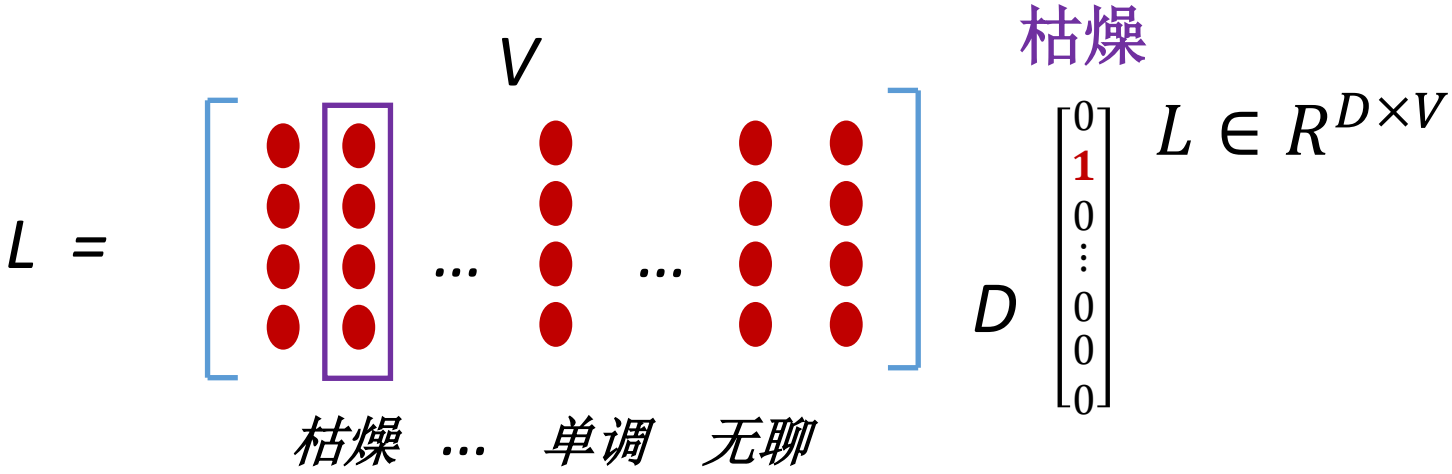
词语表示

- 如果 …



低维、稠密的连续实数空间

词向量表示

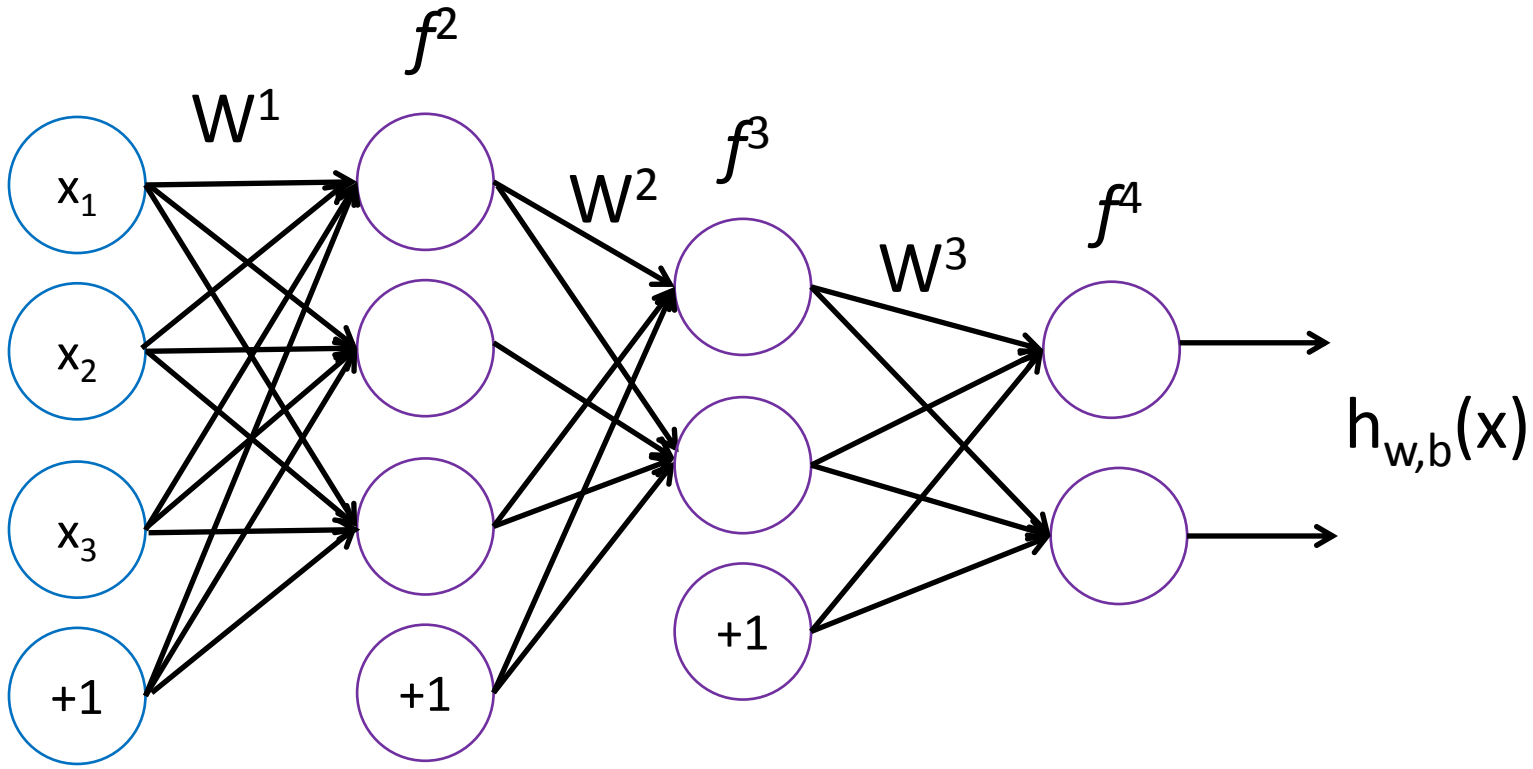


- 通常称为look-up table
  - 我们可以对L右乘一个词的one-hot表示 e 得到该词的低维、稠密的实数向量表达：$x = Le$
- 初始化
  - 通常先随机初始化，然后通过目标函数优化词的向量表达

自然语言处理常用的几种网络

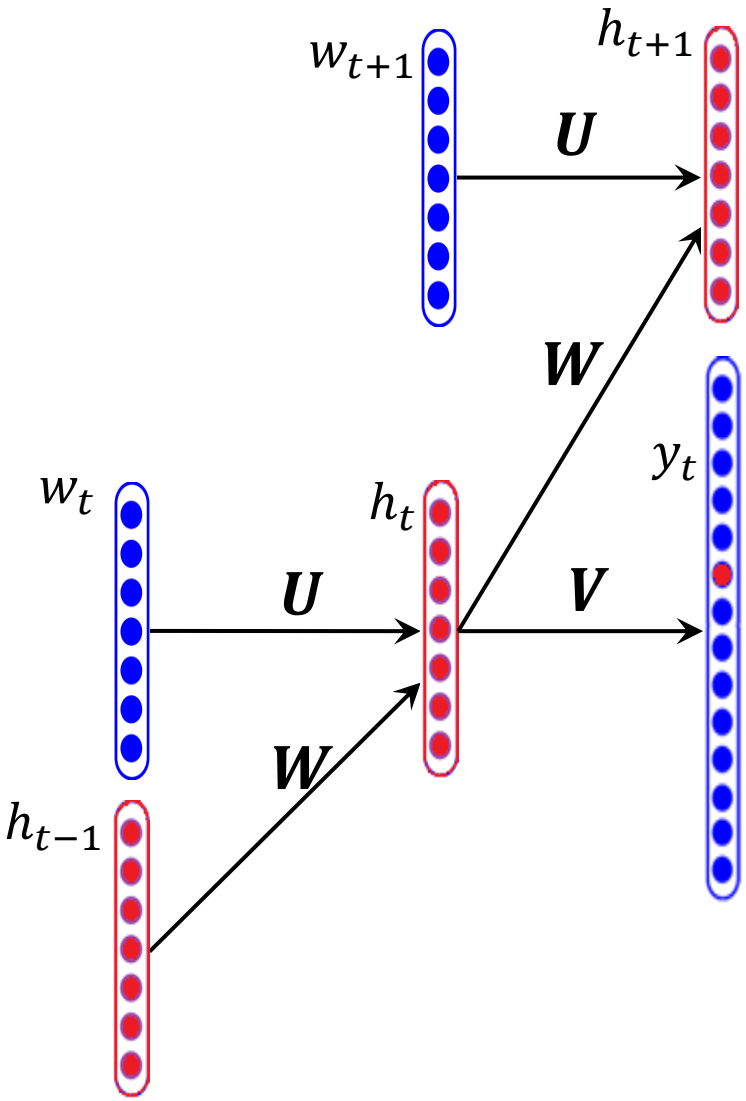
- 前馈神经网络
- 循环神经网络
- （递归）自编码器
- 递归神经网络
- 卷积神经网络

