相关问题检索

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

December 11, 2013

例子

Query: *Q1: Any cool clubs in Berlin or Hamburg?*
Expected: *Q2: What are the best/most fun clubs in Berlin?*
Not Expected: *Q3: Any nice hotels in Berlin or Hamburg?* *Q4: How long does it take to Hamburg from Berlin?* *Q5: Cheap hotels in Berlin?*

Figure : 问句检索相关例子

基本方法

- VSM(vector space model)
- LM(language model)
- Translation Model(the state of the art)

这周看的

- Cao et al., 2010 提出了一个识别 cQA 中通用问题的框架，将问句匹配分为全局匹配和局部匹配两个子问题，从而提升了结果。
- Duan et al., 2008 提出了一种基于 MDL（最小描述距离）的方法，将问句转化为对于 question topic 和 question focus 两个子问题的匹配方法。
- Cai et al., 2011 提出了将 LDA 和问句匹配相联合的方法。

VSM model

$$S_{\mathbf{q},\mathbf{d}} = \frac{\sum_{t \in \mathbf{q} \bigcap \mathbf{d}} w_{\mathbf{q},t} w_{\mathbf{d},t}}{W_{\mathbf{q}} W_{\mathbf{d}}}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln(1 + \frac{N}{f_t}), \; w_{\mathbf{d},t} = 1 + \ln(tf_{t,\mathbf{d}})$$

$$W_{\mathbf{q}} = \sqrt{\sum_t w_{\mathbf{q},t}^2}, \; W_{\mathbf{d}} = \sqrt{\sum_t w_{\mathbf{d},t}^2}$$

Figure : VSM model

Okapi BM25 Model

$$S_{\mathbf{q},\mathbf{d}} = \sum_{t \in \mathbf{q} \bigcap \mathbf{d}} w_{\mathbf{q},t} w_{\mathbf{d},t}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln\left(\frac{N - f_t + 0.5}{f_t + 0.5}\right) \frac{(k_3 + 1) tf_{t,\mathbf{q}}}{k_3 + tf_{t,\mathbf{q}}}$$

$$w_{\mathbf{d},t} = \frac{(k_1 + 1) tf_{t,\mathbf{d}}}{K_{\mathbf{d}} + tf_{t,\mathbf{d}}}$$

$$K_{\mathbf{d}} = k_1\left((1 - b) + b \frac{W_{\mathbf{d}}}{W_A}\right)$$

Figure : Okapi BM25 Model

解决了 VSM 偏向于选择短问题的问题。

Language Model

$$S_{\mathbf{q},\mathbf{d}} = \prod_{t \in \mathbf{q}} ((1-\lambda) P_{ml}(t|\mathbf{d}) + \lambda P_{ml}(t|\mathbf{Coll})), \text{ where}$$

$$P_{ml}(t|\mathbf{d}) = \frac{tf_{t,\mathbf{d}}}{\sum_{t' \in \mathbf{d}} tf_{t',\mathbf{d}}}$$

$$P_{ml}(t|\mathbf{Coll}) = \frac{tf_{t,\mathbf{Coll}}}{\sum_{t' \in \mathbf{Coll}} tf_{t',\mathbf{Coll}}}$$

Figure : Language Model

Category Enhanced Retrieval Model

目标函数

$$RS_{q,d} = (1-\alpha)N(S_{q,d}) + \alpha N(S_{q,cat(d)})$$

Global Relevance

- 一种非常 naive 的 idea 是直接将类别的词作为一个大的文章进行计算
- 由于类别间的词表长度差异非常大（467-69789），这样做会直接导致归一化系数占主导地位

$$S_{\mathbf{q},cat(\mathbf{d})} = \frac{\sum_{t \in \mathbf{q} \cap cat(\mathbf{d})} w_{\mathbf{q},t} w_{cat(\mathbf{d}),t}}{W_{\mathbf{q}}}, \text{ where}$$

$$w_{\mathbf{q},t} = \ln(1 + \frac{M}{fc_t}), w_{cat(\mathbf{d}),t} = 1 + \frac{1}{\ln(\frac{W_{cat(\mathbf{d})}}{tf_{t,cat(\mathbf{d})}})}$$

