跨语言信息检索中的查询扩展

郭文 史晓东 陈毅东

（厦门大学人工智能研究所，厦门 361000）

摘要：本文提出了把词典和马尔可夫随机域的潜在语义扩展相结合的新方法，充分的利用了现有词典资源，又克服了单一使用词典方法的局限性和单一使用语义扩展的不确定性，扩展出查询词的同义近义词，上下位词和潜在语义相关词，实验表明该方法能比较好的解决跨语言信息检索中翻译项的歧义问题。

关键字：跨语言信息检索，查询扩展，词典扩展，潜在语义扩展

Query Expansion in Cross-Language Information Retrieval

Guo Wen Shi Xiao-Dong Chen Yi-Dong

（Institute of Artificial Intelligence , Xiamen 361000）

Abstract: this paper proposes a new method which integrates with dictionary and Latent Concept Expansion Using Markov Random Fields .The method can make the best of the dictionary and overcome the shortcoming of using dictionary only and using Concept Expansion only .Expand original query with related , latent concept words and antonyms completely .The experimental result show that the new method can resolve the problem of different meanings in Cross-Language Information Retrieval..

Keywords: Cross-Language Information Retrieval, Query Expansion , Dictionary Expansion ， Latent Concept Expansion

1 引言

在当今的信息时代网上每天都有海量的数字化信息在生成存储传播和转换，这种趋势不可避免地加剧了信息获取的困难，同时语言障碍越来越成为一个严重的问题。跨语言信息检索(CLIR)提供了一种方便的途径使得用户能够使用自己熟悉的语言提交查询，检索另一种语言的文档。

目前很多研究集中在如何解决翻译项的歧义方面，通常使用的方法有基于语法分析的方法、基于统计的方法、基于查询扩展的方法。例如Davis在对25个西班牙语的查询条件利用Colins双语词典翻译成英语时提出了三种消除翻译歧义的方法;Balestero和Croft在对25个西班牙语的查询条件翻译前进行查询扩展与翻译后进行查询扩展的策略来消除翻译歧义[1]。

查询扩展指的是利用计算机语言学、信息学等多种技术，把与原查询相关的语词或者与原查询语义相关联的概念以逻辑或方式添加到原查询，得到比原查询更长的新查询，然后检索文档，以改善信息检索的性能，解决信息检索领域长期困扰的词不匹配问题，弥补用户查询信息不足的缺陷。

目前关键词查询扩展技术按照其扩展词的来源不同主要有全局分析、局部分析、基于关联规则的、和基于用户查询日志的查询扩展技术等几种[2]。目前采用比较多是以下两种扩展方式，其一是加入的扩展词与原始查询词意思相近，例如用户要检索“计算机”，用“电脑”、“微机”可以表达同样的概念；其二是扩展过程添加全新的词汇，例如用户键入“信息检索”，可以联想到“词频”、“相似度计算”等等。

本文受到国家自然科学基金（NO. 60573189），863 项目（NO. 2006AA01Z139，NO. 2006AA010108-3，NO. 2006AA010107），福建省重点科技项目（NO. 2006H0038），福建省基金项目（NO. 2006J0043）资助。

本文提出了把词典和潜在语义分析方法相结合的方法。其中词典的方法使用同义词林和HowNet获得查询词中的同义，近义和上下位词。潜在语义分析使用大规模文本用来获得与查询词有潜在语义关系的相关词。并将这两种方法扩展的词用布尔运算符连接在一起作为关键词来构造查询表达式，为下一阶段的机器翻译关键字提供充足的信息。这两种方法的结合充分的利用了现有词典资源，又克服了单一使用词典方法的局限性和单一使用语义扩展的不确定性，比较好的解决了跨语言信息检索中翻译项的歧义问题。

2 基于词典的查询扩展方法

2.1 原理

语言中大量存在的同义近义、上下义等语义关系使得用户的搜索意愿有多种不同的表达形式，用户向搜索引擎提交的只是其中的一种表达形式，而以其它形式表示的页面也应该包含在结果中[4]。例如，电脑与计算机、网络与网路是同义关系，当用户搜索“电脑”时，搜索引擎应该能够根据同义关系将与“计算机”相关的网页搜索出来；又如，搜索某人的作品，搜索引擎也应该可以根据“作品”与“著作”、“电影”、“电视剧”、“音乐”之间的上下位关系，把与作品相关的著作、电影、电视剧、音乐等网页也搜索出来。也就是说，从语义上对搜索任务进行一定的补充和扩展，使得搜索结果更加全面。

