

工作总结

林连升

September 28, 2017

- **Word Similarity at NLPCC-2016**
- **DBQA at NLPCC-2017**
Document-Based Question Answering
- **STC-2 at NTCIR-13**
Short Text Conversation

Word Similarity

Table 1. The top 10 similar word pairs

word 1	word 2	score
WTO	世界贸易组织	10
紫禁城	故宫	10
计算机	电脑	9.9
赢	胜	9.8
维他命	维生素	9.5
化肥	化学肥料	9.5
课程表	课表	9.5
互联网	因特网	9.5
假货	赝品	9.5

Table 2. The top 10 dissimilar word pairs

word 1	word 2	score
讲价	打架	1
教授	黄瓜	1
控制	恐高	1
阻力	天花板	1
结盟	无理取闹	1.1
调查	努力	1.1
玻璃	魔术师	1.1
干扰	上网	1.1
课堂	美食	1.1

Table 3. The top 10 word pairs with low standard deviation

word 1	word 2	score	Stad.
教授	黄瓜	1	0.00
控制	恐高	1	0.00
阻力	天花板	1	0.00
WTO	世界贸易组织	10	0.00
紫禁城	故宫	10	0.00
讲价	打架	1	0.22
玻璃	魔术师	1.1	0.30
干扰	上网	1.1	0.30
课堂	美食	1.1	0.30

Table 4. The top 10 word pairs with high standard deviation

word 1	word 2	score	Stad
没戏	没辙	4.9	3.03
只管	尽管	4	2.94
GDP	生产力	6.5	2.80
包袱	段子	2.6	2.71
日期	时间	6	2.67
由此	通过	3.4	2.66
爱面子	好高骛远	4	2.56
一方面	一边	5.4	2.54
托福	GRE	8	2.54

Spearman's rank correlation coefficient:

$$\rho = 1 - \frac{6 \sum_{i=1}^n (R_{Xi} - R_{Yi})^2}{n(n^2 - 1)} \quad (1)$$

where n is the number of word pairs being evaluated, R_{Xi} and R_{Yi} are the standard deviations of the rank of automatic computing results and human labelled scores, respectively.

Table 5. The overall evaluation results

Team ID	Organization	Spear.
SXUCFN-QA	Shanxi University	0.518
DLUT_NLPer	Dalian University of Technology	0.457
CQUT_AC996	Chongqing University of Technology	0.436
nlp_polyu	The Hong Kong Polytechnic University	0.421
BLCU_CNLR	Beijing Language and Culture University	0.414
Cbrain	Institute of Automation, Chinese Academy of Sciences	0.412
CIST	Beijing University of Posts and Telecommunications	0.405
wanghao.ftd	Shanghai Jiao Tong University	0.405
DUTNLP	Dalian University of Technology	0.372
BIT_CWSM	Beijing Institute of Technology	0.371
NJUST-CWS	Nanjing University of Science and Technology	0.365
TJIP	Tongji University	0.357
SWJTU_CCIT	Southwest Jiaotong University	0.349

- Word2Vec + 同义词词林 + 知网
Spearman value = 0.58

- 2G 新闻语料 (12 对词未覆盖)
Spearman value = 0.338
- 2G 新闻语料 + 补充语料
Spearman value = 0.400

参考三篇 Paper:

- 1 田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法 [J]. 吉林大学学报 (信息科学版), 2010, 06: 602-608.
- 2 吕立辉, 梁维薇, 冉蜀阳. 基于词林的词语相似度的度量 [J]. 现代计算机 (专业版), 2013, 01: 3-6+9.
- 3 朱新华, 马润聪, 孙柳, 陈宏朝. 基于知网与词林的词语语义相似度计算 [J]. 中文信息学报, 2016, 04: 29-36.

