一种面向社区型问句检索的主题翻译模型

张伟男 张 宇 刘 挺

(哈尔滨工业大学计算机科学与技术学院 社会计算与信息检索研究中心 哈尔滨 150001)

摘 要 基于统计机器翻译模型的问句检索模型,其相关性排序机制主要依赖于词项间的翻译概率,然而已有的模型没有很好地控制翻译模型的噪声,使得当前的问句检索模型存在不完善之处.文中提出一种基于主题翻译模型的问句检索模型,从理论上说明,该模型利用主题信息对翻译进行合理的约束,达到控制翻译模型噪声的效果,从而提高问句检索的结果.实验结果表明,文中提出的模型在 MAP(Mean Average Precision)、MRR(Mean Reciprocal Rank)以及 p@1(precision at position one)等指标上显著优于当前最先进的问句检索模型.

关键词 社区型问答;问句检索;主题模型;翻译模型;LDA(Latent Dirichlet Allocation);社会计算;社交网络

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2015.00313

A Topic Inference Based Translation Model for Question Retrieval in Community-Based Question Answering Services

ZHANG Wei-Nan ZHANG Yu LIU Ting

(*Research Center for Social Computing and Information Retrieval*, *School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001)

Abstract The ranking scheme of the statistical translation based question retrieval models is mainly depended on the translation probabilities between terms. However, the existing translation based models yield on the noise generated by the translation model and further impact the question retrieval results. In this paper, we proposed a topic inference based translation model for question retrieval. By leveraging the topic information, we theoretically verified that it can reasonably control the translation noise and then improves the question retrieval results. Experimental results show that the proposed model significantly outperforms the state-of-the-art question retrieval models in MAP(Mean Average Precision), MRR(Mean Reciprocal Rank) and p@1(precision at position one).

Keywords community question answering; question retrieval; topic model; translation model; LDA (Latent Dirichlet Allocation); social computing; social networks

1 引 言

社区型问答(Community Question Answering, CQA)服务逐渐成为人们在互联网上获取信息以及知识的重要途径.典型的社区型问答服务包括百度知道(http://zhidao.baidu.com/)、知乎(http://www.zhihu.com/)、Yahoo!Answers(http://answers.yahoo.com/)以及 Quora(https://www.quora.com/)等.随着用户对于 CQA 服务的广泛使用,大量的用户生成

收稿日期:2013-11-29;最终修改稿收到日期:2014-09-25.本课题得到国家"九七三"重点基础研究发展规划项目基金(2014CB340503)、国家自然科学基金重点项目(61133012)、国家自然科学基金面上项目(61472105)资助.**张伟男**,男,1985 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为社会计算、信息检索、问答.E-mail:wnzhang@ir.hit.edu.cn.**张 宇**,男,1972 年生,博士,教授,硕士生导师,主要研究领域为个性化信息检索、问答.**刘 挺**(通信作者),男,1972 年生,博士,教授,博士生导师,主要研究领域为社会计算、信息检索、自然语言处理.E-mail:tliu@ir.hit.edu.cn.

内容(User Generated Content,UGC)以问句和答案的形式被积累下来,形成了优质的数据资源.

与传统的搜索引擎检索不同,CQA 检索不是返回与用户查询相关的文档列表,而是返回用户查询的答案,从而能够直接满足用户的查询需求.但是,CQA 查询中,检索的文档是用户生成的问句和答案,其长度远小于传统意义的文档[1].其中大多数的查询词项在用户的问答对中仅出现一次,因此使得基于词频和文档频率统计的检索模型不再适用于问句检索任务.

基于 Unigram 语言模型的问句检索模型[2-3]假设查询中的词项之间是相互独立的,在检索的过程中其相似性排序主要依赖于字符串的严格匹配.尽管语言模型能够利用参数和大数据集(Collection)进行相应的平滑处理,但是其无法解决问句检索中的词不匹配问题.即语言模型无法建立用户对于同一种语义的不同表述形式之间的联系.

基于统计机器翻译模型的问句检索模型[4-7]是当前用于问句检索中的最先进的模型,能够在一定程度上克服上述不足.其利用单语言相似的句子或者双语言互为翻译的句子作为对齐句对,输入到统计机器翻译模型中,获取单语言词项间的翻译概率,并将其作为词项间的相似性度量,或者根据词项的翻译之间的相似性来度量当前词项间的相似性.然而,由于目前用于问句检索的统计翻译模型仅仅利用统计共现信息作为依据来度量词项间的相似性,因此导致语义相关和无关的词项之间的翻译概率无法区分.图 1 所示为 IBM Model 1 在两个语义相似问句之间的词项翻译关系.