前馈神经网络



输入层　　隐藏层　　输出层

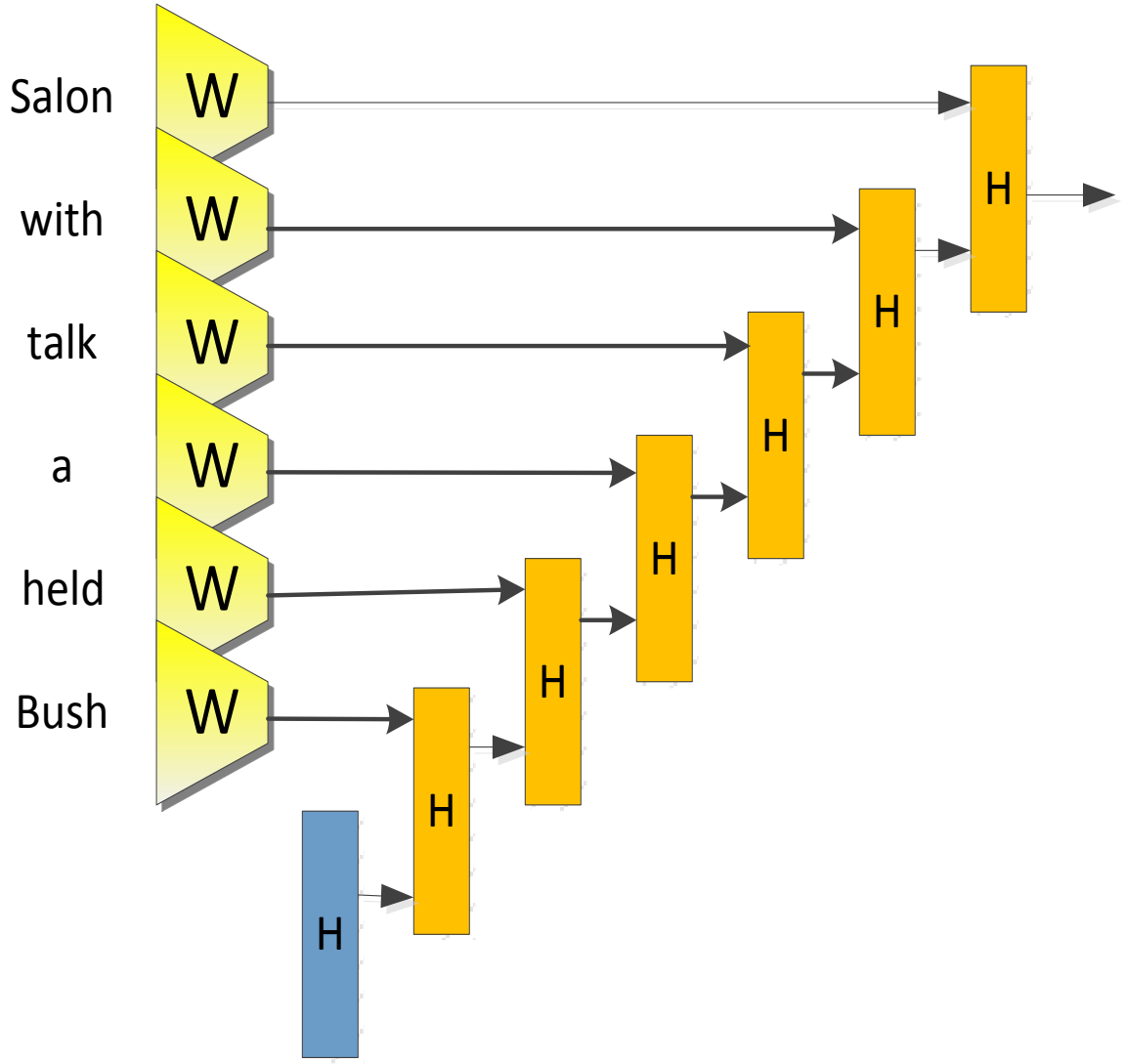
循环神经网络



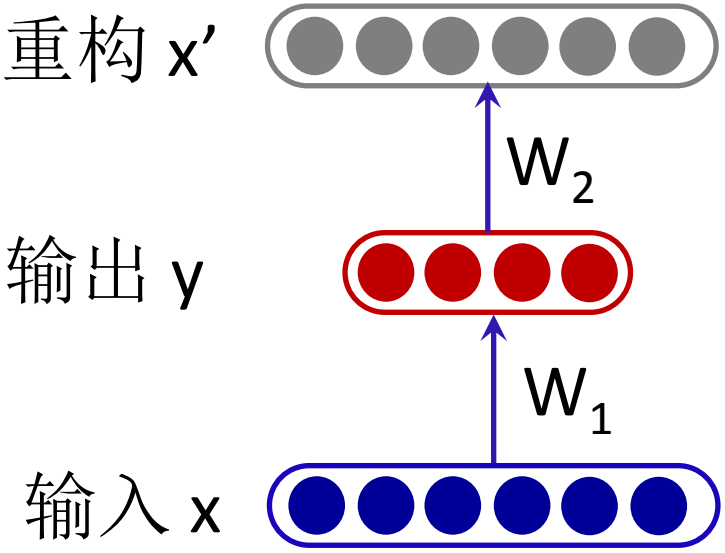
$$h_t = UL(w_t) + Wh_{t-1}$$

$$p(y_t) = \frac{e^{y_t}}{\sum_i e^{y_i}}$$

循环神经网络



自编码器

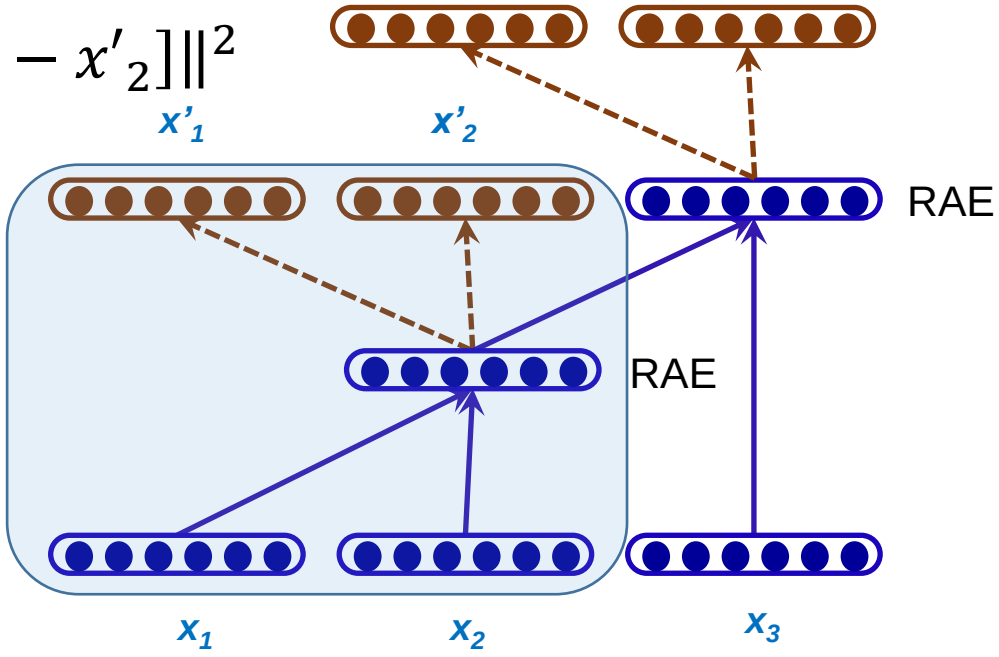


$$W_1, W_2 = argmin\frac{1}{2}\|x - x'\|^2$$

递归自编码器

- 每一层利用相同的自编码器

$$E_{Rec} = \frac{1}{2}\|[x_1, x_2] - [x'_1 - x'_2]\|^2$$

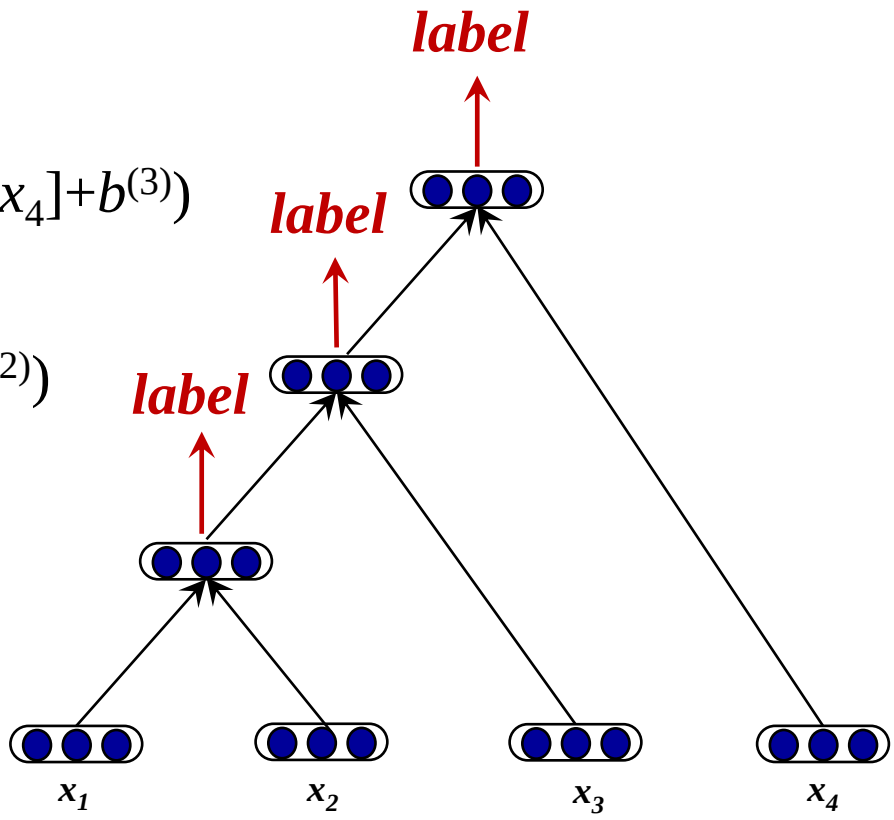


递归神经网络

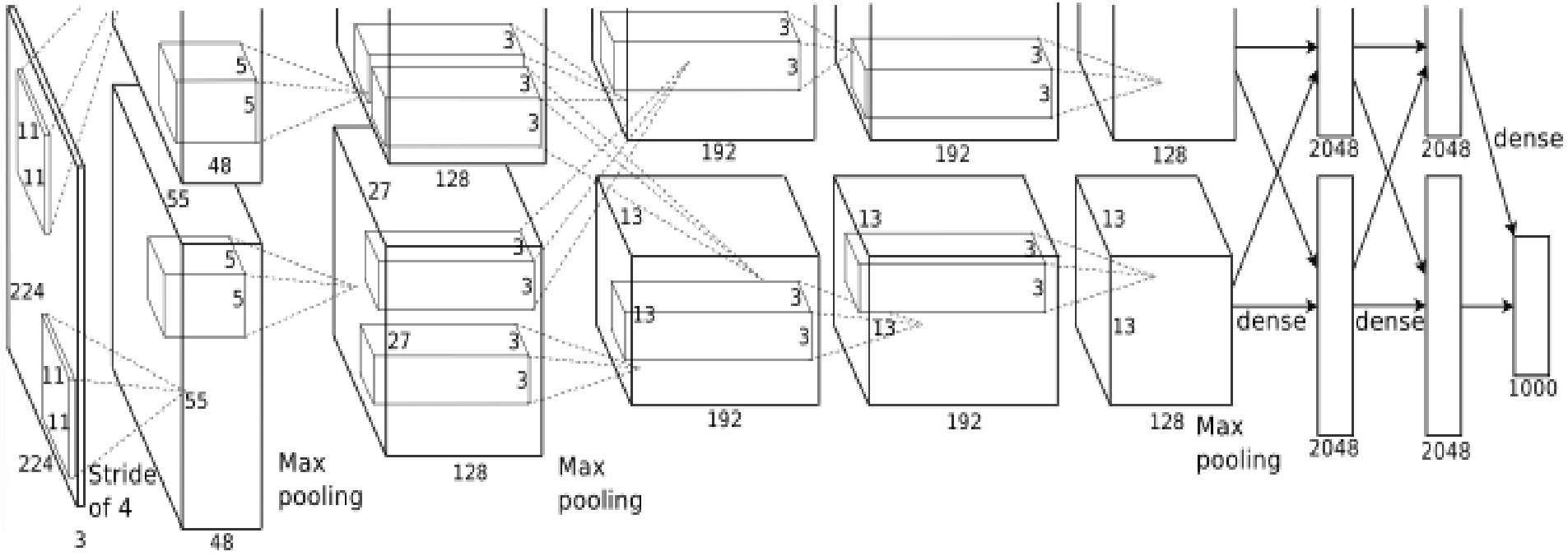
$y_3=f(W^{(3)}[y_2; x_4]+b^{(3)})$

$y_2=f(W^{(2)}[y_1; x_3]+b^{(2)})$

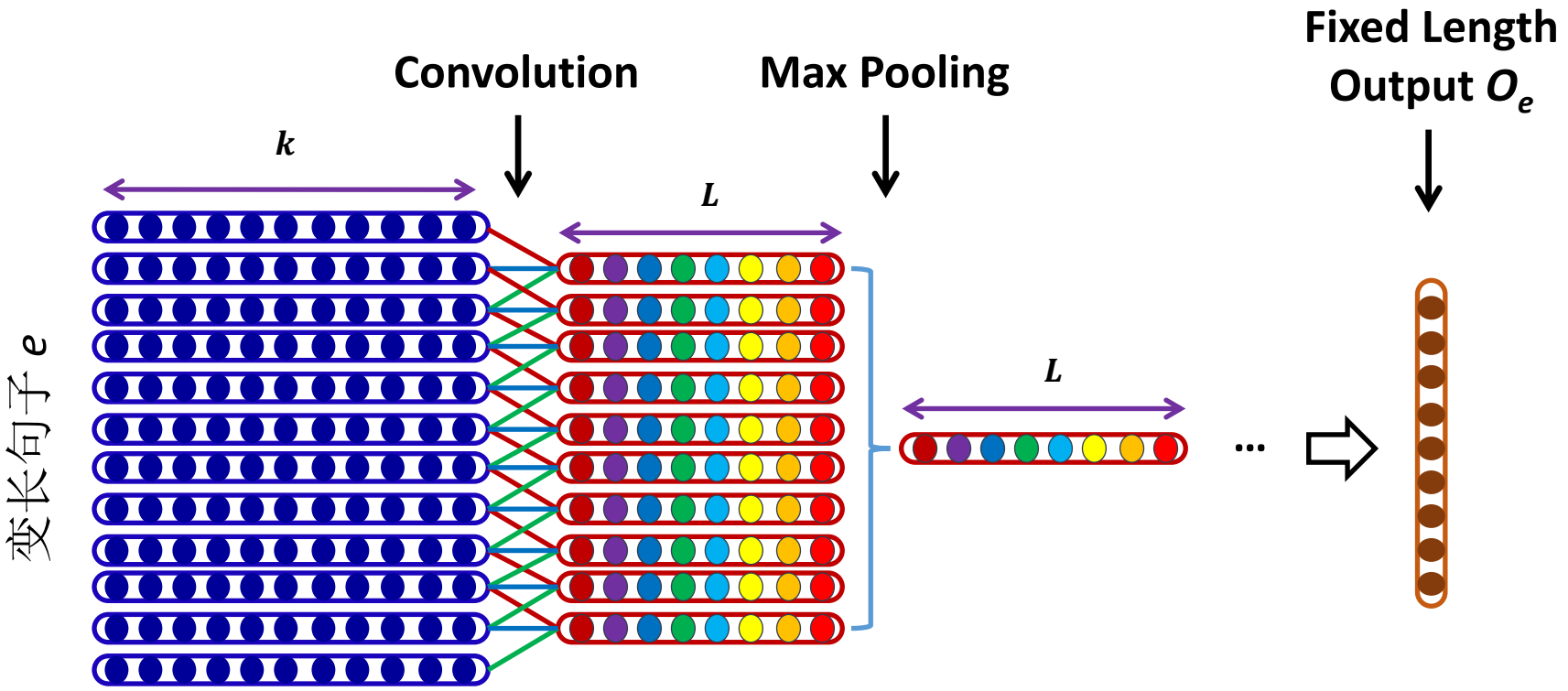
$y_1=f(W^{(1)}[x_1; x_2]+b^{(1)})$



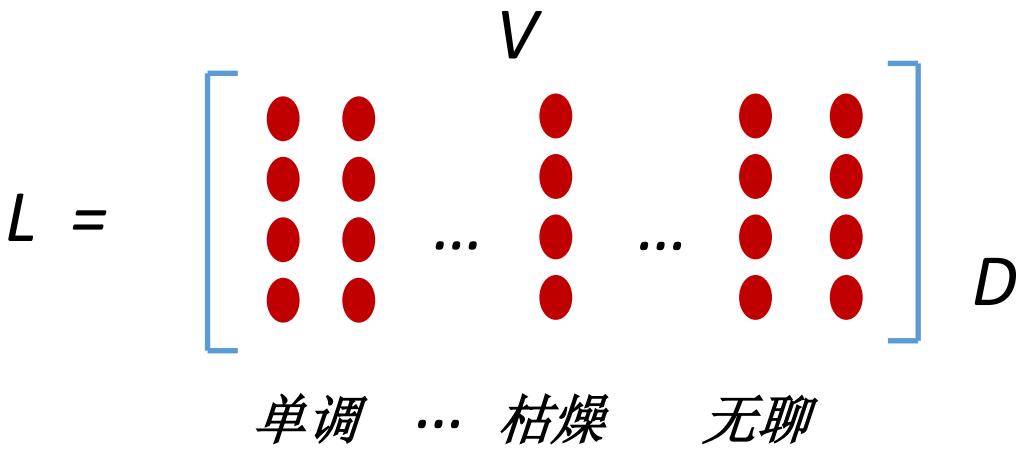
卷积神经网络


224
224
3
11
11
Stride
of 4
48
55
5
5
48
55
Max
pooling
27
128
3
3
3
27
128
Max
pooling
13
192
3
3
3
13
192
3
3
3
13
192
3
3
3
13
128
Max
pooling
dense
2048
2048
dense
dense
2048
2048
1000

卷积神经网络



词向量表示



$$L \in R^{D \times V}$$

- 训练准则: ”You shall know a word by the company it keeps” (J. R. Firth 1957)

government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

词向量表示-语言模型的副产品

- 训练准则: ”You shall know a word by the company it keeps” (J. R. Firth 1957)

