# 基于统计机器翻译模型的查询扩展

李卫疆 赵铁军 王宪刚
(哈尔滨工业大学计算机科学与技术学院语音语言教育部-微软重点实验室 哈尔滨 150001)

**摘 要**：在搜索引擎等实际的信息检索应用中，用户提交的查询请求通常都只包含很少的几个关键词，这会引起相关文档与用户查询之间的词不匹配问题，对检索性能有较严重的负面影响。该文在分析了查询产生模型的基础上，提出了一种新的基于统计机器翻译的查询扩展方法。通过统计机器翻译模型提取文档集中与查询词相关联的词，用以进行查询扩展。在 TREC 数据集上的试验结果表明：基于统计翻译的查询扩展方法不仅比不扩展的语言模型方法始终有 12 %～17% 的提高，而且比流行的查询扩展方法-伪反馈也具有可比的平均准确率。

**关键词**：信息检索；查询扩展；语言模型；统计机器翻译
**中图分类号**：TP391 **文献标识码**：A **文章编号**：1009-5896(2008)03-0725-05

# A SMT-based Approach for Query Expansion in Information Retrieval

Li Wei-jiang Zhao Tie-jun Wang Xian-gang
(*School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001, *china*)

**Abstract**: In practical applications of information retrieval, such as the search engine，the query user submitted contains only several keywords usually. This will cause unmatched issue of word of relevant files and user's query and have more serious negative effects on the performance of information retrieval. On the basis of analyzing of process of producing query, this paper puts forward a new method of query expansion on the basis of model of statistical machine translation. The approach extract related terms between documents and query through statistical machine translation model, then expand into query. The experiment result on TREC data collection shows the proposed method, SMT-based query expansion, has 12 - 17% of the improvement all the time more than the language model method without expanding. Compared to the popular approach of query expansion, pseudo feedback, the proposed method has the competed average precision.

**Key Words**: Information retrieval; Query expansion; Language model; Statistical Machine Translation(SMT)

## 1 引言

随着互联网业务的蓬勃发展，互联网上的信息也随之迅速膨胀。在这纷繁的数据海洋中，如何准确而有效地检索用户想要的有用的信息成为当今研究的热点。Ponte 和 Croft 在 1998 年[1]提出的语言模型对当今的信息检索研究影响很大。该语言模型的原理是：首先估计每个文档的语言模型；然后假设查询是由文档模型生成的，计算由估计的文档模型产生查询的概率。然而在搜索引擎等实际的信息检索应用中，用户提交的查询请求通常都只包含很少的几个关键词，这会引起相关文档与用户查询之间的词不匹配问题，对检索性能有比较严重的负面影响。如何解决词不匹配问题成为信息检索领域中一个十分重要的研究课题。查询扩展是解决词不匹配问题的有效技术手段，它以用户的初始查询为基础，通过一定的策略加入一些相关的词，以提供更多有利于判断文档相关性的信息。

2006-09-26 收到，2007-01-26 改回
国家自然科学基金重点项目(60435020)和微软亚洲研究院项目资助课题

查询扩展中最关键的技术之一就在于扩展词表的构造。目前扩展词表的构造通常有 3 种方式：第 1 种是根据语言学知识基于语义的查询扩展词表构造方法[2]，并构建了一些大规模的手工词典，例如 WordNet，HowNet 等；第 2 种是基于大规模通用语料库的统计信息，例如同现概率、互信息等构造扩展词表[3−5]；第 3 种是结合语言知识和统计信息的扩展词表构造方法[6, 7]，例如基于依存关系统计信息的扩展词表[8]。以前研究者尝试通过辞典，进行查询扩展，但效果不明显。一方面原因是：虽然这些辞典包含了人类专家标注的许多词的关系，但是这些关系并不是专门为信息检索服务而设计的。另一个原因是缺乏应用这种关系的上下文信息。

本文在分析了查询产生模型的基础上，提出了一种新的基于统计机器翻译的查询扩展方法。通过统计机器翻译模型提取文档集中与查询词有关联关系的词，并用以进行查询扩展。通过在 TREC 数据集的试验表明，本文所提出的方法优于传统的基于共现关系的方法。