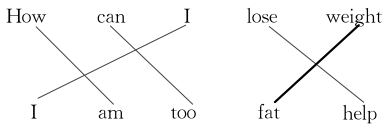

图 1 IBM Model 1 在语义相似问句之间的词项翻译关系

由图 1 我们可以看出,除了“weight”和“fat”之外,其他的词项之间在语义上是不相似的,但是由于 IBM Model 1 无法识别词项间的这种潜在语义相似关系,使得已有的基于统计翻译模型的问句检索模型在其所依赖的词项互译信息上存在较大的噪声,从而影响语义上相似问句的检索结果.因此传统的基于统计机器翻译模型的问句检索模型很难召回图 1 中所示的相似问句.

针对上述问题,我们提出了一种面向社区型问句检索的主题翻译模型,通过在传统的基于统计机器翻译模型的检索模型中,引入词项间在潜在主题下的语义相似性,从而能够解决传统翻译模型中词项翻译准确性较差的问题.具体地,我们通过利用潜在主题对互为翻译的词项进行约束,从而更加合理地度量检索模型中词项之间的相似度,并取得更好的问句检索结果.

2 相关工作

基于语言模型的信息检索模型首次由 Ponte 和 Croft[8]提出,并被广泛应用到信息检索的各个相关领域[3-4,9-13,23].Jeon 等人[2]率先将语言模型应用到问句检索中,他们采用 Unigram 模型对社区型问答服务中的问答对进行建模,从而将其应用于相似问句的发现工作中.最近,Zhang 等人[14]利用依存句法分析技术对用户提出的自然语言问句中词项间的关联性进行度量,从而将重新分配后的权重融入到已有的问句检索模型中,得到更好的问句检索结果.但是由于上述方法没有考虑用户在相同语义上的不同表述形式的信息,从而使得语义上相似但字符串表面不相似的问句无法被召回.

2.1 基于统计机器翻译模型的问句检索模型

针对已有的基于字符匹配的检索模型在匹配用户表述多样性上面的不足,研究人员将基于统计机器翻译模型引入到信息检索模型中,用以获取用户查询中的词项和候选文档中的词项之间的语义相似性.

基于统计机器翻译模型的信息检索模型最初由 Berger 和 Lafferty[15]提出,他们将 IBM Model 1 以及其简化版本(他们称其为 Model 0)应用于信息检索系统中并在 TREC 数据上验证其有效性.随后,Murdock 和 Croft[1]验证了 IBM Model 1 在句子级信息检索上的表现优于传统的 QL(Query Likelihood)检索模型.Xue 等人[4]将语言模型中的平滑机制融入到统计机器翻译模型中,从而提出了一种基于翻译模型的语言模型,并将其应用到问句检索中.然而其采用问句和自身的答案作为平行语料训练翻译模型,包含了很大的噪声.鉴于此,Bernhard 和 Gurevych[5]通过融合多个优质的单语言平行语料,从而使得基于翻译模型的检索模型效果得到了显著的提升.其采用的资源包括 WikiAnswer① 中的问答对,用户标注的相似问句对以及同一单词在不同词

① http://wiki.answers.com/

典中的解释等. Zhou 等人[6]利用短语级的统计机器翻译模型计算查询与待检索文档之间的相似性，并将其应用于问句检索中，取得优于词级别的翻译模型的问句检索结果.

尽管基于短语级翻译模型的检索模型能够为词项引入上下文的信息，从而在一定程度上解决翻译歧义的问题，但是其仍然没有合理的机制来控制统计机器翻译模型在翻译过程中产生的噪声. 本文利用主题信息作为一种隐含语义约束，借此来调整翻译模型用于问句检索时词项间的相似度，从而合理地实现对翻译噪声的控制. 进而更好地解决问句检索中的词不匹配问题.

2.2　基于主题模型的问句检索模型

主题模型作为一种文档表征模型，近年来被广泛应用于信息检索和文本挖掘的相关任务中[17]. 其中代表性的主题模型主要有 PLSA[16]（Probabilistic Latent Semantic Analysis）和 LDA[17]（Latent Dirichlet Allocation）. Steyvers 等人[18]通过将作者和主题之间建立起潜在的语义对应关系，从而提出一种新的作者主题模型. 并将此模型用于在论文数据库中发现研究主题的趋势演变以及为特定的作者生成摘要及相关论文推荐等. Wei 和 Croft[19]利用 LDA 模型对文档和查询之间的关系进行建模，通过线性结合的方式将 LDA 模型融合到最终的检索模型中，在 TREC 数据集上取得较好的检索结果. 类似地，Cai 等人[7]利用 LDA 检索模型融入到翻译模型中，并将其应用于问句检索任务. Ji 等人[20]通过对问句和答案分别进行主题建模，并假设问句和答案之间不仅属于同一个主题而且还应该共享一些词汇信息. 即利用答案作为一种查询扩展应用于原始查询中，并用其提高问句检索结果.