同义词词林简介

Ba01A02= 物质 质 素

Cb02A01= 东南西北 四方

Ba01A03@ 万物

Cb06E09@ 民间

Ba01B08# 固体 液体 气体 流体 半流体

Ba01B10# 导体 半导体 超导体

表 2 词语编码表

编码位	1	2	3	4	5	6	7	8
符号举例	D	a	1	5	B	0	2	= \ # \ @
符号性质	大类	中类	小类		词群	原子词群		
级别	第 1 级	第 2 级	第 3 级		第 4 级	第 5 级		

- 主要思想：基于同义词词林的结构，利用词语中义项的编号，根据两个义项的语义距离，计算出义项相似度。
- 首先判断在同义词林中作为叶子节点的两个义项在哪一层开始分支。例如：Aa01A01 与 Aa01B01，即在第 4 层分支。在每一层分支，分别对应一个系数，然后再乘以调节参数。

若两个义项的相似度用 Sim 表示

1) 若两个义项不在同一棵树上

$$\text{Sim}(A, B) = f \quad (1)$$

2) 若两个义项在同一棵树上:

① 若在第 2 层分支, 系数为 a

$$\text{Sim}(A, B) = 1 \times a \times \cos\left(n \times \frac{\pi}{180}\right) \left(\frac{n - k + 1}{n}\right) \quad (2)$$

④ 若在第 3 层分支, 系数为 b

$$\text{Sim}(A, B) = 1 \times 1 \times b \times \cos\left(n \times \frac{\pi}{180}\right) \left(\frac{n - k + 1}{n}\right) \quad (3)$$

④ 若在第 4 层分支, 系数为 c

$$\text{Sim}(A, B) = 1 \times 1 \times 1 \times c \times \cos\left(n \times \frac{\pi}{180}\right) \left(\frac{n - k + 1}{n}\right) \quad (4)$$

④ 若在第 5 层分支, 系数为 d

$$\text{Sim}(A, B) = 1 \times 1 \times 1 \times 1 \times d \times \cos\left(n \times \frac{\pi}{180}\right) \left(\frac{n - k + 1}{n}\right) \quad (5)$$

Figure 1: 这里 n 是分支数, k 是两个分支的距离, $abcd$ 是实验得到的系数。

举个例子：

Ae05A01= 邮递员 邮差 信差 信使 绿衣使者 通信员 投递员

Ae05A02= 交通员 交通 通讯员

Ae05A03= 联络员 联络官 联系人

其中，义项“邮递员”和“联络员”在第五层分支，分支数 $n=3$ ，
分支距离 $k=2$ ，其相似度为：

$$\begin{aligned}\text{sim} &= d * \cos(n * \pi / 180) * ((n - k + 1) / n) \\ &= 0.96 * \cos(3 * \pi / 180) * ((3 - 2 + 1) / 3) \\ &= 0.639\end{aligned}$$

特殊情况:

如果五层编码都相同, 则靠末尾的 = # @ 来计算。

- = 表示同义, 相似度取 1
- # 表示相关, 相似度取 $e(e = 0.5)$
- @ 表示封闭, 只有一个词。因此不存在两个词编号相同且末尾为 @ 的情况。

Ab02A08@ 老太公

Ab02B01= 成年人 壮年人 大人 人丁 壮丁 佬 中年人

Ab02C01= 老小 老少 大小 老幼 老老少少 白叟黄童 大大小小

Ab02C02# 遗老 遗少 遗老遗少 封建残余

- 1. 考虑路径长度信息

$$g_1(l) = e^{-\alpha l}$$

α 为待定系数, l 为距离。

比如“体制”这个单词对应的编码为“Di09D01=”，还有“阶级”这个单词对应的编码为“Di06A01=”，两个词的公共祖先在第二层，在第三层产生分支，则两个词在词林树中相隔的路径长就为 6。一个词可能有多个编码，这里取最短距离。

- 2. 考虑密度信息

- 密度信息为 d

$$d = -\log \frac{\text{freq}(c)}{N}$$

$$g_2(d) = \frac{e^{\beta d} - e^{-\beta d}}{e^{\beta d} + e^{-\beta d}}$$

- 其中 $\text{freq}(c)$ 为两个词所在的分支之间所有分支所包含的词
的个数, N 为词林中单词的总数。
- β 为待定系数, 使用双曲正切函数是为了使 g_2 的范围在
[0,1]

- 3. 最终相似度为

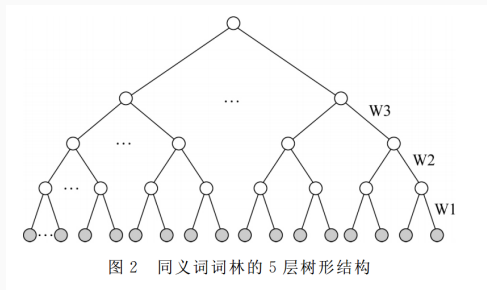
$$sim = \delta * g_1(l) + (1 - \delta) * g_2(d)$$