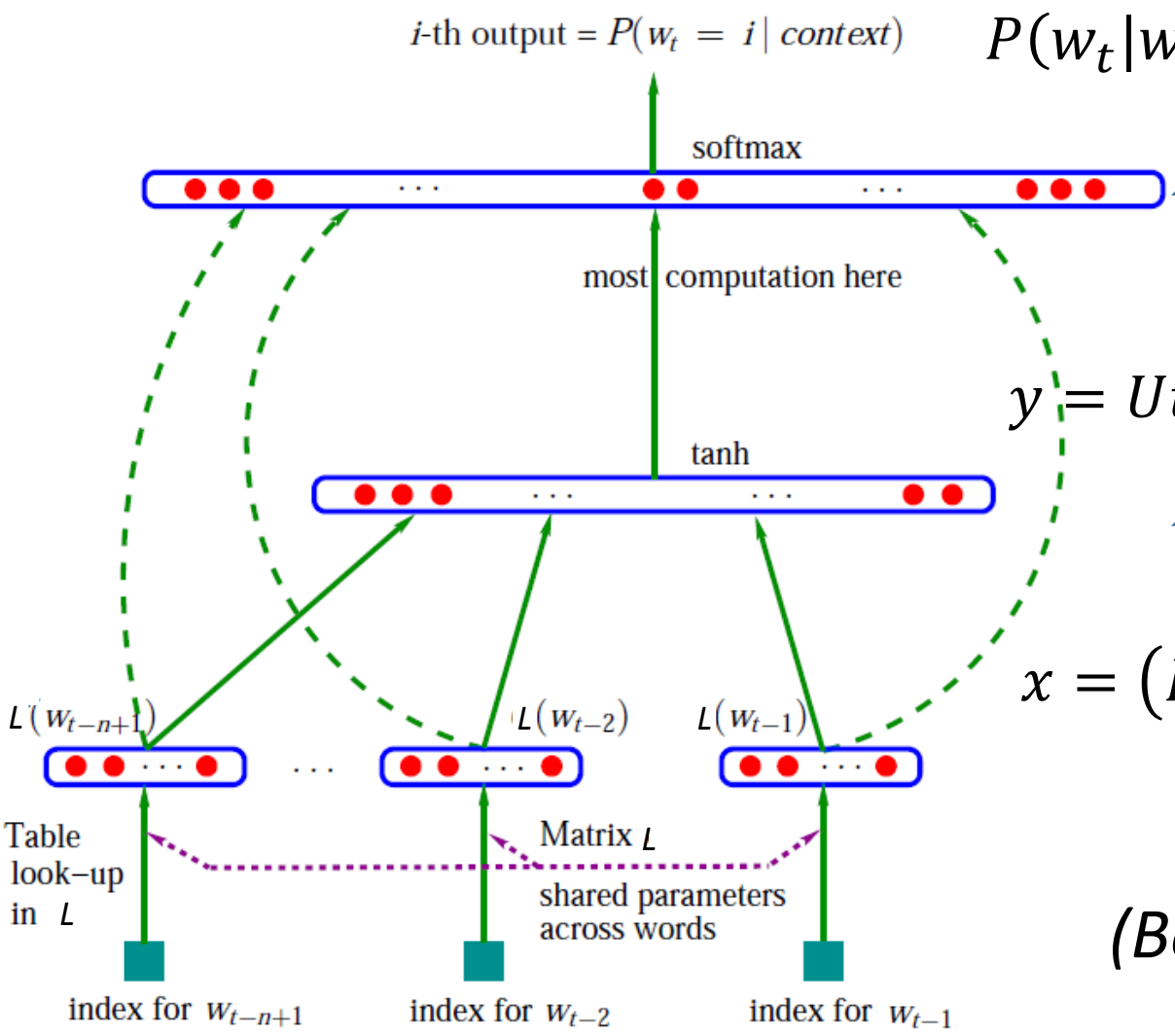
government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

$$P(w_1 w_2 \cdots w_{t-1} w_n)$$

$$= \prod_{t=1}^{n} P(w_t | w_{t-1} \cdots w_{t-n+1})$$

语言模型-前馈神经网络



$$P(w_t|w_{t-1}\cdots w_{t-n+1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

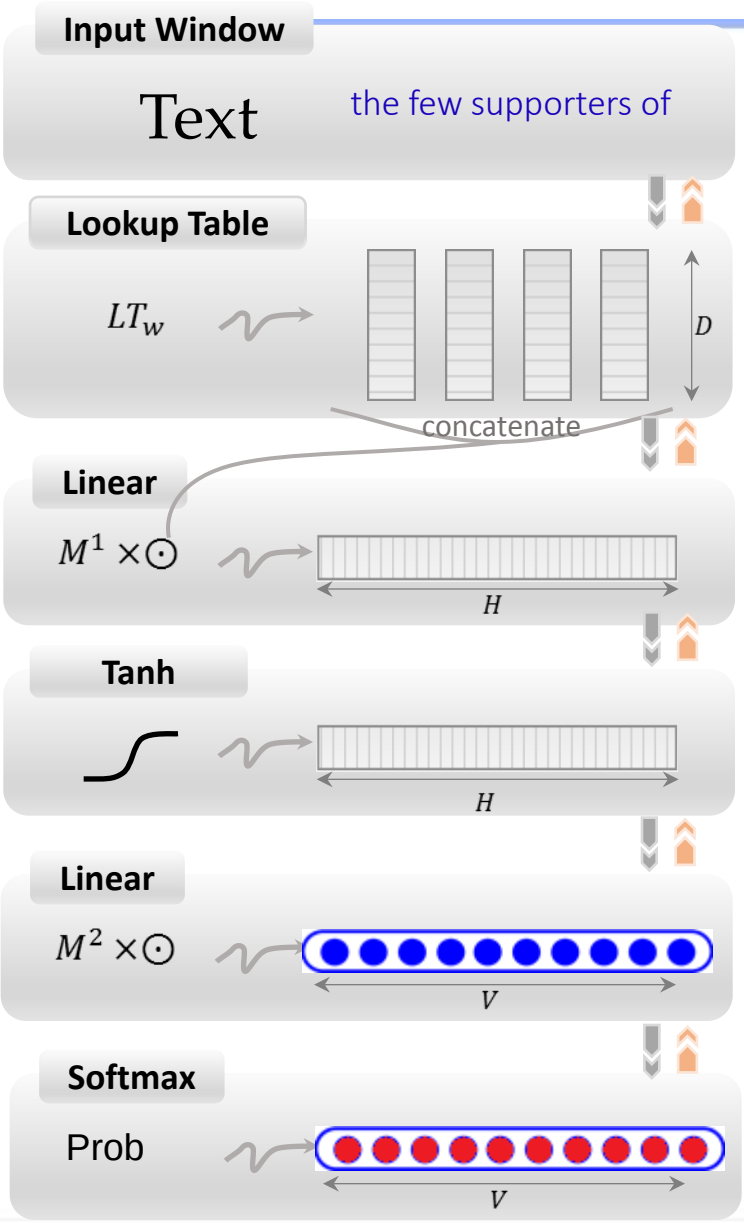
$$\theta \leftarrow \theta + \frac{\partial logP}{\partial \theta}$$

$$y = Utanh(Hx + d) + Wx + b$$

$$x = \left(L(w_{t-1}), \cdots L(w_{t-n+1})\right)$$

(Bengio et al., 2003)