尽管上述工作验证了主题信息对于问句检索的作用，但是据我们所知，通过原理性的分析并从已有的检索模型中合理地推导并融入主题模型信息的工作仍然罕有涉及. 本文从理论上验证了我们所提出的方法能够将主题模型合理地融入到当前最先进的问句检索模型中，从而提出一种新的基于主题翻译模型的问句检索模型.

2.3　基于主题模型的机器翻译模型

近年来，基于主题模型的机器翻译模型受到了广泛的关注[25-29]，其主要思想是通过主题模型解决翻译模型的适应性问题. 其中 Zhao 和 Xing[25]提出一种双语主题混合模型，并将其用于统计机器翻译的词对齐任务中. 在词和句子级别上，作者提出了三种模型来获取双语文档中的主题信息，并最终将三种模型进行融合用以提高词对齐的效果，进而提高机器翻译的效果. Tam 等人[26]提出了一种基于双语 LSA 的跨语言的语言模型以及翻译词典自适应的方法. 他们首先通过主题模型对源文本进行主题分布推断，然后将得到的主题分布用于目标语言的 n-gram 语言模型，以此来提高翻译模型的自适应性. Gong 等人[27]指出现有统计机器翻译只考虑句子级别的信息是不合理的，并由此引入文档主题信息用以提高机器翻译的效果. 具体地，他们利用 LDA 主题模型对文档片段进行建模，并在翻译过程中加入主题分布概率信息. Eidelman 等人[28]将主题模型信息作为特征融入翻译模型，用以提高翻译模型的自适应能力. Xiao 等人[29]提出一种基于规则的主题相似度计算模型，并将其应用在层次化短语结构的统计机器翻译模型中，提升翻译模型的性能.

考虑到翻译模型在问句检索上的良好表现，本文将主题模型引入到翻译模型中，通过主题的信息提高翻译模型的性能，进而提高问句检索的效果.

3　基于主题翻译模型的问句检索模型

3.1　LDA 背景简介

LDA 模型是由 Blei 等人[17]提出的一种新式的语义一致的主题模型. 它的提出迅速得到了统计机器翻译，自然语言处理以及信息检索相关研究者的关注. 同时，LDA 是一种概率图模型，其表示形式如图 2 所示. 其中 N_m 为第 m 篇文档的长度，n 为单篇文档中词项的索引号，K 为主题数，M 为文档集的规模，即文档数.

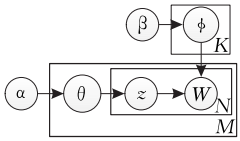

图 2　LDA 的概率图模型表示形式

LDA 模型生成文本内容的过程如下所示：

(1) 为每个主题 z（以参数 β 服从 Dirichlet 分布）选择一个多项式分布 ϕ_z；

(2) 为每个文档 w_m（以参数 α 服从 Dirichlet 分布）选择一个多项式分布 θ_m；

(3) 为每个文档 w_m 中的词项 $w_{m,n}$（$n\in[1,N_m]$）选择一个主题 z（$z\in\{1,\cdots,K\}$）；

(4) 从多项式分布 ϕ_z 中选择词项 $w_{m,n}$.

相应地，对于单篇文档 $\vec{w}_m$ 而言，LDA 生成其内

容的可能性(likelihood)如下所示：

$$p(\vec{w}_m|\vec{\alpha},\vec{\beta})=\iint p(\vec{\phi}_z|\vec{\beta})\cdot p(\vec{\theta}_m|\vec{\alpha})\cdot \prod_{n=1}^{N_m}\sum_{z_{m,n}} p(z_{m,n}|\vec{\theta}_m)\cdot p(w_{m,n}|\vec{\varphi}_{z_{m,n}}) \quad (1)$$

这里，在社区型问答服务中，一个问答对被看作是一篇文档，因此在本文中所用到的主题模型建模对象为问答对数据集合.

3.2 基于主题翻译的问句检索模型

本节将介绍我们提出的基于主题翻译的问句检索模型. 图 3 所示为该模型的系统框图. 该模型以用户的自然语言问句查询为输入，通过利用主题模型对翻译模型的质量进行提高，从而实现问句检索结果的优化.

Rank	Questions
1	Is the Apple Store online a good store to buy from?
2	Is there any online shop for Apple fruit tree to be send to Malaysia from all over the countries?
3	Is big apple pet supply reliable for live animal shipments?
4	How can I get my money wired from Apple Bank to a localbank?
5	What things will affect the apple stock?
...	
31	Should I get my ipod nano at the Apple Store or online?