Figure : Global Relevance for VSM

其他模型也做了类似的调整。

Local Relevance

对于局部匹配，只根据类别内部计算 IDF。

Result Analysis

	VSM	OptC	QC	VSM+VSM	%chg	Okapi+VSM	%chg	LM+VSM	%chg	TR+VSM	%chg	TRLM+VSM	%chg
MAP	0.2407	0.2414	0.2779	**0.3711**	54.2*	0.3299	37.1*	0.3632	50.9*	0.3629	50.8*	0.3628	50.7*
MRR	0.4453	0.4534	0.4752	**0.5637**	26.6*	0.5314	19.3*	0.5596	25.7*	0.5569	25.1*	0.5585	25.4*
R-Prec	0.2311	0.2298	0.2568	**0.3419**	48.0*	0.3094	33.9*	0.3366	45.7*	0.3346	44.8*	0.3357	45.3*
P@5	0.2222	0.2289	0.2436	**0.2789**	25.5*	0.2559	15.2*	0.2746	23.6*	0.2746	23.6*	0.2753	23.9*

Table 1: VSM vs. CE with VSM for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value $<$ 0.05)

	Okapi	OptC	QC	VSM+Okapi	%chg	Okapi+Okapi	%chg	LM+Okapi	%chg	TR+Okapi	%chg	TRLM+Okapi	%chg
MAP	0.3401	0.2862	0.3622	0.4007	17.8*	0.3977	16.9*	**0.4138**	21.7*	0.4082	20.0*	0.4132	21.5*
MRR	0.5406	0.4887	0.5713	0.6131	13.4*	0.5884	8.8	0.6214	15.0*	0.6172	14.2*	**0.6215**	15.0*
R-Prec	0.3178	0.2625	0.3345	0.3648	14.8*	0.3613	13.7*	0.3758	18.3*	0.3677	15.7*	**0.3762**	18.4*
P@5	0.2857	0.2824	0.2998	0.3140	9.9*	**0.3176**	11.2*	0.3161	10.6*	0.3111	8.8	0.3147	10.2*

Table 2: Okapi vs. CE with Okapi for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value $<$ 0.05)

	LM	OptC	QC [5]	LM+L [5]	VSM+LM	%chg	Okapi+LM	%chg	LM+LM	%chg	TR+LM	%chg	TRLM+LM	%chg
MAP	0.3821	0.3402	0.4083	0.4586	**0.4620**	20.9*	0.4599	20.4*	0.4609	20.6*	0.4603	20.5*	0.4616	20.8*
MRR	0.5945	0.5219	0.6083	0.6620	0.6630	11.5*	0.6651	11.9*	0.6622	11.4*	0.6633	11.6*	**0.6667**	12.1*
R-Prec	0.3404	0.3129	0.3624	0.4072	**0.4101**	20.5*	0.4079	19.8*	0.4087	20.1*	0.4087	20.1*	0.4100	20.4*
P@5	0.3040	0.2810	0.3230	0.3460	0.3512	15.5*	0.3498	15.1*	**0.3519**	15.8*	0.3512	15.5*	0.3513	15.6*

Table 3: LM vs. CE with LM for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value $<$ 0.05)

	TR	OptC	QC	VSM+TR	%chg	Okapi+TR	%chg	LM+TR	%chg	TR+TR	%chg	TRLM+TR	%chg
MAP	0.4010	0.3417	0.4125	**0.4592**	14.5*	0.4528	12.9*	0.4507	12.4*	0.4526	12.9*	0.4522	12.8*
MRR	0.6084	0.5392	0.6178	**0.6607**	8.6	0.6532	7.4	0.6527	7.3	0.6552	7.7	0.6540	7.5
R-Prec	0.3717	0.3175	0.3853	**0.4153**	11.7*	0.4079	9.7*	0.4054	9.1	0.4071	9.5*	0.4058	9.2
P@5	0.3168	0.2670	0.3280	0.3505	10.6*	**0.3519**	11.1*	0.3505	10.6*	0.3497	10.4*	0.3490	10.2*

Table 4: TR vs. CE with TR for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value $<$ 0.05)

	TRLM	OptC	QC	VSM+TRLM	%chg	Okapi+TRLM	%chg	LM+TRLM	%chg	TR+TRLM	%chg	TRLM+TRLM	%chg
MAP	0.4369	0.3645	0.4570	**0.4937**	13.0*	0.4823	10.4*	0.4836	10.7*	0.4876	11.6*	0.4886	11.8*
MRR	0.6316	0.5506	0.6551	**0.6704**	6.1	0.6652	5.3	0.6675	5.6	0.6685	5.8	0.6678	5.7
R-Prec	0.4008	0.3474	0.4196	**0.4407**	10.0	0.4349	8.5	0.4319	7.8	0.4331	8.1	0.4343	8.4
P@5	0.3398	0.2910	0.3487	**0.3570**	5.1	0.3556	4.6	0.3541	4.2	0.3548	4.4	0.3527	3.8