语言模型-前馈神经网络



$P(this|the, few, supporters, of)$

将每个词通过词向量矩阵L映射为低维实数向量

of → *(0.23, 0.15, 0.08, 0.31, .., 0.42)*

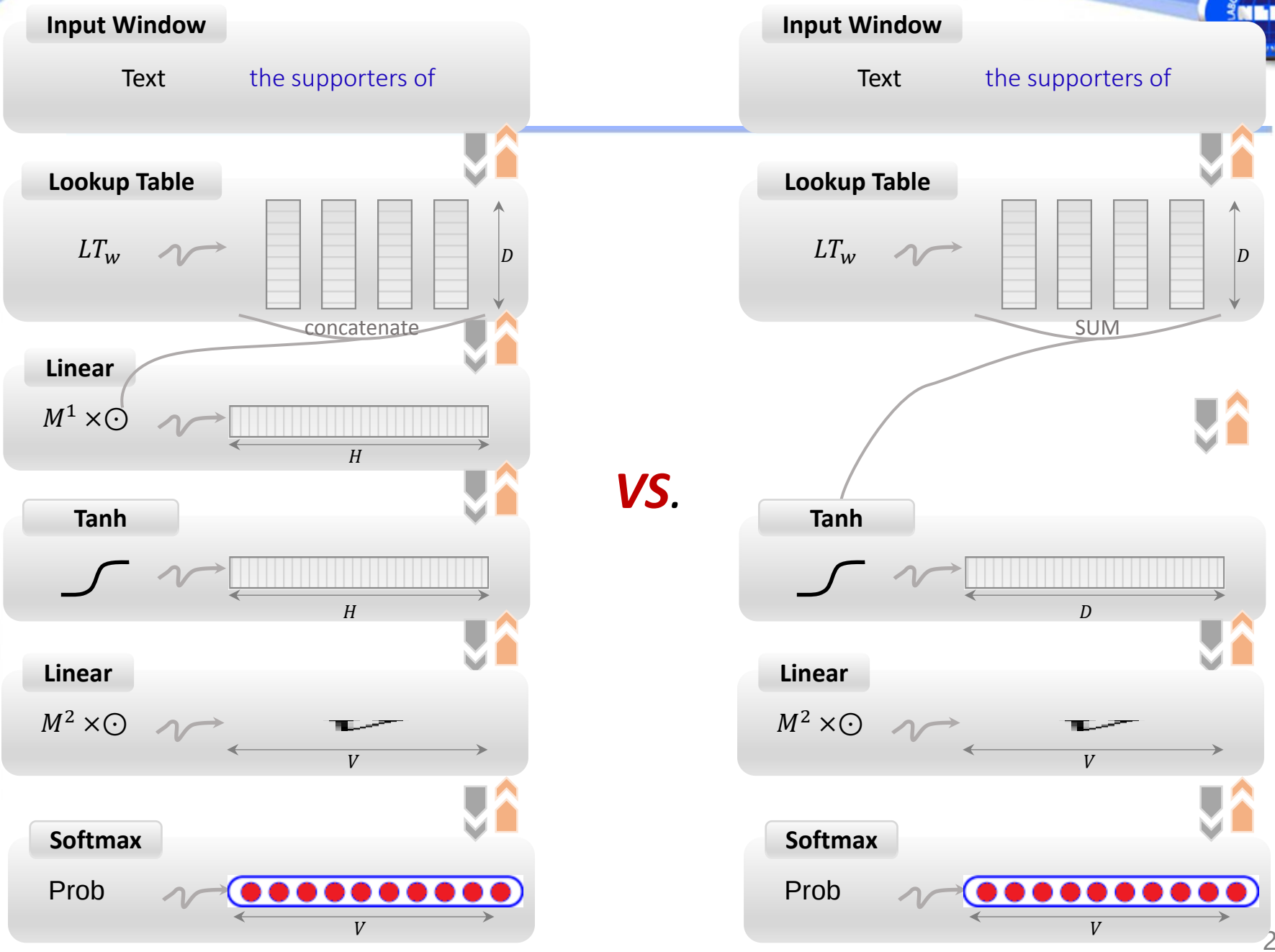
拼接所有词的向量，形成一个向量

隐藏层:
线性映射+非线性变换

⋮

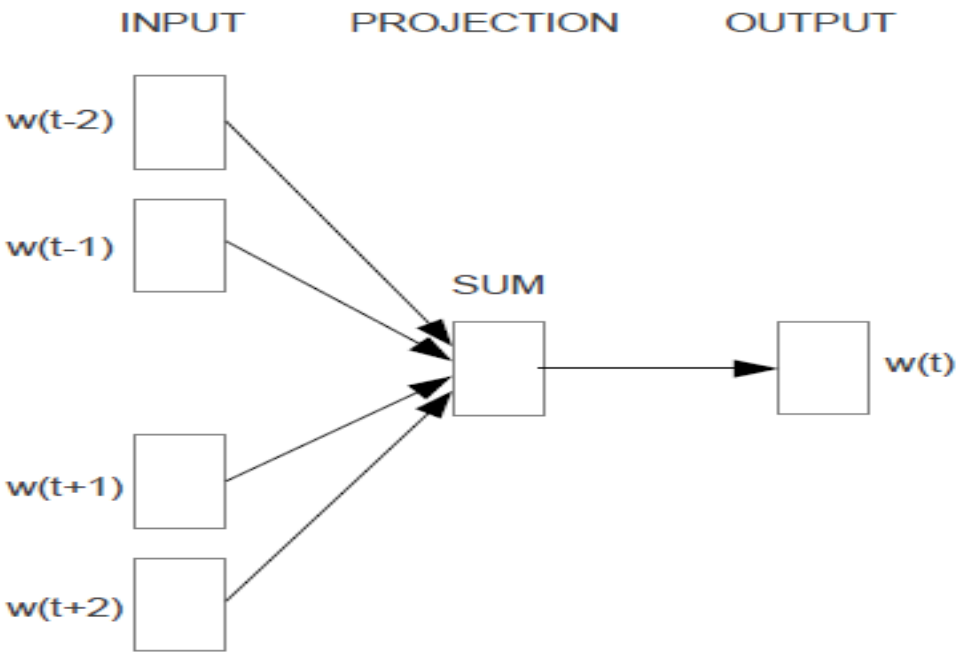
***Softmax* 输出层:**

$P(this|the, few, supporters, of)$

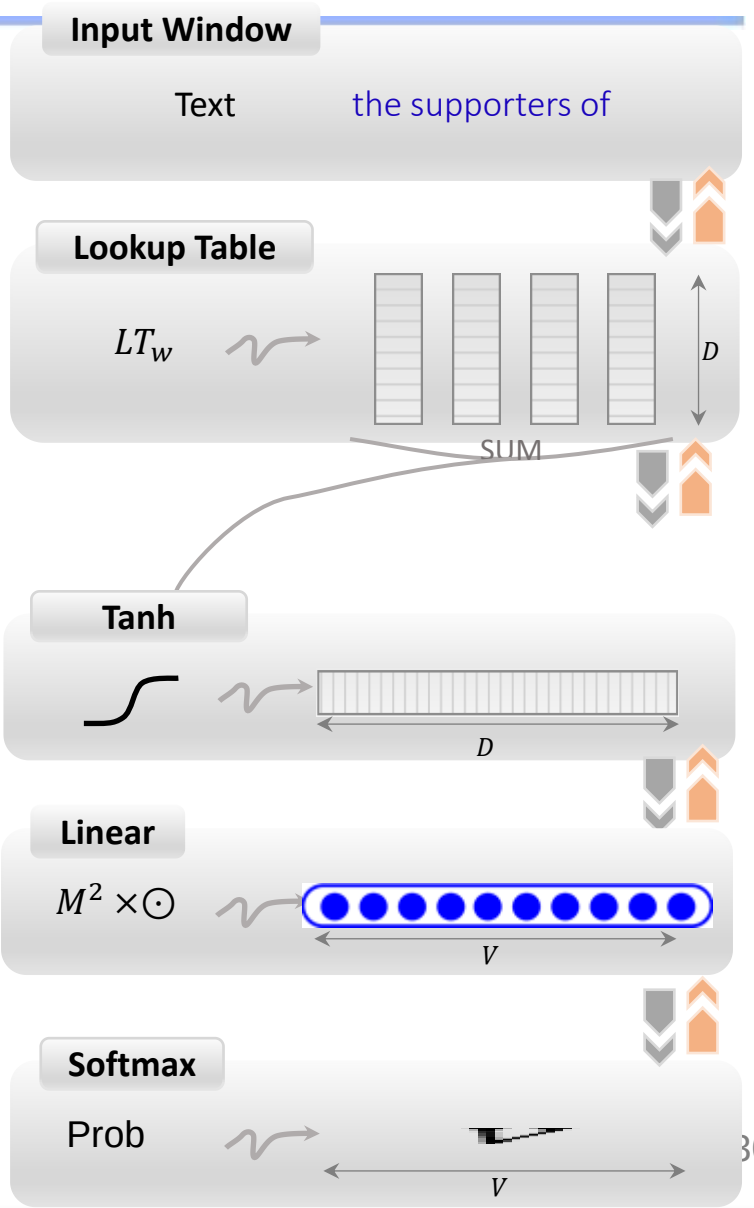

Input Window
Text
the supporters of
Lookup Table
LT_w
D
concatenate
Linear
$M^1 \times \odot$
H
Tanh
H
Linear
$M^2 \times \odot$
V
Softmax
Prob
V
VS.
Input Window
Text
the supporters of
Lookup Table
LT_w
D
SUM
Tanh
D
Linear
$M^2 \times \odot$
V
Softmax
Prob
V

Google's Word2Vec

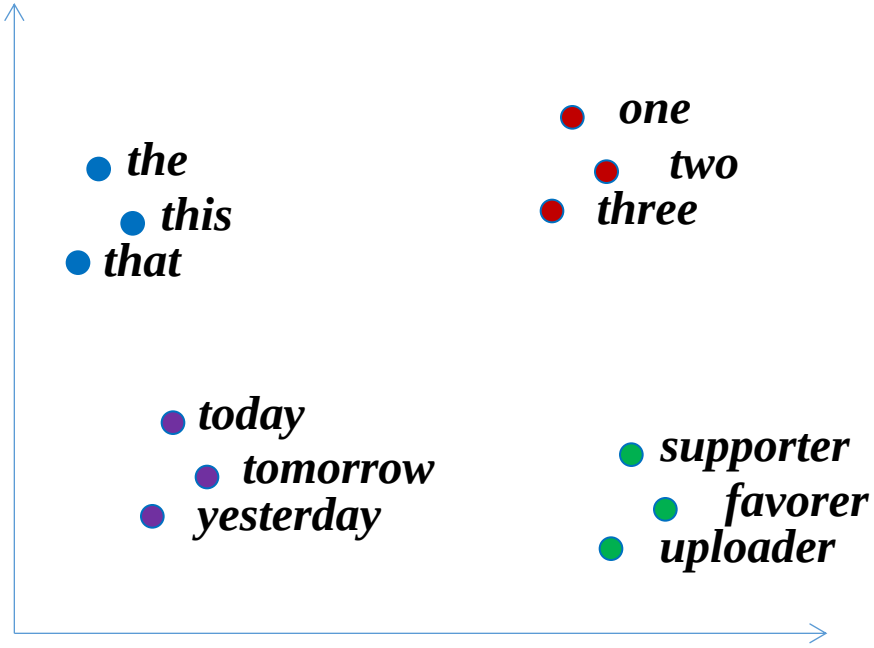
- CBOW: Continuous Bag-of-Words
  - 词序不影响预测



CBOW



词向量分布



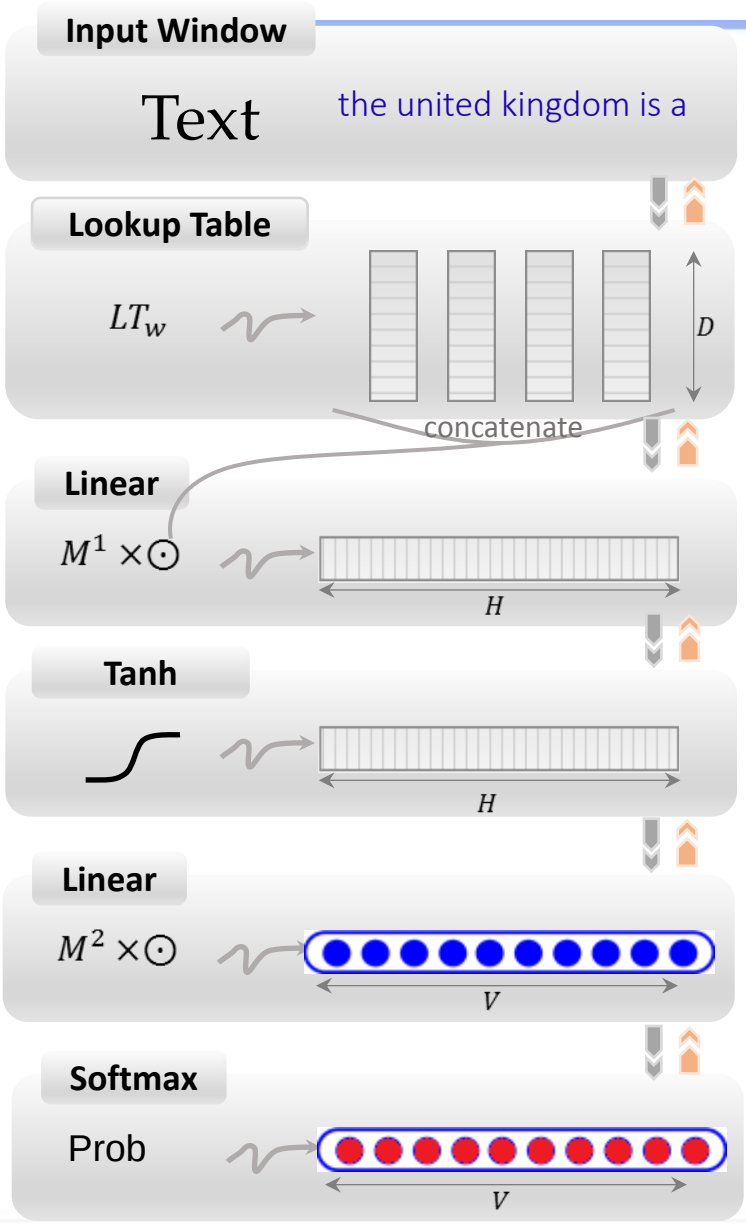
低维、稠密的实数向量空间

在低维、稠密的实数向量空间中，相似的词聚集在一起，在相同的历史上下文中具有相似的概率分布！

词向量的应用

- 神经网络语言模型
- 词性标注
- 命名实体识别
- 依存句法分析
- 统计机器翻译
- ……

命名实体识别



$$P(\boldsymbol{tag}|kingdom)$$
$$\approx P(\boldsymbol{tag}|the, united, kingdom, is, a)$$

$$\boldsymbol{tag} = \{\boldsymbol{NER_B}, \boldsymbol{NER_I}, \boldsymbol{NER_E}, \boldsymbol{NER_O}\}$$

$$P(\boldsymbol{NER_E}|kingdom)$$

统计机器翻译

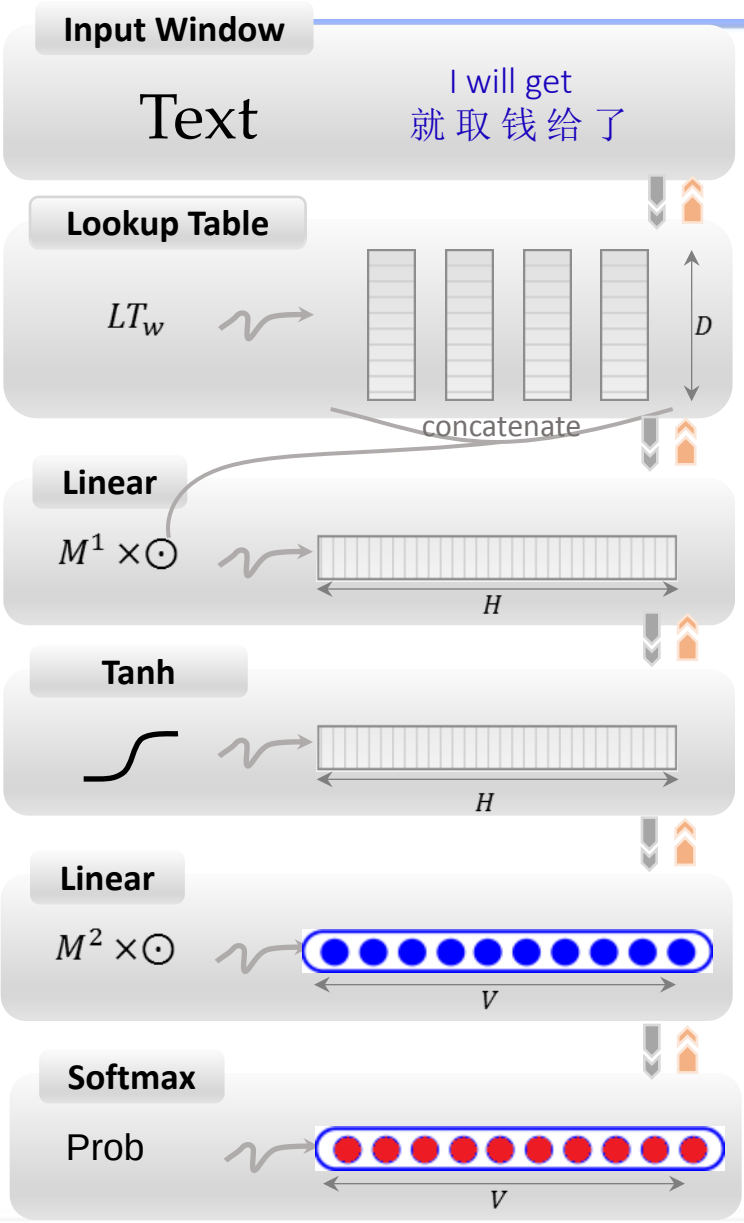
j 如何选?

$$P(e_i) \approx P(e_i|e_1 \cdots e_{i-1}, f)$$
$$\approx P\left(e_i|e_{i-3} \cdots e_{i-1}, f_{j-c} \cdots f_j \cdots f_{j+c}\right)$$

训练目标函数:

$$L = \sum_i log\left(P(e_i)\right)$$

统计机器翻译



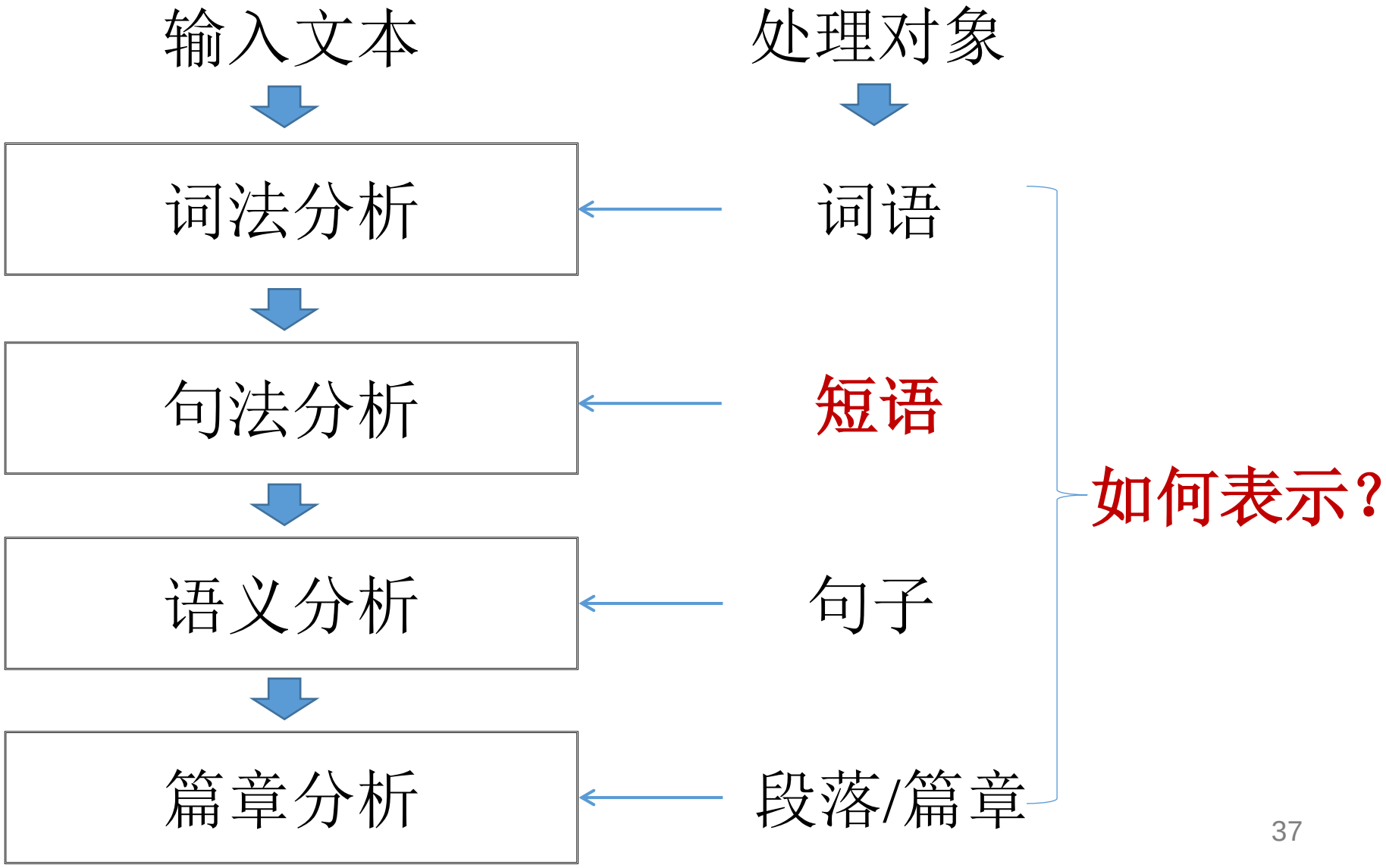
- 上下文
  - 目标语言 4-gram
  - 源语言中心词左右5个词
- 词向量 (192 维)
- 两个隐藏层 (512维)
- 输出层 softmax