Rank	Questions
1	Is the Apple Store online a good store to buy from?
2	Should I get my ipod nano at the Apple Store or online?
3	Is it cheaper to buy an iPod on the online apple store or in a regular store/shop?
4	Where can I get an Apple PowerBank G4 replacement power supply online?
5	How do you disable the itunes store and the App Store on the new iPhone?

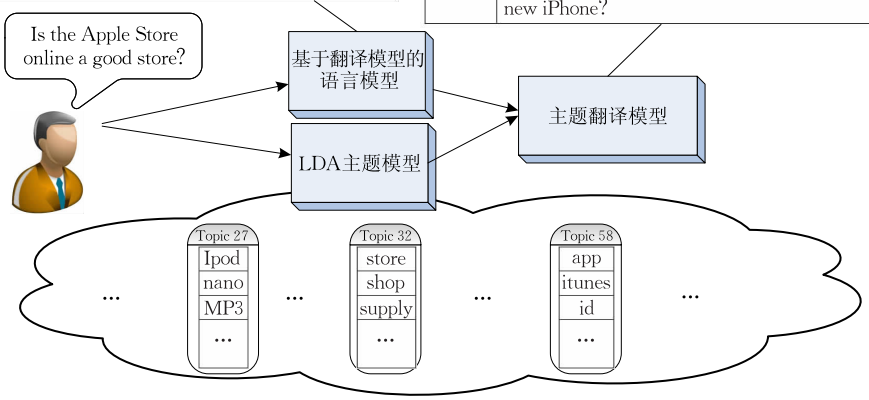

图 3 基于主题翻译模型的问句检索模型

接下来，我们从最先进的问句检索模型入手，推导出融合主题信息的新的问句检索排序模型. 目前最先进的基于统计机器翻译模型的问句检索模型[4]，其排序机制如下所示：

$$P_{\mathrm{TLM}}(w|(q,a))=\lambda_1 p_{ml}(w|q)+\lambda_2\Big(\sum_{t\in q}p(w|t)p_{ml}(t|q)\Big)+\lambda_3 p_{ml}(w|a) \quad (2)$$

其中，w 表示查询中的特定词项，q 表示待检索的问句，t 为 q 中的词项，a 为 q 相对应的答案. TLM 代表 Xue 等人[4]提出的基于翻译模型的语言模型，ml 表示极大似然估计方法. 且存在关系 $\lambda_1+\lambda_2+\lambda_3=1$.

这里，$p(w|t)$为从查询中的词项 w 到待检索问句中的词项 t 之间的翻译概率，我们在此基础之上，考虑引入主题信息，从而通过主题空间上的相似性来提高翻译模型的准确性，具体推导如下所示：

$$\begin{aligned}\hat{p}(w|t)&=\sum_{i=1}^{K}p(w|t,z_i)\cdot p(z_i|t)\\&\propto\gamma p(w|t)+(1-\gamma)\sum_{i=1}^{K}p(w|z_i)\cdot p(z_i|t)\\&=\gamma p(w|t)+(1-\gamma)\sum_{i=1}^{K}p(w|z_i)\cdot\frac{p(t|z_i)\cdot p(z_i)}{p(t)}\\&\propto\gamma' p(w|t)+(1-\gamma')\sum_{i=1}^{K}p(w|z_i)\cdot p(t|z_i) \quad (3)\end{aligned}$$

这里，$\gamma'=\dfrac{\gamma}{\gamma+\dfrac{p(z_i)}{p(t)}(1-\gamma)}$为线性插值参数，用于平衡翻译模型和主题模型之间的权重. 式(3)第 1 步到第 2 步的推导，是为了对 $p(w|t,z_i)$进行估计时便于计算所采用的一种近似处理，这里，我们采用了一种经常用于联合条件概率估计的线性插值法[19]，具体地，我们使用 $p(w|t)$和 $p(w|z_i)$对 $p(w|t,z_i)$进行估计. $p(w|t)$利用 GIZA++[21] 获得，$p(w|z_i)$和 $p(t|z_i)$由 LDA 模型通过 Gibbs 采样方法估计得出. 这样我们可以得到最终的问句检索排序模型如下所示：

$$\begin{aligned}&P_{\mathrm{T^2LM}}(w|(q,a))=\\&\mu_1 p_{ml}(w|q)+\mu_2\Big(\sum_{t\in q}p(w|t)p_{ml}(t|q)\Big)+\\&\mu_3\Big(\sum_{t\in q}p_{ml}(t|q)\sum_{i=1}^{K}p(w|z_i)\cdot p(t|z_i)\Big)+\mu_4 p_{ml}(w|a)\end{aligned} \quad (4)$$

其中，T^2LM 表示我们提出的基于主题翻译模型的问句检索模型. 且存在 $\mu_1+\mu_2+\mu_3+\mu_4=1$.