其中 δ , $1 - \delta$ 分别为 g_1 , g_2 的权重参数, $\delta \in [0, 1]$ 。

- 该方法是改进（田久乐等，2010）的算法。
- （田久乐等，2010）的算法是以分支节点数 n 和分支间隔 k 为主要考虑因素，因此会出现许多距离近的词语因分支间隔远而算出相似度过低的不合理现象。为解决这一问题，朱新华等提出了一个以词语距离 d 为主要影响因素、分支节点数 n 和分支间隔 k 为调节参数的计算公式。

$$\text{sim}(C_1, C_2) = (1.05 - 0.05 \text{dis}(C_1, C_2)) \sqrt{e^{\frac{-k}{2n}}}$$

- 其中, $dis(C_1, C_2)$ 是词语编码 C_1, C_2 在树状结构中的距离函数。 W_1 、 W_2 、 W_3 、 W_4 分别为 0.5、1、2.5、2.5。因此, 距离有 1、3、8、13 几种情况。



三种词林方法 + Word2Vec 的结果

- 田久乐等, 2010 + Word2vec Spearman value = 0.518
- 吕立辉等, 2013 + Word2vec Spearman value = 0.450
- 朱新华等, 2016 + Word2vec Spearman value = 0.538

超过了第一名的 0.518

《知网》简介

- 《知网》是一部详尽的语义知识词典。
- 《知网》的结构
 - “概念”，是用一种“知识描述语言”来描述的，可以理解为一个词的一种意思。
 - “义原”，是用于描述一个“概念”的最小语义单位。

表2：《知网》知识描述语言实例

词	概念编号	描述语言
打	017144	exercise 锻炼,sport 体育
男人	059349	human 人,family 家,male 男
高兴	029542	aValue 属性值,circumstances 境况,happy 福,desired 良
生日	072280	time 时间,day 日,@ComeToWorld 问世,\$congratulate 祝贺
写信	089834	write 写,ContentProduct=letter 信件
北京	003815	place 地方,capital 国都,ProperName 专,(China 中国)
爱好者	000363	human 人,*FondOf 喜欢,#WhileAway 消闲
必须	004932	{modality 语气}
串	015204	NounUnit 名量,&(grape 葡萄),&(key 钥匙)

义原的分类

- 《知网》一共采用了个 1500 义原，这些义原分为以下几个大类：
 1. Event| 事件
 2. entity| 实体
 3. attribute| 属性
 4. aValue| 属性值
 5. quantity| 数量
 6. qValue| 数量值
 7. SecondaryFeature| 次要特征
 8. syntax| 语法
 9. EventRole| 事件角色
 10. EventFeatures| 事件属性
- 这 10 类义原又分为三组：
 - 1-7 基本义原
 - 8 语法义原
 - 9、10 关系义原

义原的层次结构

- 在《知网》中，一共描述了义原之间的 8 种关系：上下位关系、同义关系、反义关系、对义关系、属性-宿主关系、部件-整体关系、材料-成品关系、事件-角色关系。
- 根据义原中最重要的上下位关系，所有的“基本义原”组成了一个义原层次体系。这个义原层次体系是一个树状结构，也是《知网》进行语义相似度计算的基础。

```
- entity|实体
  卜 thing|万物
  ... 卜 physical|物质
        ... 卜 animate|生物
              ... 卜 AnimalHuman|动物
                    ... 卜 human|人
                          | 卜 humanized|拟人
                          | 卜 animal|兽
                              卜 beast|走兽
                              ...
```

图2 树状的义原层次结构

- 《知网》收录的词语主要分为两类，一类是实词，一类是虚词
- 虚词的描述比较简单，用“句法义原”或“关系义原”进行描述
- 实词的描述比较复杂，由一系列用逗号隔开的“语义描述式”组成，这些“语义描述式”又有以下三种形式：
 - 基本义原描述式：用“基本义原”进行描述；
 - 关系义原描述式：用“关系义原 = 基本义原”或者“关系义原 = (具体词)”或者“(关系义原 = 具体词)”来描述；
 - 关系符号描述式：用“关系符号 基本义原”或者“关系符号 (具体词)”加以描述。