$$P(\boldsymbol{e_i}|e_{i-3}\cdots e_{i-1}, f_{j-c}\cdots f_j \cdots f_{j+c})$$

统计机器翻译

	Ar-En	Ch-En
	BLEU	BLEU
OpenMT12 - 1st Place	49.5	32.6
OpenMT12 - 2nd Place	47.5	32.2
OpenMT12 - 3rd Place	47.4	30.8
...	...	...
OpenMT12 - 9th Place	44.0	27.0
OpenMT12 - 10th Place	41.2	25.7
Baseline (w/o RNNLM)	48.9	33.0
Baseline (w/ RNNLM)	49.8	33.4
+ S2T/L2R NNJM (Dec)	51.2	34.2
+ S2T NNLTM (Dec)	52.0	34.2
+ T2S NNLTM (Resc)	51.9	34.2
+ S2T/R2L NNJM (Resc)	52.2	34.3
+ T2S/L2R NNJM (Resc)	52.3	34.5
+ T2S/R2L NNJM (Resc)	52.8	34.7
"Simple Hier." Baseline	43.4	30.1
+ S2T/L2R NNJM (Dec)	47.2	31.5
+ S2T NNLTM (Dec)	48.5	31.8
+ Other NNJMs (Resc)	49.7	32.2

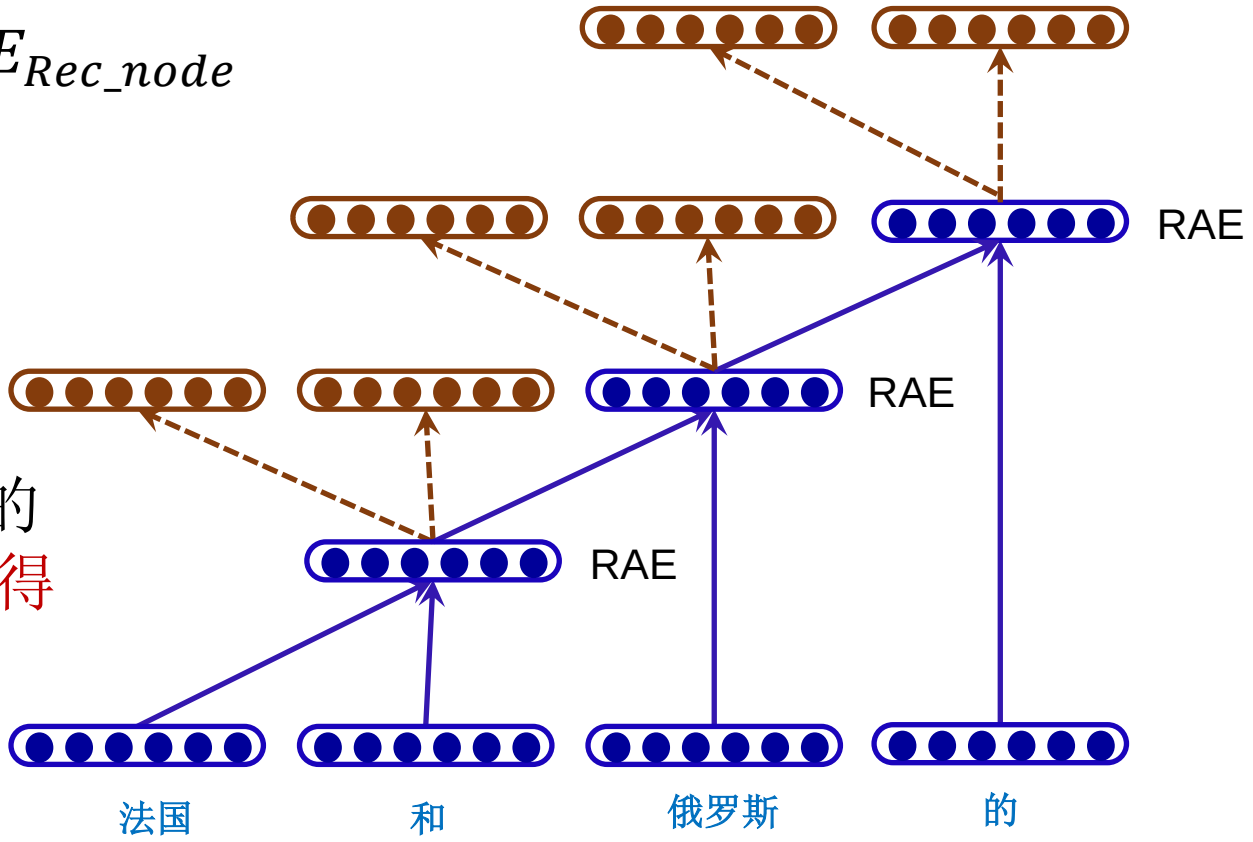
自然语言处理-最基础问题



短语向量-无监督递归自编码器

- 目标函数：最小化所有节点的重构误差

$$E_{total} = \sum_{node} E_{Rec\_node}$$

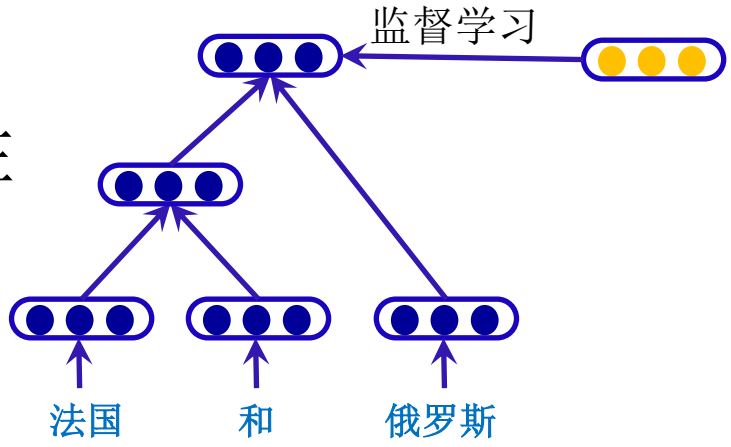


只能学到一部分短语的句法信息，,但 无法获得正确的 语义

短语向量-无监督递归自编码器

- 理想方法：有标注数据

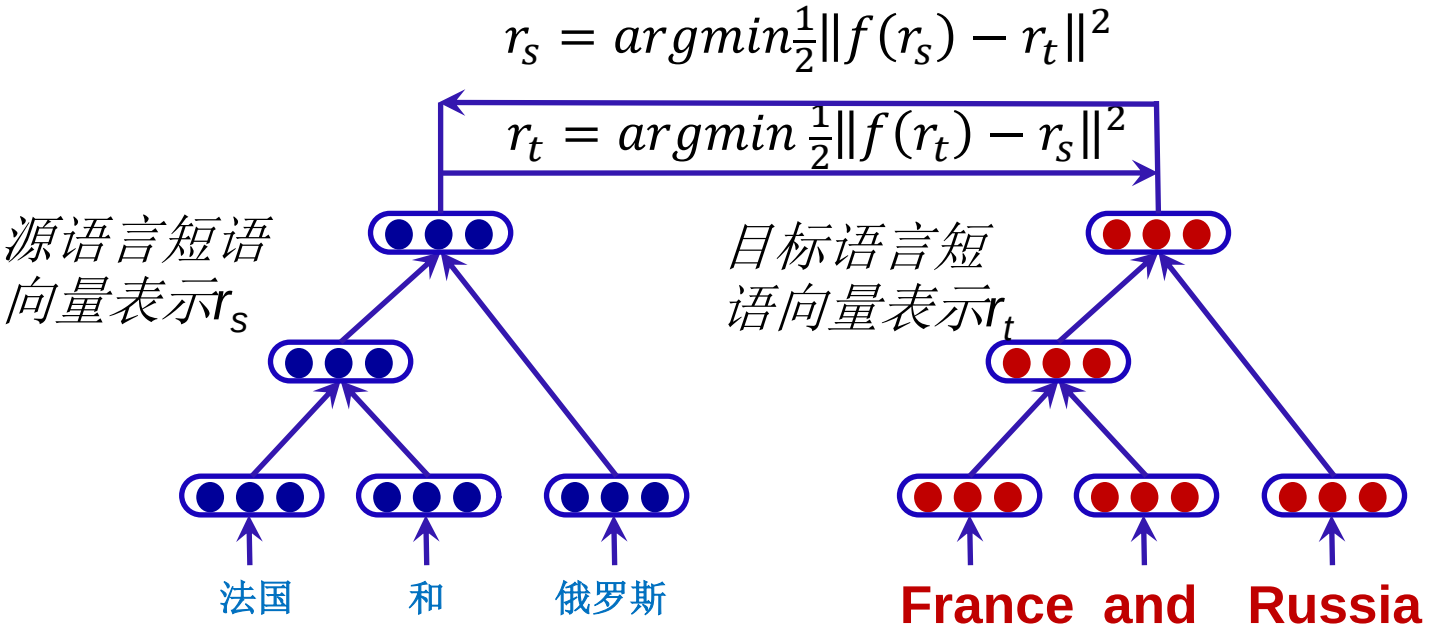
- 但是， 现实中不存在正确标注的短语向量



短语向量-双语约束的递归自编码器

- 假设
  - 短语与其翻译具有相同的语义向量表示
- 目标函数
  - 最小化短语翻译对间的语义表示误差
- 模型
  - Pre-training：无监督递归自编码器学习短语初始表示
  - Fine-tuning：相互监督学习，优化短语向量表示

短语向量-双语约束的递归自编码器



短语向量-双语约束的递归自编码器

- 目标函数

正则化项

$$J = E(S,T;\theta) + \frac{1}{2}\|\lambda\|^2$$

重构误差

双语语义误差

$$E(S,T;\theta) = \alpha E_{rec}(S,T;\theta) + (1-\alpha)E_{regression}(S,T;\theta)$$

$$E_{rec}(S,T;\theta) = E_{rec}(S;\theta) + E_{rec}(T;\theta)$$

$$E_{regression}(S,T;\theta) = E_{regression}(S|T,\theta) + E_{regression}(T|S,\theta)$$

$$E_{regression}(S|T,\theta) = \sum_{s\in S} \frac{1}{2}\|f(v_{s\_root}) - v_{t\_root}\|^2$$

短语向量-双语约束的递归自编码器

相似短语:	do not agree will definitely reject will never accept … …
相似短语:	abstract meaning real meaning intrinsic logic … …
相似短语:	what is your opinion what do you think about how do you view those … …

短语向量应用

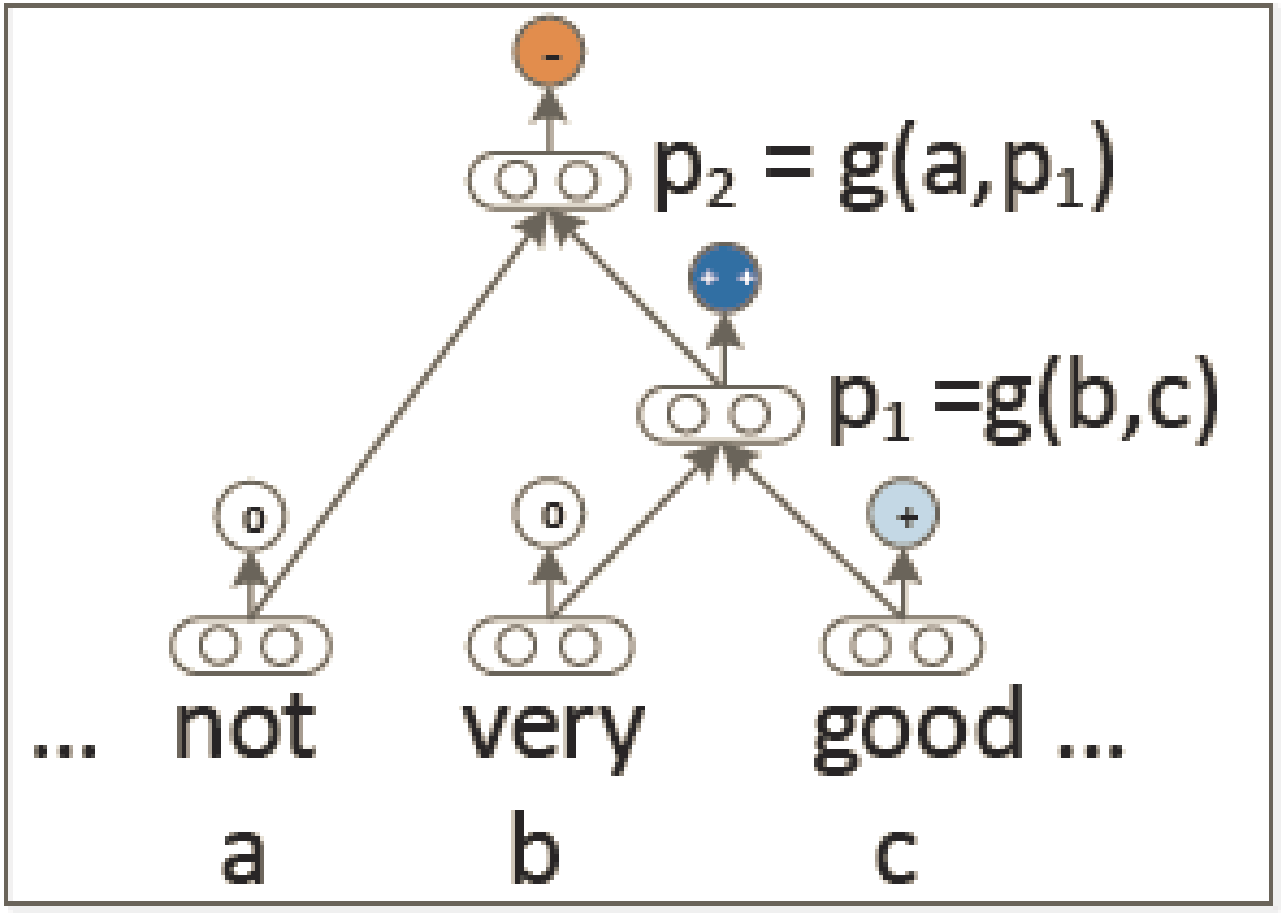
- 短语相似度计算
- 句法分析
- 情感分析
- 统计机器翻译
- …… ……

情感分析

- 任务定义

This film is not very good.