[4]中提出了一种基于语义单元的查询扩展方法，本论文的查询扩展主要是为了后一阶段机器翻译服务，考虑到复杂度效率等方面因素，这里并不构建语义单元，而是直接使用了同义词林和HowNet等语义词典来构建查询词的扩展词。以往的词典扩展方法只是单一的扩展同义词，这必然影响了扩展词的效果，因此本论文并不局限于扩展同义词。根据信息检索的特征，大部分检索涉及事物、事件、动作等类型，而不涉及修饰功能的词类型。所以重点分析如下几种关系：

（1）词的同义关系

根据语义单元的同义定义，类似的定义词的同义关系为：如果两个词在句中互相替换后不改变句子的意义，则称这两个词同义。例如：博客与部落格、网络与网路等[3]。

（2） 词的近义关系

动词之间很少有真正的同义关系，但是通常存在相同感情色彩或者相同态度等的近义关系，例如反对“台独”、打倒“台独”与批判“台独”等。

（3）词的上下位关系

指一个词相对于另一个词的有限等级，其中一个词是另一个词的次类。上下位关系产生层次语义结构，使得下位词可以继承上位词的所有特征/ 性质。例如，car 和vehicle 之间就是上下位关系，其中car 是vehicle的下位词，vehicle 是car 的上位词。

为了有效地表示这些语义关系， 构造如下的词的表示库：

编号	词	同义词	近义词	上/下位词
1	民	{r}	{n}	{q}
2	人民	{n+4}	{n+3}	{1}
…				
m	公民	{m+1}	{2}	{1}
m+1	国民	{m}	{2}	{1}

表一 词典表示库示例

2.2 算法设计

Begin

接收用户的初始查询串S t r;

取S E 的同义词集T(S t r) ,依次把T(S t r) 中的每一个同义词T (S t r) i加入到S E中,得到扩展搜索请求N ES t r = Σ T(S t r) i ;

取S E的近义词集J(V k) ,依次把J (V k) 中的每一个近义词J (V k) i 加入到S E中,得到扩展搜索请求V ES t r = Σ J (V k) i ;

取S E的上下义词集X (Nk),依次把X (Nk)中的每一个近义词X (Nk)i 加入到S E中,得到扩展搜索请求X ES t r =Σ X (Nk)i;

S t r 的扩展请求ES t r = S t r + N ES t r + V ES t r +X ES t r;

End

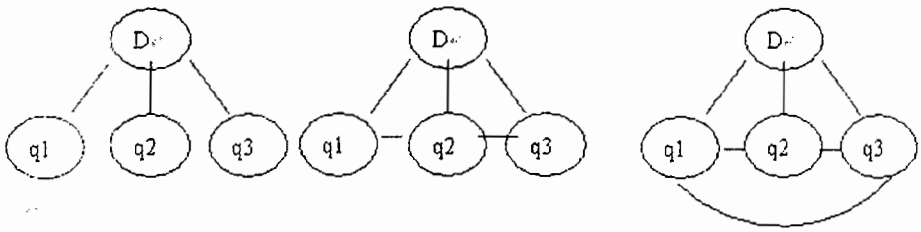
只要在前期利用词典构造出表示库,此算法可以在单位时间内实现扩展,提高了实际使用性。

3 基于马尔可夫随机域的潜在语义扩展

3.1 原理

过去大部分的倚赖模型在一致性和查准率上的效果都不好,而马尔可夫随机模型在应用中取得了很好的效果。马尔可夫随机域模型,也称做无向图模型,一般用在统计的机器学习领域来获得连接概率。在查询扩张领域,关注的是查询词Q = q1,...,qn和文档D的连接概率[3]。

在查询扩展的方面,基本的这里马尔可夫随机域模型如图二所示:



图二 具有三个查询项的随即域模型。(左) 完全独立;(中) 顺序倚赖;(右) 完全倚赖

这里我们构建三个潜在语义函数来获得P(D|Q);这三个函数为:

1) fT (c) = λ_T log P(qi|D)= λ_T log[(1-α) $\frac{tfqi,d}{|D|}$ + α $\frac{cfqi}{|C|}$]

这里,tfw,D是项w在文档D中出现的次数,|D| 是文档D中项的总数, cfw是项W出现在集合中的次数, |C|表示集合的长度[4]。

2) fo (c) = λ_O log P(qi。。。qi+k|D)

= λ_O log[(1-α) $\frac{tf\#1(qi..qi+k),D}{|D|}$ + α $\frac{cf\#1(qi..qi+k)}{|C|}$]