表2：《知网》知识描述语言实例

词	概念编号	描述语言
打	017144	exercise 锻炼,sport 体育
男人	059349	human 人,family 家,male 男
高兴	029542	aValue 属性值,circumstances 境况,happy 福,desired 良
生日	072280	time 时间,day 日,@ComeToWorld 问世,\$congratulate 祝贺
写信	089834	write 写,ContentProduct=letter 信件
北京	003815	place 地方,capital 国都,ProperName 专,(China 中国)
爱好者	000363	human 人,*FondOf 喜欢,#WhileAway 消闲
必须	004932	{modality 语气}
串	015204	NounUnit 名量,&(grape 葡萄),&(key 钥匙)

知识描述语言的符号及其含义

表1: 《知网》知识描述语言中的符号及其含义

,	多个属性之间, 表示“和”的关系
#	表示“与其相关”
%	表示“是其部分”
\$	表示“可以被该‘V’处置, 或是该“V”的受事, 对象, 领有物, 或者内容
*	表示“会‘V’或主要用于‘V’, 即施事或工具
+	对V类, 它表示它所标记的角色是一种隐性的, 几乎在实际语言中不会出现
&	表示指向
~	表示多半是, 多半有, 很可能的
@	表示可以做“V”的空间或时间

基于《知网》的词语语义相似度计算方法

- 词语相似度计算
- 义原相似度计算
- 虚词概念的相似度计算
- 实词概念的相似度计算

对于两个汉语词语 W_1 和 W_2 ，如果 W_1 有 n 个概念： $S_1^1, S_2^1, \dots, S_n^1$ ， W_2 有 m 个概念： $S_1^2, S_2^2, \dots, S_m^2$ ，规定它们的相似度是各个概念的相似度之最大值：

$$Sim(W_1, W_2) = \max_{i=1 \dots n, j=1 \dots m} Sim(S_i^1, S_j^2) \quad (2)$$

- 由于所有的概念都最终归结于用义原来表示，所以义原的相似度计算是概念相似度计算的基础。
- 假设两个义原在树状的层次体系中的路径距离为 d ，我们可以得到这两个义原之间的相似度为：

$$Sim(p_1, p_2) = \frac{\alpha}{d + \alpha} \quad (3)$$

- 我们认为，在实际的文本中，虚词和实词总是不能互相替换的，因此，虚词概念和实词概念的相似度总是为零。
- 由于虚词概念总是用“语法义原”或“关系义原”这两种方式进行描述，所以，虚词概念的相似度计算归结为，计算其对应的“句法义原”或“关系义原”之间的相似度即可。

集合的相似度计算

特征结构的相似度计算

实词概念的描述可以表示为一个特征结构，包括以下四个特征：

- 第一基本义原描述：其值为一个基本义原，这一部分的相似度记为 $Sim_1(S_1, S_2)$ ；
- 其它基本义原描述：除第一基本义原描述式以外的所有基本义原描述式，其值为一个基本义原的集合，这一部分的相似度记为 $Sim_2(S_1, S_2)$ ；
- 关系义原描述：对应所有的关系义原描述式，其值是一个特征结构，对于该特征结构的每一个特征，其属性是一个关系义原，其值是一个基本义原，或一个具体词。这一部分的相似度记为 $Sim_3(S_1, S_2)$ ；
- 关系符号描述：对应所有的关系符号描述式，其值也是一个特征结构，对于该特征结构的每一个特征，其属性是一个关系义原，其值是一个集合，该集合的每个元素是一个基本义原，或一个具体词。这一部分的相似度记为 $Sim_4(S_1, S_2)$ 。

- 两个概念的总体相似度为：

$$Sim(S_1, S_2) = \sum_{i=1}^4 \beta_i Sim_i(S_1, S_2) \quad (4)$$

- 其中， $\beta_i (1 \leq i \leq 4)$ 是可调节的参数，且有：
 $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$
- 后者反映了 Sim_1 到 Sim_4 对于总体相似度所起到的作用依次递减。

Open Domain Question Answering

- **NLPCC**

Natural Language Processing and Chinese Computing

- **Open Domain Question Answering**

including three tasks:

- Knowledge-Based Question Answering (KBQA)
- **Document-Based Question Answering (DBQA)**
- Table-Based Question Answering (TBQA)

The task of DBQA is to answer Chinese questions by selecting one or multiple sentences from a given document as answers.

俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖, 中国古代称为北海, 位于俄罗斯西伯利亚的南部。 \t 0
俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖是世界上最深, 容量最大的淡水湖。 \t 0
俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖贝加尔湖是世界上最深和蓄水量最大的淡水湖。 \t 0
俄罗斯贝加尔湖的面积有多大? \t 它位于布里亚特共和国 (Buryatiya) 和伊尔库茨克州 (Irkutsk) 境内。 \t 0
俄罗斯贝加尔湖的面积有多大? \t 湖型狭长弯曲, 宛如一弯新月, 所以又有“月亮湖”之称。 \t 0
俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖长 636 公里, 平均宽 48 公里, 最宽 79.4 公里, 面积 3.15 万平方公里。 \t 1
俄罗斯贝加尔湖的面积有多大? \t 贝加尔湖湖水澄澈清冽, 且稳定透明 (透明度达 40.8 米), 为世界第二。 \t 0

Figure 2: An example for DBQA

Method

1. Find question word in question

俄罗斯 贝加尔湖 的 面积 有 多大

2. Remove stopwords

俄罗斯 贝加尔湖 面积 多大

3. Assign weight for each word in question

$w_i = 2^{-distance(word_i, word_q)}$ $word_q$: question word

4. Scoring sentence A

$$score(A) = \sum_{i=1}^l s_i, \quad s_i = \begin{cases} w_i & word_i \in A \\ 0 & word_i \notin A \end{cases}$$

l is the number of the words in question.

Final Evaluation Results: Document-based QA Task

Rank	Team	TeamID	MRR	MAP	ACC@1
1	复旦大学	KEngine	0.720194	0.716594	0.592
2	北京邮电大学	Prisers	0.689619	0.68576	0.5556
3	同济大学	CU-KG	0.685011	0.680963	0.5512
4	DeepIntell	run2	0.683674	0.680067	0.5492
5	天津大学	TJUNLP	0.677203	0.673271	0.54
6	DeepIntell	run3	0.675772	0.670828	0.5356
7	DeepIntell	primary	0.672872	0.668659	0.5372
8	DeepIntell	run4	0.664586	0.660256	0.5244
9	华中师范大学	CCNU-NLP-Blaze	0.660674	0.658893	0.5144
10	北京大学	ICL-WLL	0.652062	0.649218	0.5056
11	浙江大学	CS-NLP	0.583311	0.580741	0.4284
12	网龙网络有限公司	Nders	0.557158	0.556341	0.3996
13	北京航空航天大学	BHU	0.54846	0.545021	0.372
14	国防科技大学	FuRongWang	0.533575	0.531831	0.3692
15	华东师范大学	ECNU	0.506718	0.503114	0.3404
16	北京联合大学	TensorRQ	0.494292	0.491736	0.3288
17	北京大学	name_system	0.436557	0.434162	0.2696
18	大连理工大学-1	DLUT_NLPer	0.402115	0.40085	0.2172
19	大连理工大学-2	DLUT_NLPer	0.384112	0.382343	0.2016
20	大连理工大学-3	DLUT_NLPer	0.384112	0.382343	0.2016
21	重庆大学	CQUT_AC996	0.353259	0.352269	0.1744

Figure 3: Result for DBQA

Short Text Conversation

- **NTCIR**

NII Testbeds and Community for Information access Research

- **STC**

Short Text Conversation

which consists of two subtask:

- Chinese task
- Japanese task

IR-based method

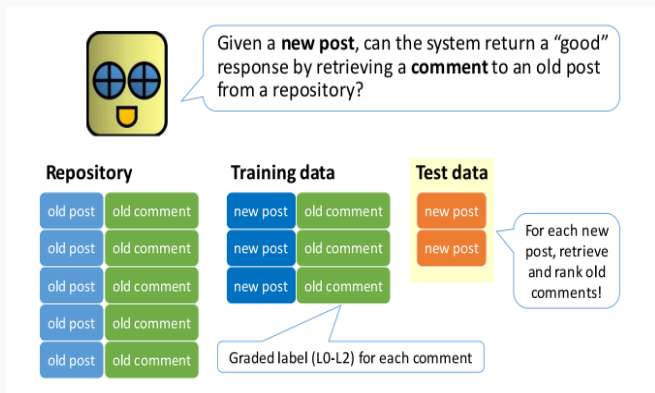


Figure 4: IR-based method

To make the annotation task operable, the appropriateness of retrieved comments is judged from the following four criteria:

1. Fluent
2. Coherent: logically and topically relevant
3. Self-sufficient
4. Substantial

If either (1) or (2) is untrue, the retrieved comment should be labeled "L0"; if either (3) or (4) is untrue, the label should be "L1"; otherwise, the label is "L2".