情感分析



将词、短语与句子利用递归神经网络表示为连续向量，从而预测词、短语与句子的情感极性

情感分析

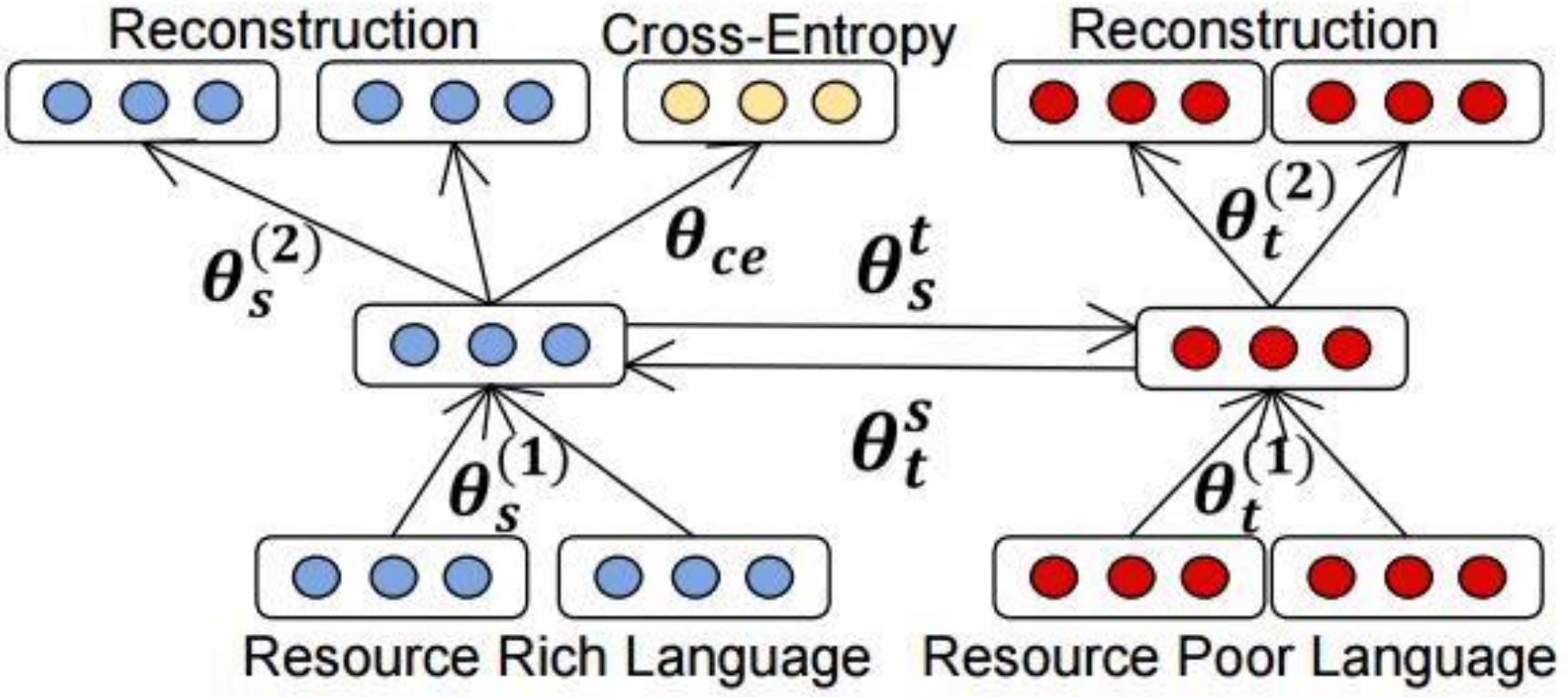
Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.6**	**87.6**	**85.4**

跨语言情感分析

This film is not very good. 

这 部 手机 不是 很 好.

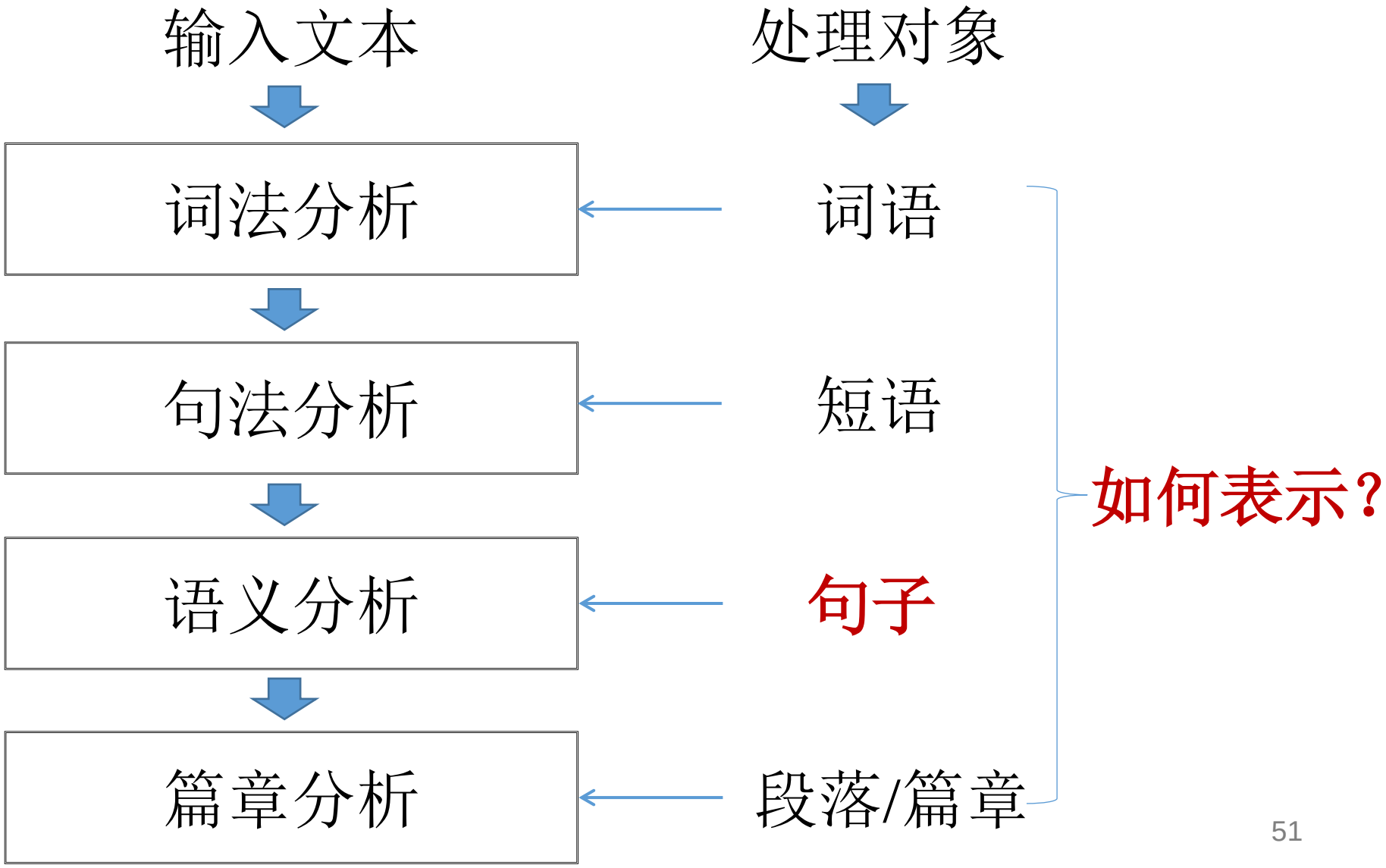
跨语言情感分析


Reconstruction
Cross-Entropy
Reconstruction
$\theta_s^{(2)}$
θ_{ce}
θ_s^t
$\theta_t^{(2)}$
θ_t^s
$\theta_s^{(1)}$
$\theta_t^{(1)}$
Resource Rich Language
Resource Poor Language

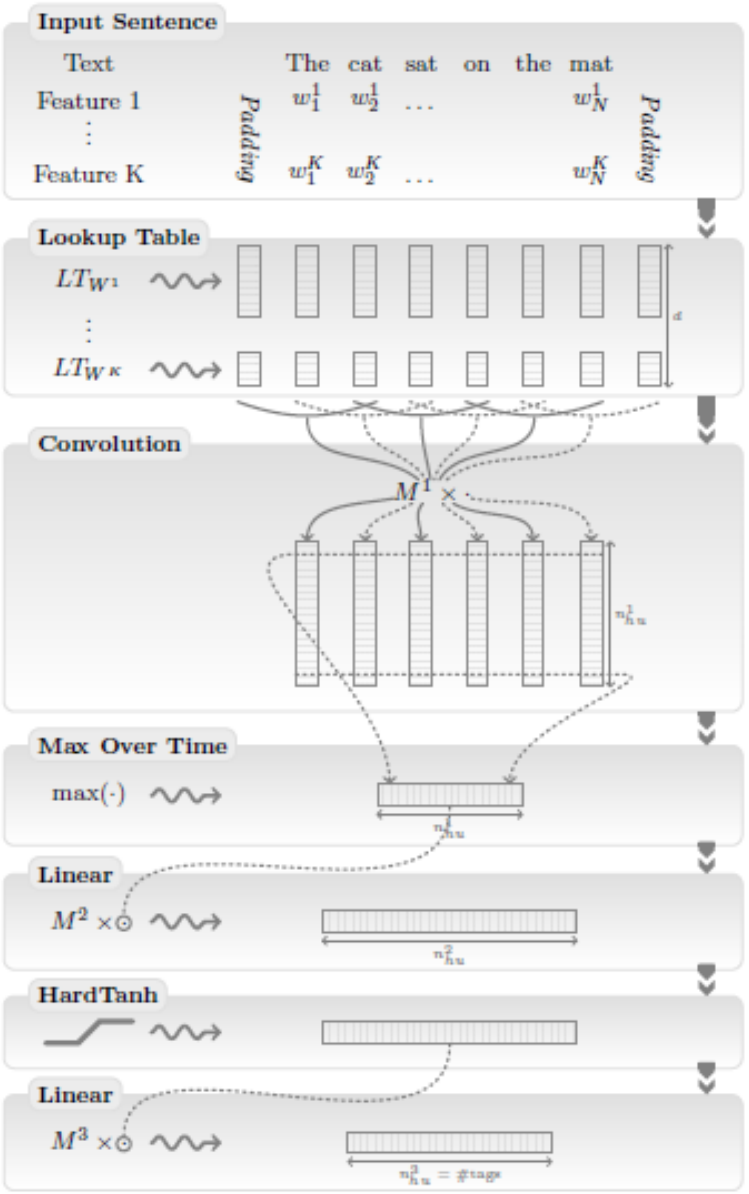
跨语言情感分析

Dataset	RHMR		SMRD
Classifier	Ratings	Polarity	Polarity
Majority class	35.19	51.83	52.34
Bag-of-Words	51.98	62.52	68.47
WordNet based	55.47	67.29	75.5
XL Clustering	72.34	84.46	84.71
Basic RAE	75.53	79.31	81.06
BRAE-U	76.01	82.66	84.83
BRAE-P	79.70	84.85	87.00
BRAE-F	81.22	90.50	90.21