从直观上看，我们可以得出结论，当 $p(w|z_i)$ 和 $p(t|z_i)$ 的值越接近时，则式(4)中第 3 项的值越高，即查询中的词项 w 以及待检索问句中的词项 t 属于同一主题的可能性越大，则其相似性值越高，从而实现利用主题信息对词项间相似性的度量机制的更新.

4　实验结果及分析

4.1　实验数据集

我们利用 API① 从 Yahoo! Answers 中获取了共 1123134 个完整问答对，其中包括问句的 title、content 以及 answers. 该数据集覆盖了较为广泛的主题，例如 Health、Internet 等. 我们从中随机的选择了 200 个问句作为我们的查询集合，在去除停用词之后，我们手工地过滤掉长度小于 2 个词的查询，最终得到了 168 个问句查询，我们在其中随机选择了 140 个问句作为测试查询，剩余的 28 个作为开发集用于参数调整. 另外，由于我们需要的是在大规模数据上对主题进行建模，因此本文主题模型建模的对象是整个问答对数据集.

为了获取查询相关性问句集合，对于查询测试集中的每个查询，我们汇集多个搜索模型(如向量空间模型、BM25 模型、语言模型以及翻译模型等)的前 20 个检索结果，并聘请两位以英语为母语，且不熟悉当前实验方法设计的学生进行手工标注检索结果为相关(数字 1)或不相关(数字 0)，当标注出现冲突时，由第 3 位标注人员对标注结果进行判定.

我们采用 p@1(precision at position one)、MAP(Mean Average Precision)和 MRR(Mean Reciprocal Rank)[24] 作为评价指标.

4.2　实验对比系统

我们选取了在问句检索方面的经典模型及当前最先进模型作为对比系统，具体设置如下：

(1) 语言模型(LM). Jeon 等人[2] 提出的基于语言模型的问句检索模型.

(2) 翻译模型(TRM). Murdock 和 Croft[1] 提出的基于统计机器翻译模型的句子检索模型.

(3) 基于翻译模型的语言模型(TLM). Xue 等人[4] 提出的基于翻译模型的语言模型.

(4) 基于词项赋权的 TLM(drTLM). Zhang 等人[14] 提出的利用依存句法分析图进行问句词项重新赋权的 TLM.

(5) 主题模型(TM). Wei 和 Croft[19] 提出的基于 LDA 的信息检索模型.

此外，还有很多在问句检索方面的杰出工作，如 Cao 等人[12]、Cai 等人[7] 和 Zhou 等人[6] 等，前两项工作依赖于问句所在的类别信息，后一项工作基于短语级翻译模型. 然而，一方面，我们所提出方法的目标是构建一个通用的问句检索模型，使其可以不受应用场景的影响. 因此，我们并没有利用 Yahoo! Answers 的类别信息作为辅助信息来指导问句检索模型. 另一方面，考虑到短语识别本身存在一定的错误，以及短语级别的问句检索存在数据稀疏问题，都会影响问句检索的效果，因此我们采用词作为基础单元进行问句检索的相关研究. 文献[20]利用主题模型度量问题与答案词项之间主题分布的相似性，指导问句检索模型，而本文的方法更加注重于度量问句与问句词项之间的主题分布的相似性，以此对问句检索模型进行改进，由于本文与文献[20]在任务的基本假设上存在差异，以及文献[20]与文献[19]在主题建模形式上相似，又同时应用于检索任务中，因此，我们对比了本文方法和文献[19]的实验结果，而没有与文献[20]进行直接的比较.

我们在开发集中，利用 WEKA[22] 提供的 Grid Search 工具将上述对比系统以及我们系统的参数分别调至最优，其中 $\mu_1=0.3$，$\mu_2=0.1$，$\mu_3=0.4$，$\mu_4=0.2$，主题数 $K=80$. 表 1 所示为实验结果对比. 本文所有的结果都是在 $p<0.05$ 的条件下进行 t-test 统计显著性检验的结果. 在统计显著性检验中，p 值表示当假设成立时，获得一个测试统计至少被观测到一次的概率. p 值通常与接受假设检验结果成立的置信度相对应，即 $p<0.05$ 意味着我们可以以高于 0.95 的置信度接受假设检验的结果. 本实验中，我们以步长为 10，变化范围为 10～50，对 Gibbs

表 1　问句检索实验结果对比

	LM	TRM	TLM	TM	drTLM	**T^2LM**
MAP	0.2635	0.2678	0.2889	0.3043	0.4170	**0.4375**
% of MAP improvements over						
LM	N/A	+1.63	+9.64	+15.48	+58.25	**+66.03**
TRM	N/A	N/A	+7.88	+13.63	+55.71	**+63.37**
TLM	N/A	N/A	N/A	+5.33	+44.34	**+51.44**
TM	N/A	N/A	N/A	N/A	+37.04	**+43.77**
drTLM	N/A	N/A	N/A	N/A	N/A	**+4.92**
p@1	0.2071	0.2143	0.1928	0.2143	0.2771	**0.2929**
MRR	0.1929	0.1940	0.1889	0.2018	0.2583	**0.2734**

注：粗体为我们所提出的方法及相应的结果，其中 T^2LM 在 $p<0.05$ 的情况下，在统计上显著优于 LM、TRM、TLM 和 TM.