这里, tf#1 (qi,qi+k) ,D表示项qi...qi+k顺序的出现在文档D中的次数, |D| 是文档D中项

的总数，cf#1(qi,qi+k)表示项qi...qi+k顺序的出现在集合中的次数，|C|表示集合的长度[5]。

3) fu(c) = λ_U log P(#uwN (qi ... qi+k) |D)

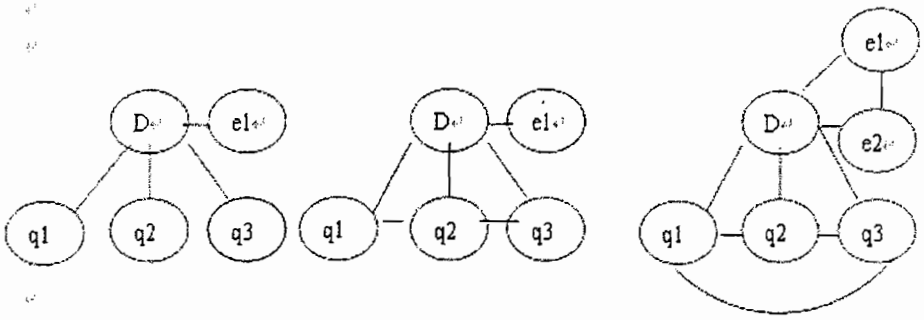
= λ_u log[(1-α) $tf\#uwN(qi..qi+k),D/|D| + \alpha cf\#uwN(qi..qi+k)/|C|$]

tf#uwN (qi,qi+k),D表示项qi...qi+k顺序或者非顺序的出现在窗口的次数，其余参数和上面定义的类似[6]。

有了这些潜在的语义函数，我们可以得到：

P(D|Q)= $\sum \lambda_{cf}(c) = \sum_{c\in T} \lambda_T f_T(c) + \sum_{c\in O} \lambda_{of_O}(c) + \sum_{c\in O\cup u} \lambda_U f_U(c)$，其中限制$\lambda_T + \lambda_O + \lambda_U = 1$，T是一个查询项和一个文档构成的集合，O 是两个或多个顺序出现的查询项和文档构成的集合，U表示两个或多个不按顺序出现的查询项和文档构成的集合。

潜在语义扩展（Latent Concept Expansion）模型正是在MRF(Markov Random Field)的基础上引入潜在词，扩展后的模型如图三：



图三 引入潜在语义的MRF模型

这里我们选择的D是相关文档或者伪相关文档，因此扩展后的概率公式为：

P(E|Q) $\approx \sum_{D\in R}$ exp[FQD(Q,D)+ FD(D)+ FQD(E,D)+ FQ(E)]，R表示相关文档或者伪相关文档的集合，这一集合可以通过MRF求出[8]。

3.2 潜在语义扩展的步骤

1) 选择适合的潜在语义的MRF模型，这里我们选择扩展使用单一项的模型，即图三中间的模型。

2) 选择K个潜在的项做为潜在的语义词进行扩展

3) 扩展词的概率如下：

$$P(e|Q) \propto \sum_{D \in R} \exp[\lambda_T \sum_{qi \in Q} \log[(1-\alpha) \frac{tfqi,d}{|D|} + \alpha \frac{cfqi}{|C|}]$$

$$+\lambda_o \sum_{qi..qi+k \in Q} \log[(1-\beta) \frac{tf\#1(qi..qi+k),D}{|D|} + \alpha \frac{cf\#1(qi..qi+k)}{|C|}]+$$

$$\lambda_u \sum_{qi..qi+k \in Q} \log[(1-\beta) \frac{tf\#uwN(qi..qi+k),D}{|D|} + \alpha \frac{cf\#uwN(qi..qi+k)}{|C|}]$$

$$+\log[(1-\alpha) (\frac{tfe,D}{|D|} + \alpha \frac{cfe}{|C|}) \lambda_{TD} ' / (\frac{cfe}{|C|}) \lambda_{TQ} ']]$$