自然语言处理-最基础问题

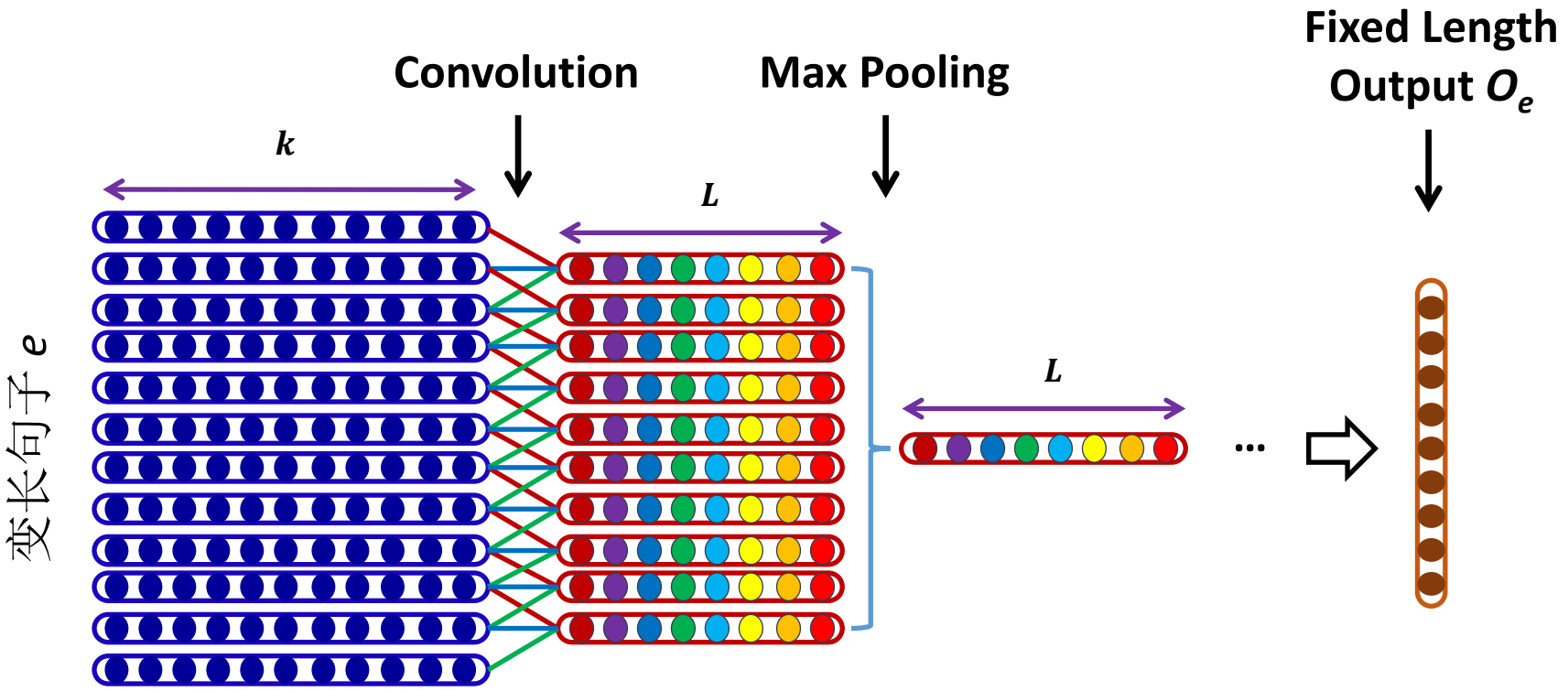


卷积神经网络

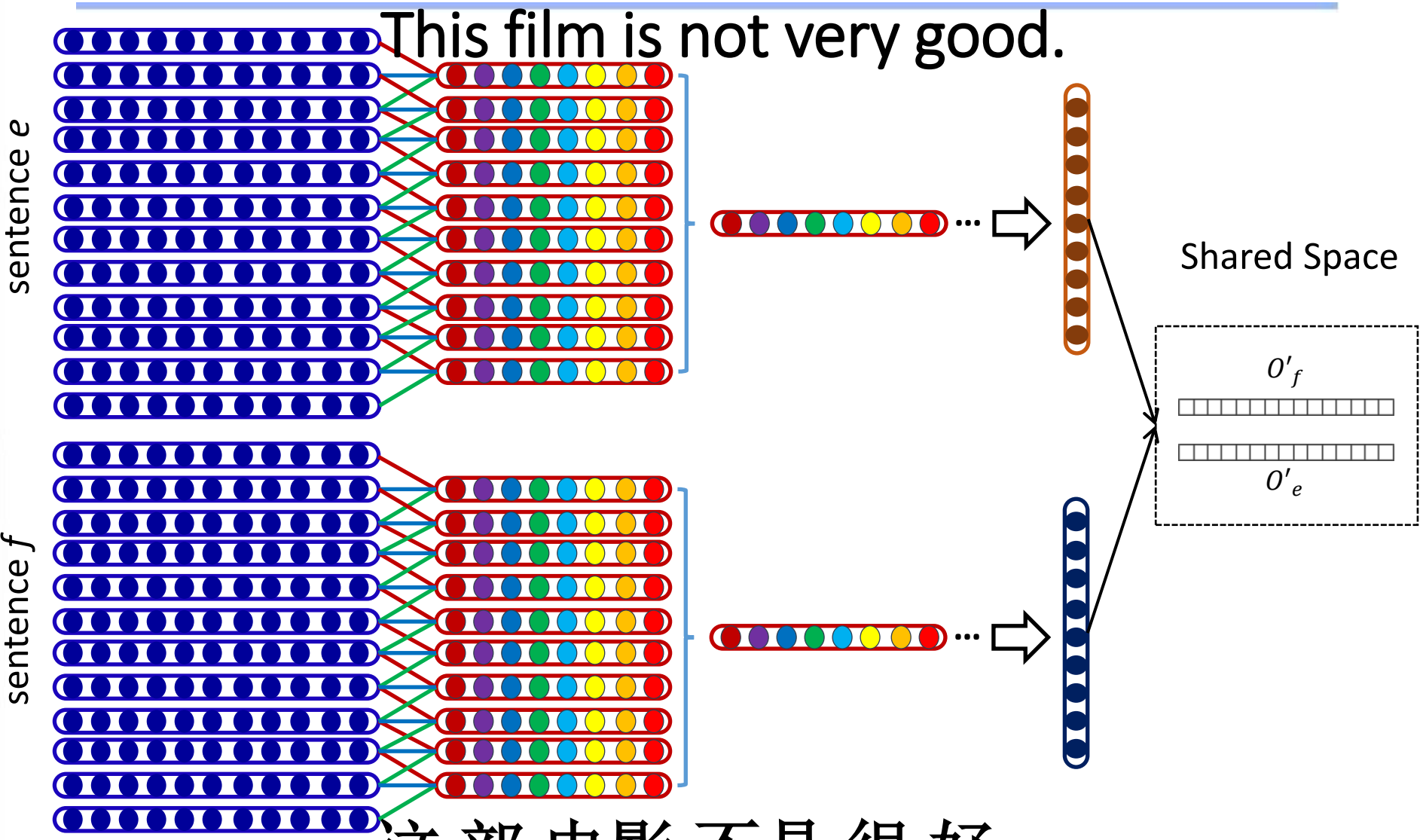


- Convolution:
  - 假设每个词语50维向量
  - 设计矩阵M (150*100)
  - $M*L_{i:i+2}$ (i=0, …, N)
- Pooling:
  - 每行选择K=1最大值
  - 100维输出