① http://developer.yahoo.com/answers/

采样的样本量进行了测试，测试结果表明，采样样本量的变化在实验结果上的差异性可以忽略不计，本文采用的采样样本量为 40.

由表 1 我们可以得出以下的分析：

(1) TM 和 T^2LM 在问句检索上的效果要优于 LM、TRM 和 TLM. 这是因为前两个模型引入了主题信息辅助的问句检索模型. 主题信息对于 TM 而言其作用相当于一种查询扩展，而对于 T^2LM 来说，由于其引入了词项间的翻译信息，因此主题信息既可以看作是基于翻译模型的扩展，同时通过主题的限定也能够帮助控制翻译模型在词项互译过程中产生的噪声，因此主题模型在应用于基于翻译模型的问句检索模型上更有优势.

(2) 对比 TLM 和 TM 的结果我们可以看出，虽然这两个模型都是部分地基于语言模型的问句检索模型，但是基于主题模型的语言模型(TM)的实验结果要优于基于翻译模型的语言模型(TLM). 从而说明对于问句查询扩展而言，主题模型的效果要优于翻译模型. 这是因为主题模型的工作原理更类似于词项聚类，即将语义上相似的词项聚类成若干个主题类别. 但翻译模型则是基于词项间的统计共现性的，而共现频度高的词项未必是同一类别的，因此在问句检索的查询扩展方面主题模型更有优势.

(3) 对比 T^2LM 和 TLM 的结果我们可以看出，融入主题信息的 T^2LM 模型在 MAP 上面高出 TLM 模型 51.44%. 说明其对基于翻译模型的问句检索模型有较大的帮助. 同时我们注意到，在 T^2LM 中 μ_3 的值最大，这说明了其在问句检索上面性能的提升主要来自于主题模型部分. 此外，由于直接对词项翻译结果进行评价需要大量的人力，因此本实验中没有讨论在加入主题信息之后的词项翻译结果和原始的词项翻译结果的比较. 由于 TLM 方法的原始文献中，没有发布实验数据集，因此我们无法在其原始数据集上重现实验结果. 另外，TLM 在我们的数据集上的表现低于其原始文献中的结果，这是因为我们实验中所采用的数据集是来自 Yahoo!Answers，而 TLM 原始文献中的数据集来自于 Wondir①，这是数据集上的差异. 此外，TLM 原始文献中的测试集是来自于 TRECQA 的 50 个问题，TRECQA 与 Yahoo!Answers 的数据差异很大，TRECQA 的数据是人工构造的问题，而我们采用的测试数据是完全由用户生成的数据. 因此导致了 TLM 在本文上的表现以及与 T^2LM 之间结果的差异.

(4) 值得注意的是，我们对比了目前最先进的基于词项重要性赋权的问句检索模型 drTLM. 通过比较发现，T^2LM 的结果要优于 drTLM，这主要因为 T^2LM 一方面是通过改变查询中词项的权重从而提升问句检索的效果，更重要的是另一方面其能够从翻译模型中获得词项扩展增益.

由于在使用 LDA 的过程中，主题数需要在实验前给定，因此我们考虑了主题数对于实验结果的影响，图 4 所示为在开发集上 MAP 随着主题数的变化曲线.

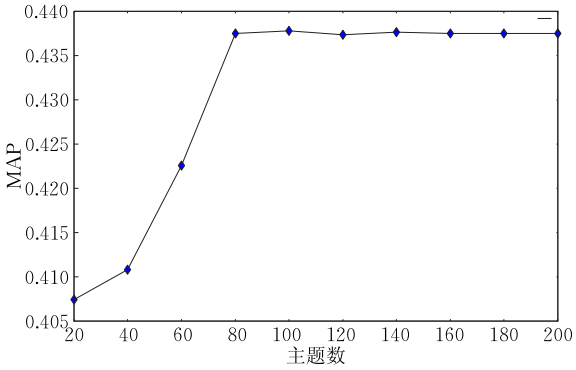

图 4 T^2LM 模型的主题数与 MAP 的变化曲线

由图 4 可以看出，在主题数小于 80 的时候 MAP 的值随着主题数增长而增长，在 80 之后则趋于平稳. 通过观察主题分析后的输出数据我们可以看出，当设定的主题数超过 80 之后，主题区分度的模糊性便显现出来，以 100 个主题为例，我们通过输出每个主题的高频词列表中观察到，有 20 个主题的高频词都能够在其他 80 个主题中找到，而在这 20 个主题中，除去其高频词之外的其他主题词则多数为噪声词，因此，我们可以推断这 20 个主题能够被包含在其余的 80 个主题中. 对其他主题数的观察结果与上述结果相似. 因此在我们的实验中，主题数选取为 80.