## 2 相关工作

相关研究已经发现决定查询扩展的关键因素是适当的

扩展词的选择[9] 。为了确定扩展词，一种方法是通过手工建立的词典，像 Wordnet、Hownet 等。词典包括人为建立的 term 关系表，它们可以用来指示 terms 之间的关系。 文献[10]就利用 Wordnet 词典选择同义词和歧义词来进行查询扩展。然而，他的实验并没有显示对检索效果的改进。可能有下列原因：虽然 Wordner 包含了人类专家标注的许多词的关系，但是这些关系并不是专门设计用来为信息检索服务的。另一个原因是缺乏应用这种关系的上下文信息。

另一种方法是关联关系抽取：如果两个词是频繁地同时出现，那么认为是互为相关的。但是，共现关系总是引入很多噪声，这导致检索结果返回很多不相关的文档，从而大大降低召回率。同时可能丢失真正的关联关系词。

为了选择更好的词来扩展查询，文献[11]提出了一种方法：通过计算与整个查询的关系来选择扩展词，与查询的关系通过与每个查询中的词的关系的累加计算出。因此，与多个查询此相关的词将被优先选中。文献[12]也尝试通过混合一个词语一组词的关系来确定该词与一组词的关联关系。在文献[13]提出的方法中，作者尝试通过一个上下文相关词来确定查询词的词义。然而，对于查询扩展来说，并不需要明确的词义消歧，也就是说，不必准确地知道词的词义再来做查询扩展，而仅仅需要确定相关的扩展词。文献[14]对上述方法进行了扩展，通过该词出现在同一上下文来计算词的相似度。一旦词的关系确定了，相关联的词将被做查询扩展。在本论文中，基本的信息检索框架采用语言模型[1]。之所以选择这个模型主要是因为该模型可以很灵活地整合词的关联关系。实际上，以往的研究已经证明语言模型有能力整合词的关系和查询扩展[15]。然而上述研究并没有涉及文档集中词的上下文关系的抽取。

近年来，国内也对查询扩展进行了深入的研究。张敏、马少平等人提出了根据词之间的语义关系进行扩展和替换的文档重构方法[7]。 通过对隐含语义索引的分析，文献[16]提出了语义双重查询扩展的方法。文献[4]提出了一种基于局部共现的查询扩展方法。在文献[6]的研究中，通过对文档进行潜在语义分析，引入计算词语间语义相似度的方法。文献[5]提出相关文档词和非相关文档词分别建立语言模型,然后通过计算某个词出现在相关文档词和非相关文档词中分布的变化去估计这个词和相关文档词的相关度。

## 3 基于统计机器翻译模型的查询扩展

下面首先描述检索语言模型，分析查询产生过程。然后引出利用统计翻译模型产生估计查询词产生扩展词的概率。

### 3.1 查询模型

根据 Ponte 和 Croft 在 1998 年[1]提出的语言模型。文档 $D$ 的相关度是由该文档产生查询 $Q$ 的概率决定。 假设查询 $Q$ 是由相互独立的一组词组成，$q_i$ 为查询 $Q$ 中的一个词，则

$$\log p(Q \mid D) = \sum_i \log p(q_i \mid D) = \sum_{i:c(q_i,d)>0} \log p_s(q_i \mid d) + \sum_{i:c(q_i,d)=0} \log p_u(q_i \mid d)$$

上式中前项表示文档模型生成查询词的概率，后项表示文档模型未生成查询中的词的概率。此模型假定查询模型与文档模型服从同一的概率分布。 在实际应用中，查询普遍很短，因此用文档模型来估计查询模型是不确切的。

在此基础上提出的另一个模型是基于文档模型和查询模型的相对熵(KL-divergence)来计算查询和文档的相关度：

$$\mathrm{Rank}(D,Q) = \sum_{q_i \in Q} P(q_i \mid Q) \log P(q_i \mid D) \tag{1}$$

在传统语言模型中，为了计算上的简化，假设词与词之间是相互独立的。如果一个查询词出现在文档中，那么它在一定程度上与该文档相关。但是，如果一个查询词没有出现在文档中，未必就能断定此查询与该文档无关。也许有一些与该查询词密切相关的词包含在文档中，例如，同义词等。因此，仅仅通过查询词是否在文档中出现来判断文档的相关性是不完全的，必须通过一种方法找出这些与查询词相关的词。

