A Deep Architecture for Matching Short Texts

赵惜墨

哈尔滨工业大学
计算机学院
智能技术与自然语言处理实验室

November 21, 2013

从 bilinear 的模型匹配开始：

$$match(x, y) = x^T A y = \sum_{m=1}^{D_x} \sum_{n=1}^{D_y} A_{nm} x_m y_n$$

1. A 是提前计算好的，相当与权重
2. 每一个子元素的乘积 $x_n y_m$ 都可以看作是一个 x和y的 局部决策（local decision）。
3. 上面加和的向量积 $M = xy^T$ 可以看作是 x和y 的部分决策的空间表示。最终的决策考虑了所有的局部决策，因此在 bilinear 中有：$match = \sum_{nm} A_{nm} M_{nm}$，也就是所有局部权重的线性加和。

概念

parallel text 需要匹配的两个文本（问答）对

局部性 在底层用词的共现 (co-occurrence) 来匹配语义。

混合性 决策有不同层次的抽象。局部决策, 抓住词义相近的词之间的关系, 将会逐层形成最后和全局决策。

localness

局部匹配与图像匹配相类似，平行文本的文本块由两个文本之间的关联词决定。像处理图像一样，可以用 $(\Omega_{x,p}, \Omega_{y,p})$ 来决定匹配的范围，$(\Omega_{x,p}, \Omega_{y,p})$ 分别代表 X，Y 的子集。与图像一样，用块“patch”来表示。

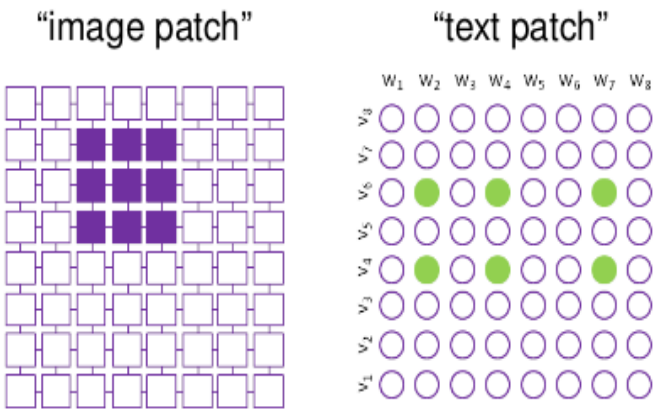


Figure : fig: 图像块 vs 文本块

localness 续

1. 文本块不能用给定的连续空间。因为词不一定和其周围词相关，因此需要通过匹配文本的共现对来发现。
2. 用 bilingual topic model 来发现共现对，这种方法可以成功获取同领域和不同领域的共现对, 这种方法可以成功获取同领域和不同领域的共现对。基本的思想是:
    1. 当词对多次在跨领域出现的时候（例如，感冒——抗体），它们在决定匹配的时候有很高的得分
    2. 当词对多次在相同领域出现的时候（例如，夏威夷——度假），它们对该领域匹配有很高的帮助。
    3. 例如，从 QA 对中就可以发现，夏威夷和 RAM 就不可能作为一个共现对。也就是说，模型只在块中匹配底层次的语义关系。

形成层次性的决策过程

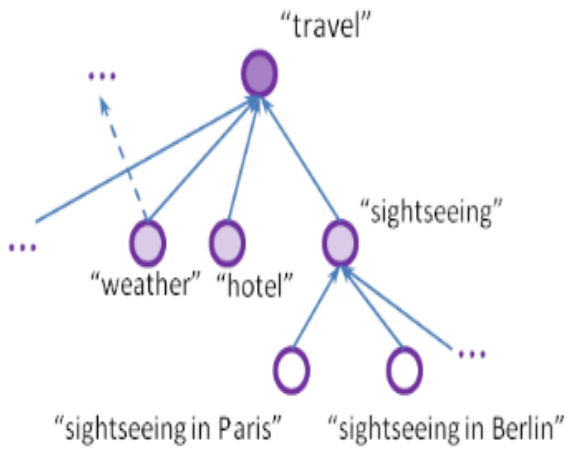


Figure : 层次决策

1. 决策完成之后，大多数词之间的相关性都是没有的。
2. “sightseeing in paris”和“sightseeing in berlin”
3. “sightseeing”来形成一个更高的决策。
4. 也能够相对的形成“hotel”和“transportation”，这些都能形成一个更高的层次“travel”。
5. 注意到高层次主题对底层次主题没有包含关系。

Topic Modeling for Parallel Texts

1. 具体方法是用的 LDA+Gibbs sampling
2. 将问答对放到同一篇文章中，对每一个 topic 建一个词表，避免混合。
3. 允许词表之间有重叠，例如，希望词（hotel，price）能够出现在不同的 topic 中。
4. 按照 topic 数量，找到逐层递减的 topic 集合，
   $H = \{T_1, \cdots, T_L\}$