表 2 所示为 TLM 和 T^2LM 在“Is the Apple Store online a good store?”查询上前 3 位检索结果的对比，其中粗体为相关检索结果. 由表 2 我们可以

表 2 问句检索结果比较示例

Rank	TLM	T^2LM
1	**Is the Apple Store online a good store to buy from?**	**Is the Apple Store online a good store to buy from?**
2	Is there any online shop for Apple fruit tree to be send to Malaysia from all over the countries?	**Should I get my ipod nano at the Apple Store or online?**
3	Is big apple pet supply reliable for live animal shipments?	**Is it cheaper to buy an iPod on the online apple store or in a regular store/shop?**

① http://www.wondir.com/

直观地看出 T^2LM 检索结果明显优于 TLM，这是因为在 T^2LM 中，其问句检索结果被限定在 3 个主题中，在本实验中为第 27，32 和 58，如图 3 中所示，可见这 3 个主题都与 Apple Store 有关，且都是电子产品类别，因此 T^2LM 能够召回更多相关的检索结果且排序靠前.

5　实现和应用时的关键技术点

本文所提出的基于主题翻译模型的问句检索模型在实现和应用时，主要依赖的是主题分布信息和词的互译概率信息. 前者是通过 LDA 主题模型训练得到的词的主题分布信息，后者是通过 Giza++ 词对齐后得到的词与词的互译信息. 此外，为了保证实际效果的准确性，计算主题分布信息时，需要根据特定的数据集调整主题的数量，同时，需要在相应的数据集上获取单语平行语料作为翻译模型输入. 最后，在应用时，需要根据特定的应用来调整整个问句检索模型的各个参数，以达到最优的问句检索效果.

6　结论及未来工作

本文提出了一种基于主题翻译模型的问句检索模型，通过在基于统计机器翻译模型的问句检索模型中引入主题信息，从而解决了由于翻译模型产生的噪声而影响问句检索结果的问题. 同时我们在理论上说明我们所提出的主题模型可以合理地融合到已有的最先进的检索模型中，实验结果证实了其有效性.

尽管主题模型能够作为一种潜在语义扩展增强问句检索的效果，但是我们也应当发现，目前的技术没有很好地解决主题间的歧义关系问题，因此在后续工作中，我们会进一步深入探讨如何解决主题的歧义性问题，以期获得更好的问句检索效果.

致　谢　编辑及审稿老师给了宝贵意见，在此表示感谢！

参　考　文　献

[1] Murdock V, Croft W B. A translation model for sentence retrieval//Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP). Vancouver, Canada, 2005: 684-691

[2] Jeon J, Croft W B, Lee J H. Finding similar questions in large question and answer archives//Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM). Bremen, Germany, 2005: 84-90

[3] Duan H, Cao Y, Lin C Y, et al. Searching questions by identifying question topic and question focus//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL). Columbus, USA, 2008: 156-164

[4] Xue X, Jeon J, Croft W B. Retrieval models for question and answer archives//Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR). Singapore, 2008: 475-482

[5] Bernhard D, Gurevych I. Combining lexical semantic resources with question & answer archives for translation-based answer finding//Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL). Singapore, 2009: 728-736

[6] Zhou G, Cai L, Zhao J, et al. Phrase-based translation model for question retrieval in community question answer archives//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL). Portland, Oregon, USA, 2011, 1: 653-662

[7] Cai L, Zhou G, Liu K, et al. Learning the latent topics for question retrieval in community QA//Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). Chiang Mai, Thailand, 2011: 273-281

[8] Ponte J M, Croft W B. A language modeling approach to information retrieval//Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR). Melbourne, Australia, 1998: 275-281

[9] Song F, Croft W B. A general language model for information retrieval//Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM). Kansas City, USA, 1999: 316-321

[10] Zhai C X. Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies, 2008, 1(1): 1-141

[11] Ming Z Y, Chua T S, Cong G. Exploring domain-specific term weight in archived question search//Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM). Toronto, Canada, 2010: 1605-1608

[12] Cao X, Cong G, Cui B, et al. Approaches to exploring category information for question retrieval in community question-answer archives. ACM Transactions on Information Systems (TOIS), 2012, 30(2): 7