设 $e_i$ 是查询词 $q_i$ 的扩展词， $d$ 表示文档， $q$ 表示查询。假设查询是由这样的随机过程产生的：首先依照概率 $P(q_i \mid Q)$ 从 $Q$ 中取一查询词 $q_i$ ，再根据 $P(e_i \mid q_i)$ 产生扩展词 $e_i$ ，则

$$P(q \mid Q) = \sum_{q_i \in Q} P(e_i, q_i \mid Q) = \sum_{q_i \in Q} P(e_i \mid q_i) P(q_i \mid Q) \tag{2}$$

其中 $P(e_i \mid q_i)$ 表示两个词之间的关联关系。$P(q_i \mid Q)$ 可以通过最大似然估计计算出来，用 $P_{\mathrm{ML}}(q_i \mid Q)$ 表示。 下一节将介绍如何通过统计翻译模型估计 $P(e_i \mid q_i)$ 。

### 3.2 查询扩展词的产生

根据文献[17]提出的模型，可以利用此模型生成词到词的翻译概率。把查询和文档构成的句对$<q, d>$作为翻译的训练数据，通过 $d$ 到 $q$ 翻译找出查询词对应于文档中的同义词和关联词。这种同义关系是统计意义上的关系， 因此，更能反映该词与上下文间的关联关系。 把这些与查询有关联的词扩展到查询中，从而提高查询的效率。 因为，在检索语言模型中并不考虑词在文档和查询中的位置，所以本文采用统计机器翻译的 IBM 模型 1。

设 $d =< d_1, d_2, \cdots, d_m >$ 表示文档句， $q =< q_1, q_2, \cdots, q_l >$ 表示查询句， $a$ 表示对齐，则

$$\Pr(q, a \mid d) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} t(q_i \mid d_{a_i})$$

加上所有可能的对齐

$$\Pr(q \mid d) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(q_i \mid d_{a_i})$$

通过 Lagrange 变换，

$$t(q_i \mid d_j) = \lambda_e^{-1} c(q_i \mid d_j; q, d) \tag{3}$$

实际中，训练数据由一组句对组成，$(q^{(1)} \mid d^{(1)}, q^{(2)} \mid d^{(2)},$

$\cdots, q^{(S)} \mid d^{(S)})$，那么翻译概率是：

$$t(q_i \mid d_j) = \lambda_e^{-1} \sum_{s=1}^{S} c(q_i \mid d_j; q^{(s)}, d^{(s)}) \tag{4}$$

其中 $\lambda_e$ 是 Lagrange 乘子。

$$c(q_i \mid d_j; q, d) = \frac{t(q \mid d)}{t(q \mid d_0) + \cdots + t(q \mid d_l)} \overbrace{\sum_{j=1}^{m} \delta(q, q_j)}^{\text{count of } q_i \text{ in } q} \underbrace{\sum_{i=0}^{l} \delta(d, d_i)}_{\text{count of } d_i \text{ in } d} \tag{5}$$

用式(4)和式(5)，通过 EM 算法可以获得词对词的翻译概率 $t(q_i \mid d_j)$ 表。 根据翻译列表，选择与查询候选次翻译概率较大的前 $n$ 个词扩充到查询中去。

### 3.3 训练集构造

在实际应用中，很难找到大量的<$q$,$d$>训练语料。为了解决这个问题，本文采用人工构造的查询句，与文档构成训练语料对。假设查询是由从文档中抽取的词组成的，什么样的词最有可能成为查询词？一种方法可以直接通过极大似然估计。

$$P_{\mathrm{ML}}(w_i \mid D) = \frac{c(w_i, D)}{\mid D \mid}$$

另一种方法是抽取那些与文档最相关的词，这可以通过互信息来衡量。本文选用后一种方法。

$$\begin{aligned} I(w, D) &= H(w) - H(w \mid D) \\ &= p(w, D) \lg \frac{p(w \mid D)}{p(w)} \propto p(w \mid D) \lg \frac{p(w \mid D)}{p(w)} \end{aligned} \tag{6}$$