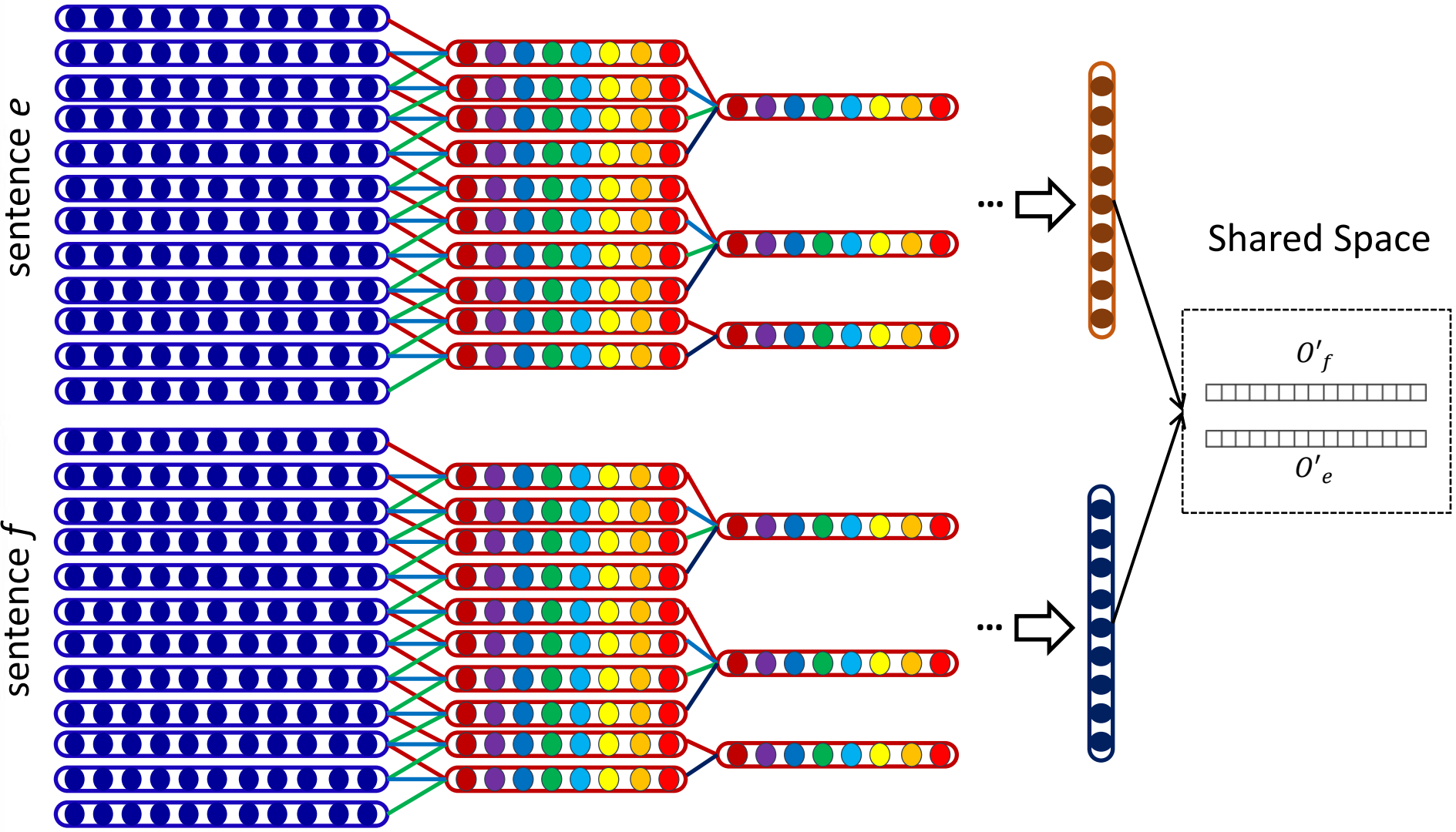
卷积神经网络



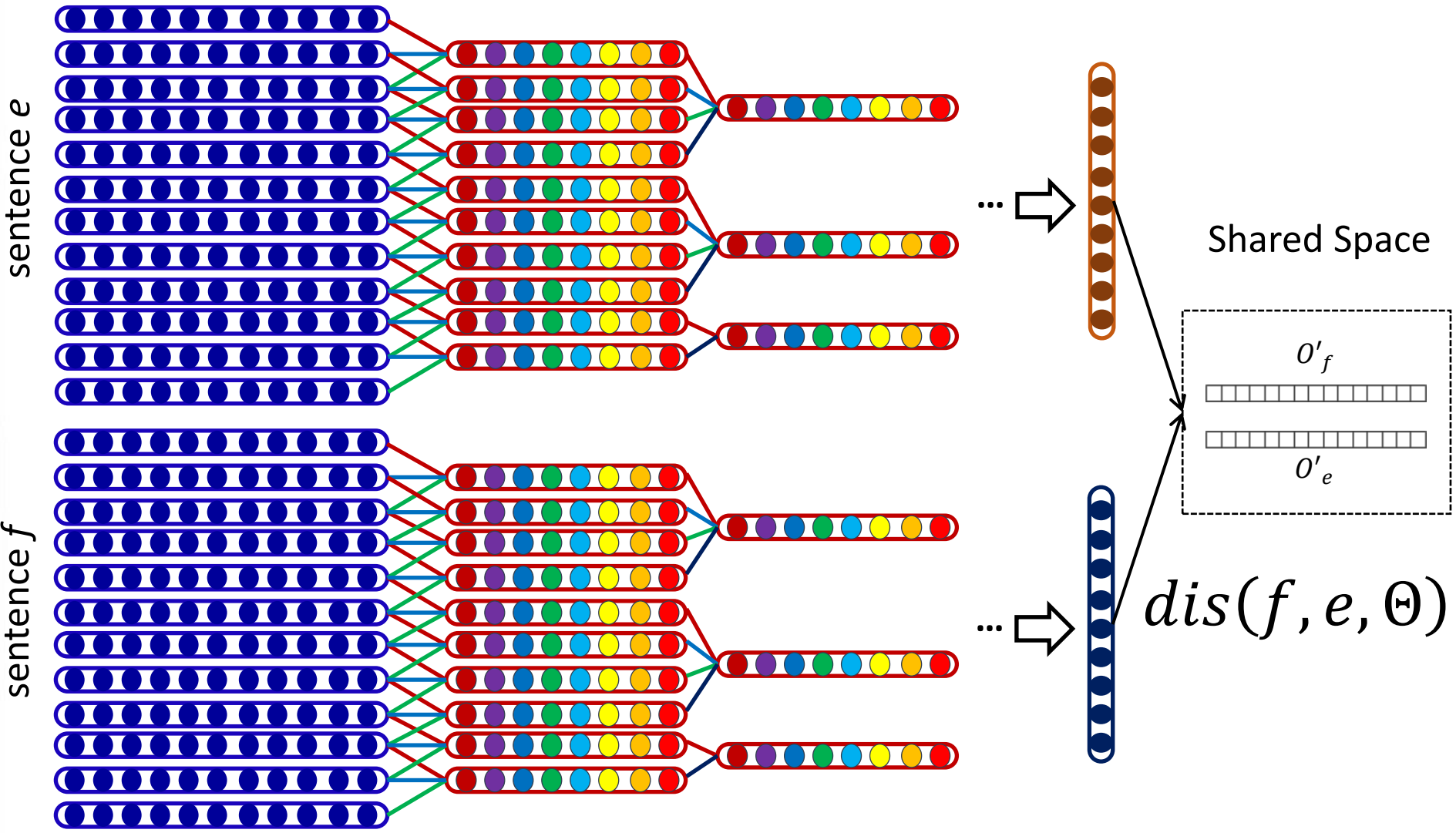
Bilingually-constrained CNN



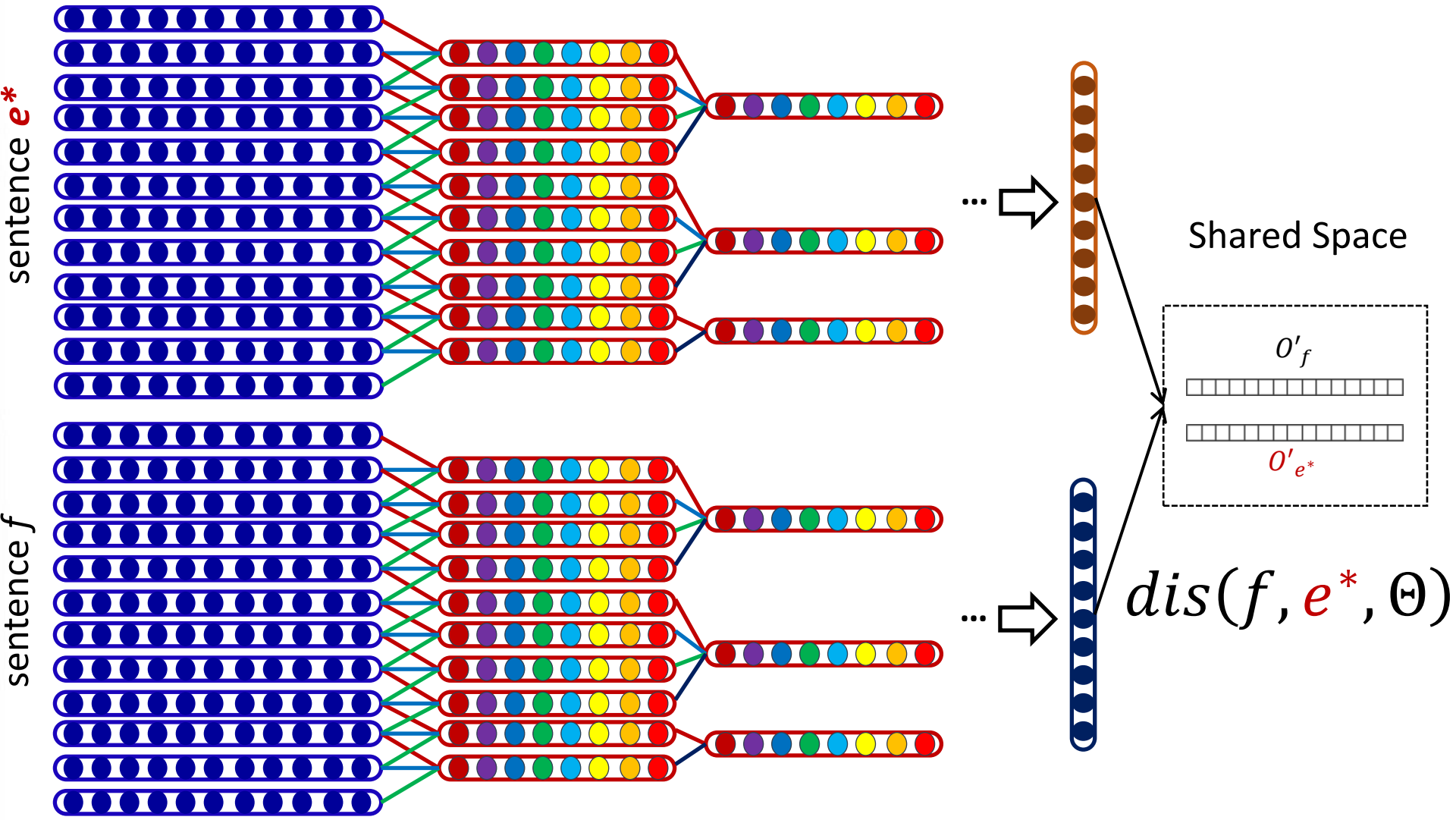
Chunk-based CNN



Max-Margin Training

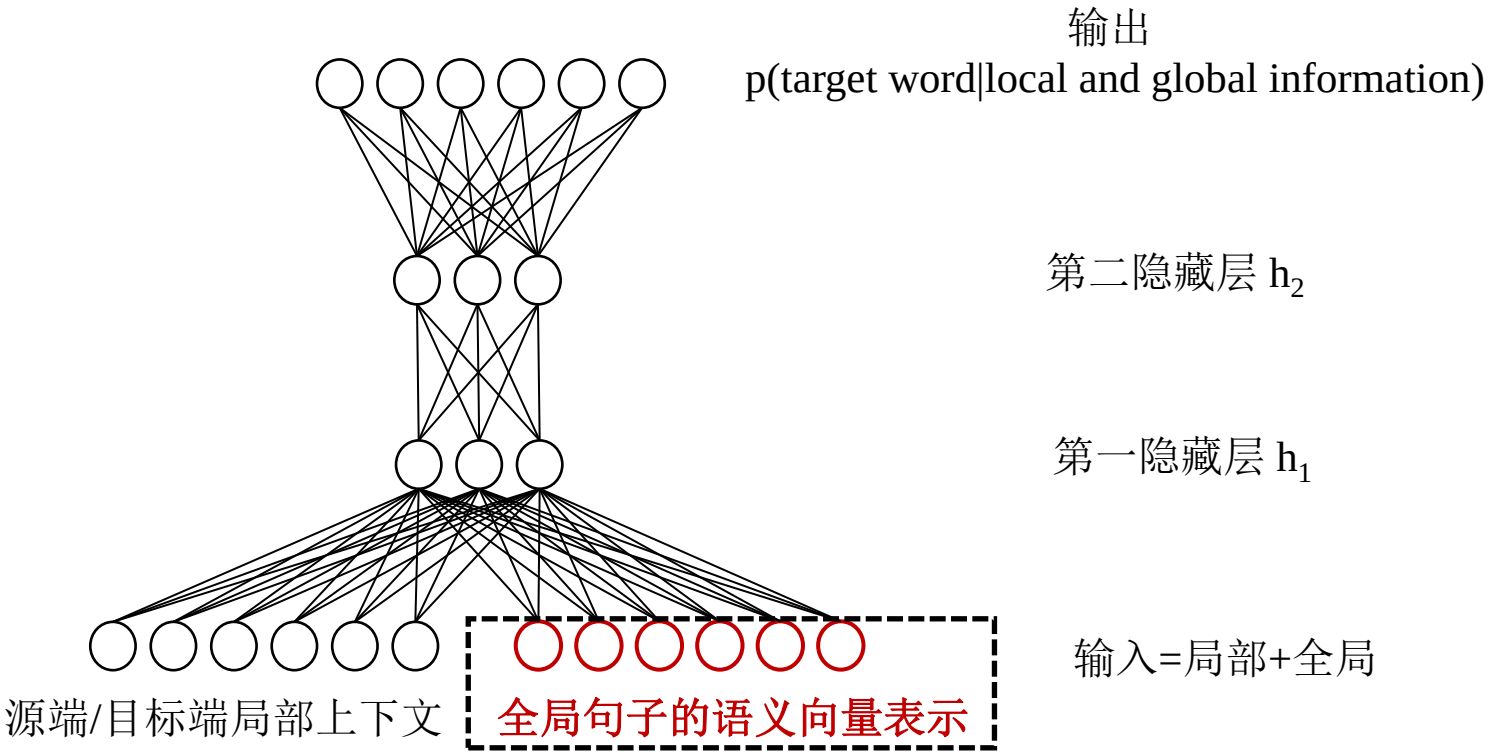


Max-Margin Training



卷积神经网络–统计机器翻译

$$P(e_i) \approx P\left(e_i | e_{i-3} \cdots e_{i-1}, f_{j-c} \cdots f_j \cdots f_{j+c}\right)$$
$$\approx P\left(e_i | e_{i-3} \cdots e_{i-1}, f_{j-c} \cdots f_j \cdots f_{j+c}, \boldsymbol{f}\right)$$

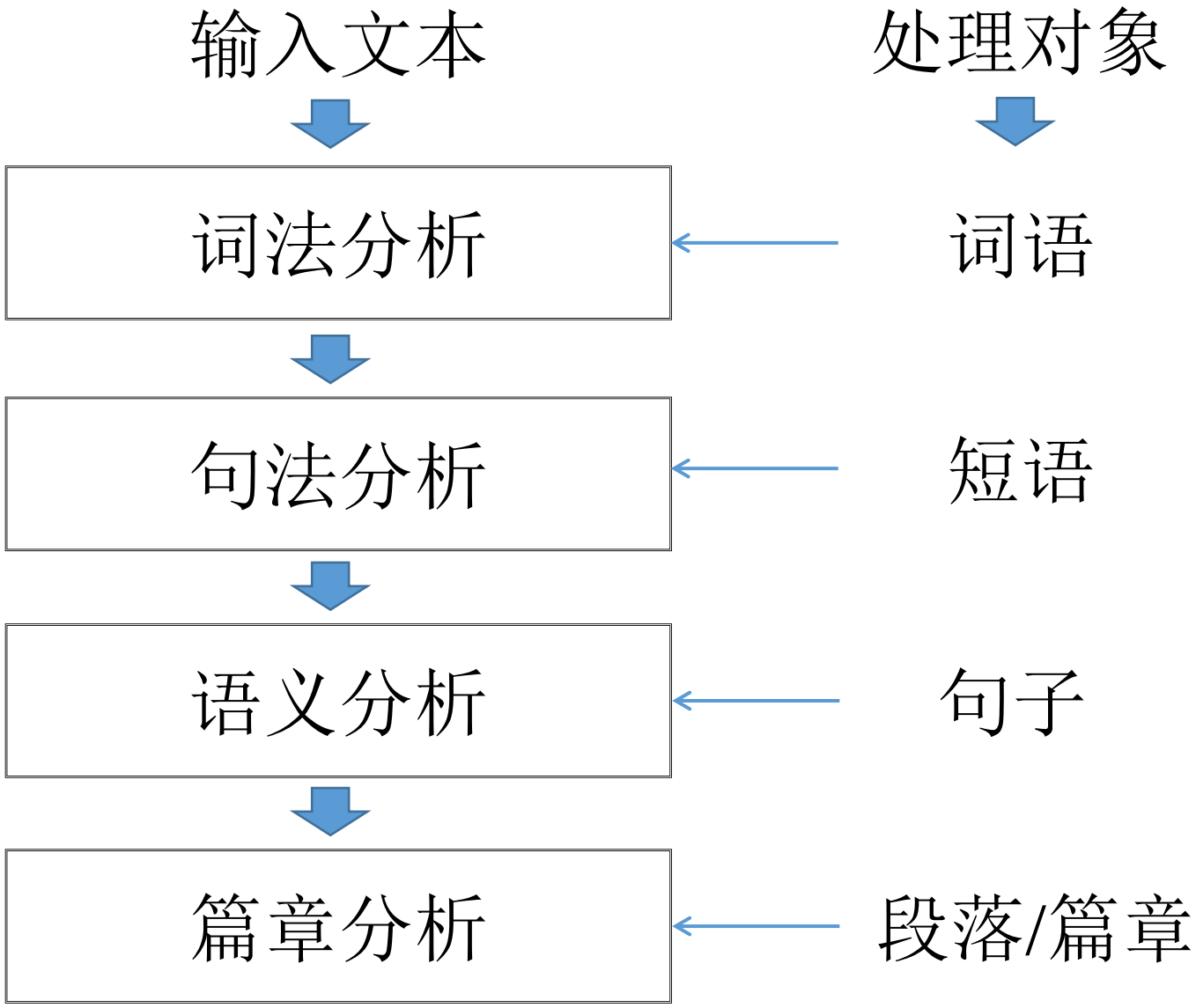


卷积神经网络–统计机器翻译

System	MT03	MT05	MT06	MT08
HPB	35.98	34.66	35.25	27.80
+NNJM	36.93	35.55^{+}	35.77	28.64^{+}
+AVE_SENT	37.16	35.88^{+}	36.07^{+}	29.19^{+}
+BCCNN-1	37.32	36.06^{+}	36.42^{+}	29.35^{+*}
+BCCNN-2	37.75	36.24^{+}	36.65^{+*}	29.97^{+*}
+BCCNN-4	**37.98**	36.22^{+}	$\mathbf{36.78}^{+*}$	$\mathbf{30.02}^{+*}$
+BCCNN-8	37.64	$\mathbf{36.29}^{+*}$	36.49^{+}	29.98^{+*}

超过2 BLEU值的显著提升！

总结



参考文献

1, Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. A neural probabilistic language model. Journal of Machine Learning Research 3:1137–1155.

2, Socher, R.; Pennington, J.; Huang, E. H.; Ng, A. Y.; and Manning, C. D. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proc. of EMNLP, 151–161.

3, Socher, R.; Perelygin, A.; Wu, J. Y.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP.

4, Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proc. of ICML, 160–167.

5, Mikolov, T.; Karafiat, M.; Burget, L.; Cernocky, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In INTERSPEECH, 1045–1048.

6, Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Proc. of NIPS.

7, Devlin J., Zbib R., Huang Z., Lamar T., Schwartz R., and Makhoul J. 2014. Fast and robust neural network joint models for statistical machine translation. In Proc. of ACL, 2014.

8, Zhang, J.; Liu, S.; Li, M.; Zhou, M.; and Zong, C. 2014. Bilingually-constrained phrase embeddings for machine translation. In Proc. of ACL.

9, Zhang, J.; Liu, S.; Li, M.; Zhou, M.; and Zong, C. 2014. Mind the Gap: Machine Translation by Minimizing the Semantic Gap in Embedding Space. In Proc. of AAAI.

10, Zhang, J.; Liu, S.; Li, M.; Zhou, M.; and Zong, C. 2015. Towards Machine Translation in Semantic Vector Space. TALLIP.

11, Zhang, J.; Zhang, D.; and Hao, J. 2015. Local Translation Prediction with Global Sentence Representation. In Proc. of IJCAI.

… …

谢谢!
Q&A