Table 5: TRLM vs. CE with TRLM for computing local relevance (%chg denotes the performance improvement in percent of each model in CE; * indicates a statistically significant improvement over the baseline using the t-test, p-value $<$ 0.05)

Figure : Result Analysis

Result Analysis - Global Relevance

TR and TRLM 在计算全局相关度时，不如 LM 好，在 baseline 的情况，基于翻译的模型能够找到语义相同的不同词，但是在类内计算全局相关度时，在相关类别上，一般词表都差不多，减弱了语义不同词的影响。

intuition

Query: *Q1: Any cool clubs in Berlin or Hamburg?*
Expected: *Q2: What are the best/most fun clubs in Berlin?*
Not Expected: *Q3: Any nice hotels in Berlin or Hamburg?* *Q4: How long does it take to Hamburg from Berlin?* *Q5: Cheap hotels in Berlin?*

Figure : intuition

intuition 续

- 对于问句 Any clubs in Berlin or Hamburg?
- 将问句分成两部分：

  question topic Berlin Hamburg
  question focus clubs
- 在选词的时候，选定一些基本的词 (BaseNP - Base Noun Phrase)，和一些疑问词开头的句式 (WH-ngram)。

Topic Chain

$$p(c|t) = \frac{count(c,t)}{\sum_{c \in C} count(c,t)}$$

对于词 t 和类别 c。

specificity

$$s(t) = \frac{1}{-\sum_{c \in C} p(c|t) \log(p(c|t)) + \epsilon}$$

specificity 越大表明词的区分度越大。

构建 topic chain

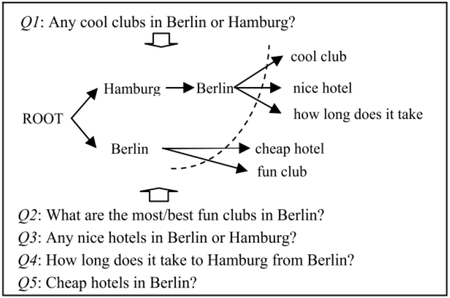

Figure : question tree

tree cut

- 通过 tree cut 的方式将问句 q 分成两部分：$H(q), T(q)$。
- H(q) 为 question topic
- T(q) 为 question focus

检索

$$p(q|\tilde{q}) = \lambda \cdot p(H(q)|H(\tilde{q})) + (1-\lambda) \cdot p(T(q)|T(\tilde{q}))$$

Figure : 检索

$$p(q|\tilde{q}) = \lambda \cdot \prod_{t \in H(q)} p(t|H(\tilde{q}))^{count(q,t)} + (1-\lambda) \cdot \prod_{t \in T(q)} p(t|T(\tilde{q}))^{count(q,t)}$$

Figure : 检索

Result

Methods	Results
VSM	1. How cold does it usually get in Charlotte, NC during winters? 2. How long and cold are the winters in Rochester, NY? **3. How cold is it in Alaska?**
LMIR	**1. How cold is it in Alaska?** 2. How cold does it get really in Toronto in the winter? 3. How cold does the Mojave Desert get in the winter?
LMIR-CUT	**1. How cold is it in Alaska?** **2. How cold is Alaska in March and out-door activities?** 3. How cold does it get in Nova Scotia in the winter?

Table 4. Search Results for
"How cold does it get in winters in Alaska?"

Figure : 检索结果

Result2

Methods	MAP	R-Precision	MRR
VSM	0.198	0.138	0.228
LMIR	0.203	0.154	0.248
LMIR-CUT	**0.236**	**0.192**	**0.279**

Table 3. Searching Questions about 'Travel'

Figure : 检索结果

Methods	MAP	R-Precision	MRR
VSM	0.236	0.175	0.289
LMIR	0.248	0.191	0.304
LMIR-CUT	**0.279**	**0.230**	**0.341**

Table 5. Searching Questions about 'Computers & Internet'

Figure : 检索结果

Topic Model based Method

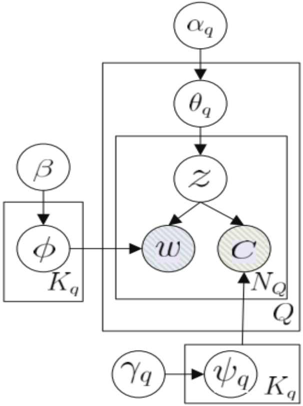

Figure : Topic Model

The end

Thanks!