按照式(6)计算文档中每个词与文档的互信息，然后根据每个词在文档中的互信息分布从中取出 $l$ 个词构成查询句。

## 4　实验结果与分析

通过对查询文档句对<$q$,$d$>进行统计翻译训练得出翻译概率表如表 1，其中 AP 训练集大约 120 万句对，SJM 训练集和 WSJ 训练集大约 70 万句对。从表 1 中可以看出这些词是统计意义上的关联关系，并且这些词反映了上下文的关系。 同时也用 WordNet 进行了相应的同义词查找，由于篇幅关系这里就不详细地列举。WordNet 只反映了被查询词的词义上的同义关系。可见通过翻译的方法能够获得 WordNet，HowNet 等一些辞典所不能表达的关联词。

本文在 3 个不同的 TREC 数据集上分别对本文提出的查询扩展方法进行了试验。 这 3 个 TREC 数据集的情况如下表 2。 所有文档按以下方式统一处理：用 Porter 对 term 进行 stemming；去除 stopwords。查询选用 TREC disk2 和 3 的 topics 50–100，在这些查询中只有 title 域的内容作为查询用。查询长度平均为 3，4 词。

本文采用的主要评测指标是平均准确率。

在本文的试验中，语言模型保持不变。只是查询相应的变化。目的就是考察查询变化对查询结果的影响。 首先，文档语言模型用 Dirichlet prior 平滑；然后，对查询模型作插值。

$$P(q_i \mid Q) = (1-\lambda) \sum_{q_j} P(e_i \mid q_j) P_{\mathrm{ML}}(q_j \mid Q) + \lambda P_{\mathrm{ML}}(q_i \mid Q) \tag{7}$$

$$P(q_i \mid D) = \frac{tf(q_i, D) + \mu P_{\mathrm{ML}}(q_i \mid C)}{\mid D \mid + \mu} \tag{8}$$

有两个平滑参数 $\lambda$ 和 $\mu$，根据经验设 $\lambda = 0.4$ 和 $\mu = 1000$。另外本文还作了一组对比试验-伪反馈。目的是与流行的查询扩展方法作一下比较。 试验结果如表 3，表 4，表 5 中：LM 表示传统的 unigram 语言模型，PFB 表示伪反馈，MTQE 表示基于统计翻译的查询扩展。Increase 显示 MTQE 与 LM 的比较。

通过以上几组 TREC 数据集的试验表明，与未进行查询扩展的语言模型相比，本文提出的方法在 3 个数据集上都有一致地大幅度的提高。就平均准确率而言，平均准确率在 AP 数据集提高 17.48%，WSJ 数据集提高 12.51%，SJM 数据集提高 16.90%。同时，从以上 3 个数据集可以看出，无论是哪个召回率点上，查询扩展都优于传统的语言模型。与

**表 1　翻译词表**

| law | | Investor | | industri | |
|---|---|---|---|---|---|
| 目标词 | 翻译概率 | 目标词 | 翻译概率 | 目标词 | 翻译概率 |
| law | 0.0484 | investor | 0.0163 | industri | 0.0257 |
| crimin | 0.0041 | stock | 0.0149 | steel | 0.0013 |
| anti | 0.0039 | bond | 0.0137 | fuel | 0.0013 |
| trademark | 0.0038 | note | 0.0137 | oil | 0.0012 |
| lawyer | 0.0036 | yield | 0.0131 | foreign | 0.0011 |
| patent | 0.0035 | rate | 0.0123 | semiconductor | 0.0011 |
| copyright | 0.0023 | market | 0.0104 | emiss | 0.0009 |
| gun | 0.0022 | treasuri | 0.0065 | auto | 0.0008 |
| court | 0.0019 | broker | 0.0059 | trade | 0.0008 |

**表 2 实验 TREC 数据集信息**

| 数据集 | 描述 | 平均文档长度 | 文档数量 | 词汇总数 |
|---|---|---|---|---|
| WSJ | Wall Street Journal (1990,1991, 1992) | 313 | 74,520 | 126,446 |
| SJM | San Jose Mercury News (1991) | 237 | 90,257 | 146,529 |
| AP | Associated Press (1988,1989,1990) | 261 | 158,240 | 194,799 |