如何利用 topic

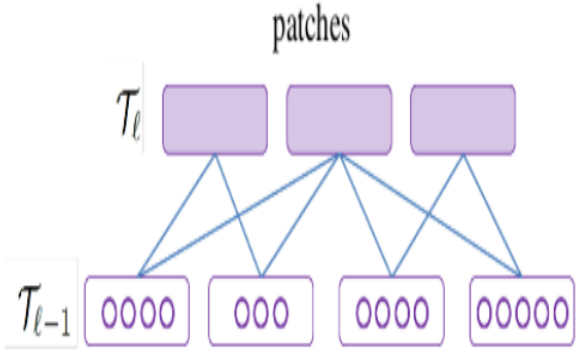
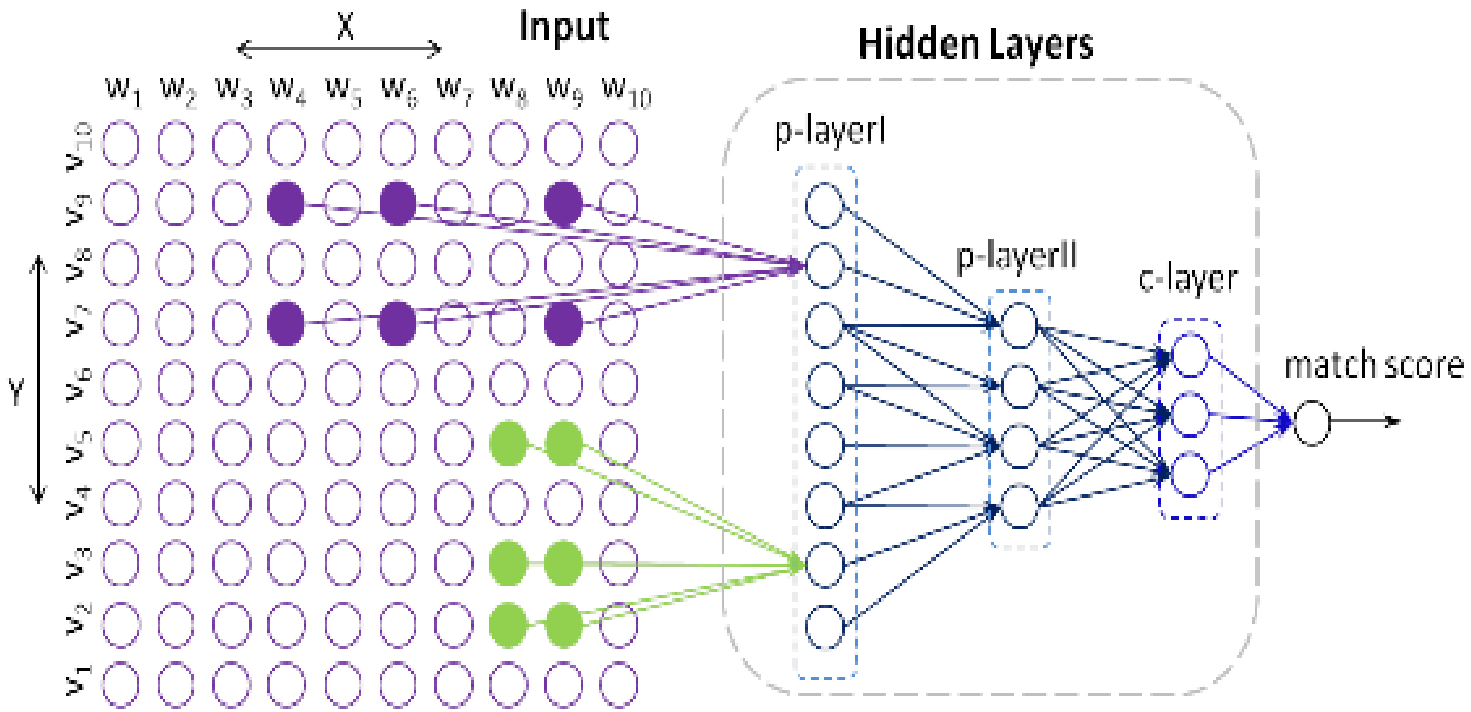


Figure : 层次决策

1. 剪掉在所有主题概率都很低的词。剩下的词在每一个主题确定一个块。
2. 根据 H, 建立一个 DAG G, 根据在 T_l 和 T_{l-1} 共同出现的词, 确定连接。
3. 重复此过程, 建立神经网络。

最终得到的模型



Figure： 网络结构

训练

训练用的 BP

实验

	Question-Answer		Weibo-Response	
	nDCG@1	nDCG@6	nDCG@1	nDCG@6
RANDOM GUESS	0.167	0.550	0.167	0.550
PLS	0.285	0.662	0.171	0.587
RMLS	0.282	0.659	0.165	0.553
SIAMESE NETWORK	0.357	0.735	0.175	0.574
DEEPMATCH	**0.723**	**0.856**	**0.336**	**0.665**

Figure： 实验结果