Post	意大利禁区里老是八个人...太夸张了吧 There are always 8 Italian players in their own restricted area...Unbelievable!	Related Criteria	Labels
Comment1	我是意大利队的球迷，等待比赛开始。 I am a big fan of the Italy team, waiting for the football match to start	(2) Coherent	L0
Comment2	意大利的食物太美味了 Italian food is absolutely delicious.	(2) Coherent	L0
Comment3	太夸张了吧! Unbelievable!	(4) Substantial	L1
Comment4	哈哈哈哈哈仍然是0: 0 。还没看到进球。 Haha, it is still 0:0, no goal so far.	(3) Self-sufficient	L1
Comment5	这正是意大利式防守足球。 This is exactly the Italian defending style football game	——	L2

Figure 5: Example of a post and its five candidate comments with human annotation. The content of the post implies that the football match has already started, while the author of Comment 1 is still waiting for the match to start. Comment 2 talks about the food of Italy. Comment 3 is a widely used response, but it is appropriate for this post. Comment 4 states that the current score is still 0:0, it is an appropriate comment only for this specific scenario.

1. Preprocessing
2. Candidates Generation
3. Ranking

- Traditional-Simplified Conversion
- Full-width to half-width
- Replace with token <_NUM> <_TIME> <_URL>
- Word Segmentation
- Filter meaningless words and special symbols.

Similarity Feature

- LSA (Latent Semantic Analysis)
- LDA (Latent Dirichlet allocation)
- Word2Vec
- LSTM-Sen2Vec

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (8)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t * \tanh(C_t) \quad (10)$$

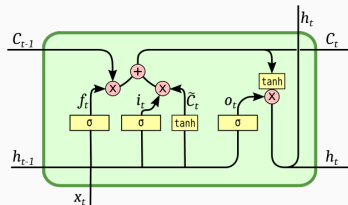


Figure 6: The repeating module in an LSTM contains four interacting layers.

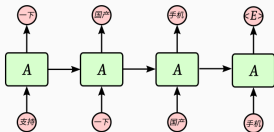


Figure 7: The unidirectional LSTM

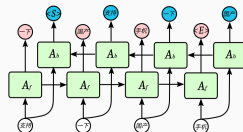


Figure 8: The Traditional Bidirectional LSTM

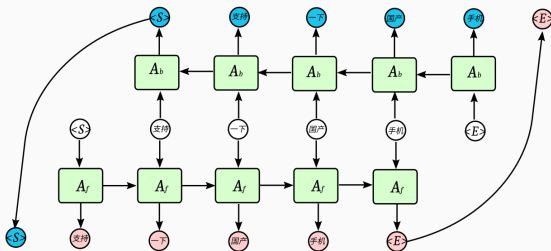


Figure 9: The Modified Bidirectional LSTM

- Similar Posts

$$Score_{q,p}^1(q, p) = Sim_{LDA}(q, p) * Sim_{W2V}(q, p) * Sim_{LSTM}(q, p) \quad (11)$$

$$Score_{q,p}^2(q, p) = Sim_{LSA}(q, p) * Sim_{W2V}(q, p) * Sim_{LSTM}(q, p) \quad (12)$$

- Comment Candidates

$$Score_{q,c}^1(q, c) = Sim_{LSA}(q, c) * Sim_{W2V}(q, c) \quad (13)$$

$$Score_{q,c}^2(q, c) = Sim_{LDA}(q, c) * Sim_{W2V}(q, c) \quad (14)$$

- TextRank
- Pattern-IDF
- TextRank + Pattern-IDF

Ranking

Inspired by PageRank.

For N candidates, we construct $N \times N$ matrix C .

$C_{ij} = \text{sim}(\text{candidate}_i, \text{candidate}_j)$. Then we compute iteratively

$$R(t+1) = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \dots\dots\dots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} C_{11} & C_{12} & C_{13} & \dots & C_{1N} \\ C_{21} & C_{22} & C_{23} & \dots & C_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{N1} & C_{N2} & C_{N3} & \dots & C_{NN} \end{bmatrix} R(t)$$

$$R(0) = [1/N \ 1/N \ \dots \ 1/N]^T$$

Stop when $|R(t+1) - R(t)| < \epsilon$

Questions?