**表 3 AP 数据集试验结果**

| Recall | LM | MTQE | Increase | PFB |
|---|---|---|---|---|
| Recall 0.00 | 0.6609 | 0.6876 | 4.04 | 0.6834 |
| Recall 0.10 | 0.5248 | 0.5705 | 8.71 | 0.5617 |
| Recall 0.20 | 0.4465 | 0.4879 | 9.27 | 0.4827 |
| Recall 0.30 | 0.3917 | 0.4350 | 11.05 | 0.4338 |
| Recall 0.40 | 0.3312 | 0.3922 | 18.42 | 0.3835 |
| Recall 0.50 | 0.2838 | 0.3372 | 18.82 | 0.3397 |
| Recall 0.60 | 0.2196 | 0.2721 | 23.91 | 0.2718 |
| Recall 0.70 | 0.1759 | 0.2283 | 29.79 | 0.2225 |
| Recall 0.80 | 0.1038 | 0.1527 | 47.11 | 0.1335 |
| Recall 0.90 | 0.0601 | 0.1001 | 66.56 | 0.0839 |
| Recall 1.00 | 0.0260 | 0.0364 | 40.00 | 0.0289 |
| **Avg. Pre.** | **0.2740** | **0.3219** | **17.48** | **0.3124** |

**表 4 WSJ 数据集试验结果**

| collection | LM | MTQE | Increase | PFB |
|---|---|---|---|---|
| Recall 0.00 | 0.6516 | 0.6804 | 4.42 | 0.6097 |
| Recall 0.10 | 0.5085 | 0.5273 | 3.70 | 0.5076 |
| Recall 0.20 | 0.4126 | 0.4227 | 2.45 | 0.4139 |
| Recall 0.30 | 0.3321 | 0.3655 | 10.06 | 0.3463 |
| Recall 0.40 | 0.2573 | 0.3015 | 17.18 | 0.3076 |
| Recall 0.50 | 0.1895 | 0.2454 | 29.50 | 0.2698 |
| Recall 0.60 | 0.1539 | 0.1828 | 18.78 | 0.2043 |
| Recall 0.70 | 0.1076 | 0.1324 | 23.05 | 0.1395 |
| Recall 0.80 | 0.0767 | 0.0963 | 25.55 | 0.1081 |
| Recall 0.90 | 0.0489 | 0.0663 | 35.58 | 0.0727 |
| Recall 1.00 | 0.0299 | 0.0366 | 22.41 | 0.0398 |
| **Avg. Pre.** | **0.2287** | **0.2573** | **12.51** | **0.2538** |

普遍采用的伪反馈相比，本文提出的基于统计翻译的查询扩展方法，也表现不错，在 AP 和 WSJ 数据集上略优于伪反馈，但是在 SJM 数据集上略低于伪反馈。

## 5 结论

针对信息检索中，短查询和长文档之间的不匹配问题。

**表 5 SJM 数据集试验结果**

| collection | LM | MTQE | Increase | PFB |
|---|---|---|---|---|
| Recall 0.00 | 0.5371 | 0.5780 | 7.61 | 0.5833 |
| Recall 0.10 | 0.3999 | 0.4741 | 18.55 | 0.5129 |
| Recall 0.20 | 0.3286 | 0.3941 | 19.93 | 0.4273 |
| Recall 0.30 | 0.2686 | 0.3117 | 16.05 | 0.3314 |
| Recall 0.40 | 0.2098 | 0.2288 | 9.06 | 0.2705 |
| Recall 0.50 | 0.1847 | 0.1996 | 8.07 | 0.2404 |
| Recall 0.60 | 0.1371 | 0.1585 | 15.61 | 0.1818 |
| Recall 0.70 | 0.1138 | 0.1353 | 18.89 | 0.1413 |
| Recall 0.80 | 0.0866 | 0.1123 | 29.68 | 0.1101 |
| Recall 0.90 | 0.0681 | 0.0886 | 30.10 | 0.0706 |
| Recall 1.00 | 0.0554 | 0.0687 | 24.01 | 0.039 |
| **Avg. Pre.** | **0.2006** | **0.2345** | **16.90** | **0.2484** |