[13] Wang K, Ming Z, Chua T S. A syntactic tree matching approach to finding similar questions in community-based QA services//Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR). Boston, USA, 2009: 187-194

[14] Zhang W-N, Ming Z-Y, Zhang Y, et al. The use of dependency relation graph to enhance the term weighting in question retrieval//Proceedings of the International Conference on Computational Linguistics (COLING). Mumbai, India, 2012: 3105-3120

[15] Berger A, Lafferty J. Information retrieval as statistical translation//Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR). Berkeley, USA, 1999: 222-229

[16] Hofmann T. Probabilistic latent semantic indexing//Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR). Berkeley, USA, 1999: 50-57

[17] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[18] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for information discovery//Proceedings of the ACM Knowledge Discovery and Data Mining (SIGKDD). Seattle, USA, 2004: 306-315

[19] Wei X, Croft W B. LDA-based document models for ad-hoc retrieval//Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR). Seattle, USA, 2006: 178-185

[20] Ji Z, Xu F, Wang B, et al. Question-answer topic model for question retrieval in community question answering//Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM). Sheraton Maui, Hawaii, 2012: 2471-2474

[21] Och F J, Ney H. Improved statistical alignment models//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Hong Kong, China, 2000: 440-447

[22] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18

[23] Gao Y, Wang M, Zha Z, et al. Visual-textual joint relevance learning for tag-based social image search. IEEE Transactions on Image Processing (TIP), 2013, 22(1): 363-376

[24] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: ACM Press, 1999

[25] Zhao B, Xing E P. BiTAM: Bilingual topic admixture models for word alignment//Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics on Main Conference Poster Sessions. Association for Computational Linguistics (ACL). Sydney, Australia, 2006: 969-976

[26] Tam Y C, Lane I, Schultz T. Bilingual LSA-based adaptation for statistical machine translation. Machine Translation, 2007, 21(4): 187-207

[27] Gong Z, Zhang Y, Zhou G. Statistical machine translation based on LDA//Proceedings of the IEEE 4th International Universal Communication Symposium (IUCS). Beijing, China, 2010: 286-290

[28] Eidelman V, Boyd-Graber J, Resnik P. Topic models for dynamic translation model adaptation//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics (ACL). Jeju Island, Korea, 2012: 115-119

[29] Xiao X, Xiong D, Zhang M, et al. A topic similarity model for hierarchical phrase-based translation//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics (ACL). Jeju Island, Korea, 2012: 750-758

ZHANG Wei-Nan, born in 1985, Ph. D., lecturer. His research interests include social computing, information retrieval and question answering.

ZHANG Yu, born in 1972, Ph. D., professor, M. S. supervisor. His research interests include question answering and personalized information retrieval.

LIU Ting, born in 1972, Ph. D., professor, Ph. D. supervisor. His research interests include social computing, information retrieval and natural language processing.

Background

Question retrieval in CQA is different from general Web search. Unlike the Web search engines that return a long list of ranked documents, question retrieval returns several relevant questions with possible answers directly. Meanwhile, question retrieval can also be considered as a traditional Question Answering (QA) problem, but the focus of the QA task is transformed from answer extraction, answer matching and answer ranking to searching for relevant questions with good ready answers.

A major challenge is the word verboseness in the queries where important words may be surrounded by other additional words. These additional words are more likely to confuse the

current search engines rather than help them. The other major challenge is the word mismatch between the queries and the candidate questions for retrieval. This makes it difficult for the two questions to match each other in the question retrieval task. In applications based on user generated content (UGC), such as CQA services, where the users tend to use a more diverse and informal vocabulary to express their information needs, the word mismatch problem is even more common and severe than in general search.

In order to solve the word verboseness in queries, previous work mainly focused on core term discovery, query reformulation, key concept identification on verbose queries, etc. Despite the great success achieved these papers mainly focused on distinguishing the key concepts from the non-key ones and the importance among the key concepts was not taken into consideration. In this paper, we propose a ranking based method for key concept identification, which not only distinguishes the key concepts from the non-key ones, but also captures the differences among key concepts.

To tackle the word mismatch problem, previous work mainly resorts to query expansion. However, the former approach overlooks concept level evidences for query expansion and the latter approach fails to assign explicit weights to the expanded aspects.

In this paper, we proposed a topic inference based translation model to tackle the word mismatch problem in question retrieval. By leveraging the topic information, we theoretically verified that it can reasonably control the translation noise and then improves the question retrieval results. Experimental results show that the proposed model significantly outperforms the state-of-the-art question retrieval models in MAP, MRR and p@1.

This work is supported by the National Basic Research Program (973 Program) of China (Grant No. 2014CB340503) and the National Natural Science Foundation of China (Grant Nos. 61133012, 61472105).