本文提出了基于统计机器翻译的查询扩展方法。通过统计翻译模型，估算与查询词相关联的词，用以进行查询扩展。该方法考虑了词项在语料集中的全局统计信息,使得选取的扩展词与初始查询所表征的主题或概念具有更好的相关性。克服了传统的基于词典的查询扩展的缺少上下文关系的不足，试验表明，本文提出的方法能够大大地提高检索性能。 与未进行查询扩展的语言模型相比，分别提高了 12% ～17% 不等。与流行的基于伪反馈的查询扩展方法相比，本文提出的方法也具有可比的平均准确率。总的来说，本文提出的基于统计翻译的查询扩展的是有效的。

## 参 考 文 献

[1] Ponte J and Croft W. A language modeling approach to information retrieval. In Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval(SIGIR'98), Melbourne, Australia, 1998: 222–229.

[2] Richardson R and Smeaton A. Using wordnet in a knowledge-based approach to information retrieval. Trinity College Dublin, Working paper ca-0395, 1995.

[3] Lin D K and Zhao S J. Identifying synonyms among distributionally similar words. Proceedings of International Joint Conference of Artificial Intelligence (IJCAI2003), Mexico, 2003: 1492–1493.

[4] 丁国栋, 白硕. 一种基于局部共现的查询扩展方法. 中文信息学报, 2006, 20(3): 84–91.
Ding Guo-dong and Bai Suo. Local co-occurrence based query expansion for information retrieval. *Journal of Chinese Information Processing*, 2006, 20(3): 84–91.

[5] 吕碧波. 基于相关文档池建模的查询扩展. 中文信息学报, 2005, 20(3): 78–83.
Lv Bi-bo. Query expansion based on modeling of relevant documents pool. *Journal of Chinese Information Processing*.

2005, 20(3): 78–83.

[6] Xu J and Croft W. Query expansion using local and global document analysis. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 1996: 4–11.

[7] 张敏. 基于语义关系查询扩展的文档重构方法. 计算机学报, 2004, 27(10): 1395–1401.
Zhang Min. Document refinement based on semantic query expansion. *Chinese Journal of Computers*, 2004, 27(10): 1395–1401.

[8] Kang De L. Dependency-based evaluation of MINIPAR. Proceedings of the Workshop on the Evaluation of Parsing Systems, Granada, Spain, 1998: 298–312.

[9] Peat H and Willett P. The limitations of term co-occurrence data for query expansion in document retrieval systems. *JASIS*, 1991, 42(5): 378–383.

[10] Voorhees E. Query expansion using lexicalsemantic relations. ACM SIGIR, Dulin, Ireland, 1994: 61–69.

[11] Qiu Y and Frei H. Concept based query expansion. ACM SIGIR, Pittsburgh, PA, USA, 1993: 160–169.

[12] Bai J, Song D, Nie J Y, and Cao G. Query expansion using term relationships in language models for information retrieval. ACM CIKM, Bremen, Germany, 2005: 688–695.

[13] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. ACL, Cambridge, Massachusetts, USA, 1995: 403–410.

[14] Schjtze H and Pedersen J O. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, 1997, 33(3): 307–318.

[15] Berger A and Lafferty J. Information retrieval as statistical translation. In Proceedings of SIGIR'99, Berkeley, CA,USA, 1999: 222–229.

[16] 曹华梁, 朱星. 适用于 P2P 的系统查询扩展优化方法. 上海交通大学学报, 2005, 39(10): 1706–1710.
Cao Hua-liang and Zhu Xing. SDQE: A semantic query optimization in P2P system. *Journal of Shanghai Jiaotong University*, 2005, 39(10): 1706–1710.

[17] Brown P, Della Pietra S, Della Pietra V, and Mercer R. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*. 1993, 19(2): 263–311.

李卫疆： 男，1969 年生，博士生，研究领域为信息检索、网络信息处理.

赵铁军： 男，1962 年生，教授，博士生导师，主要研究领域为自然语言处理、网络信息处理、人工智能.

王宪刚： 男，1984 年生，硕士生，研究领域为信息检索.