实际运用中，考虑到采用的语料，设置的窗口大小为一个文档的长度，实验发现窗口的大小设置为一个句子长度，一个文档的长度和无限制，对扩展的结果的影响相差不大[5]。同时为了避免当Q中查询项长度为一时第二个和第三个潜在函数不起作用的问题，把潜在项做为查询项的一部分引入第二个和第三个潜在函数，此时修改概率计算公式如下：

$$P(e|Q) \propto \sum_{D \in R} \exp[\lambda_T \sum_{qi \in Q} \log[(1-\alpha) \frac{tfqi,d}{|D|} + \alpha \frac{cfqi}{|C|}]$$

$$+\lambda_o \sum_{qi..qi+k \in Q \cup e} \log[(1-\beta) \frac{tf\#1(qi..qi+k),D}{|D|} + \alpha \frac{cf\#1(qi..qi+k)}{|C|}]$$

$$+\lambda_u \sum_{qi..qi+k \in Q \cup e} \log[(1-\beta) \frac{tf\#uwN(qi..qi+k),D}{|D|} + \alpha \frac{cf\#uwN(qi..qi+k)}{|C|}]$$

$$+\log[(1-\alpha) (\frac{tfe,D}{|D|} + \alpha \frac{cfe}{|C|}) \lambda_{TD} ' / (\frac{cfe}{|C|}) \lambda_{TQ} ']]$$

实验表明修改后的能获得更好的结果。

4 两种方法的结合

通过把基于词典的查询扩展方法和基于马尔可夫随机域的潜在语义扩展方法相结合，我们能够实现对查询词的全方位的扩展。

我们采用以下的结合步骤：

a) 获得查询字符串，进行分词获得查询项。

b) 对查询项同时使用基于词典的查询扩展方法和基于马尔可夫随机域的潜在语义扩展方法进行扩展。

c) 原查询项和扩展词用布尔运算符连接在一起作为关键词来构造查询表达式，期间记录扩展的词的来源，例如属于同义近义词，上下位词或相关词，为下一阶段翻译做准备。

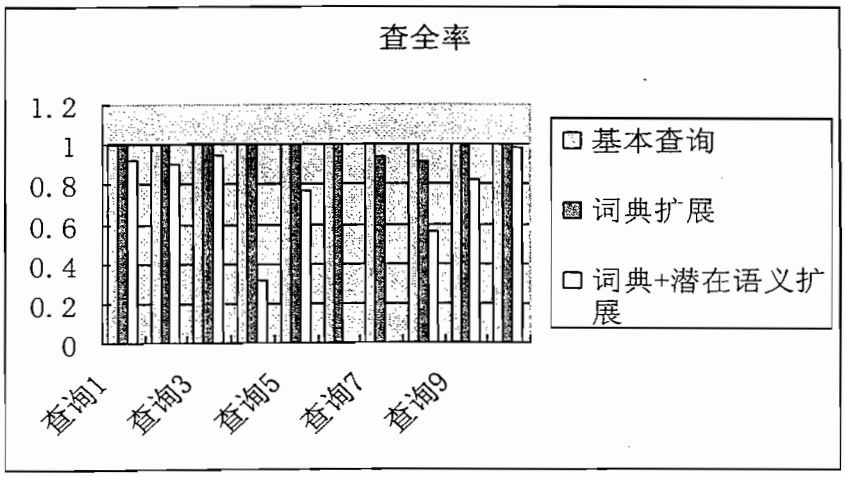
5 实验设计和实验结果分析

5.1 实验所用资源和语料预处理

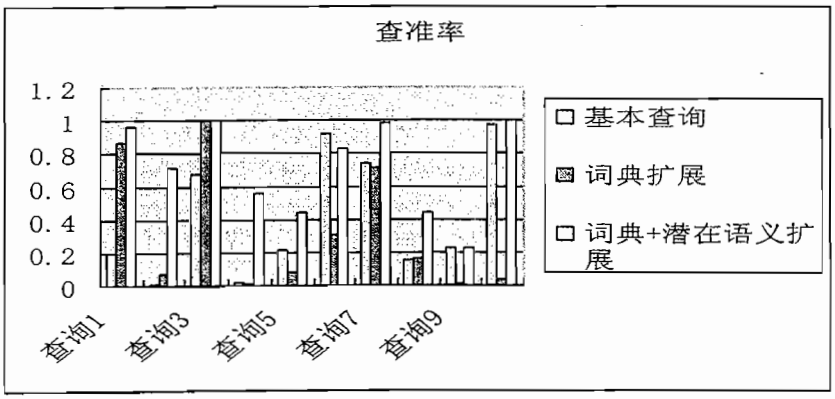
为了测试基于本文提出的查询扩展检索性能，采用同义词林，HowNet2002和人民日报一月份19483个文档做中文扩展实验。设计了10 个带有歧义的查询(Q1 ，Q2 ， …，Q10)作为查询集供实验用。对原始训练文档集经过分词、去掉停用词等文档预处理，提取每一篇文档的特征词和整个文档集的总特征词库以及相应的特征词出现次数。

5.2 实验结果及其分析

本文把不基于任何扩展的查询和单独基于词典的方法以及词典加潜在语义扩展的方法进行检索性能比较。实验采用google搜索引擎对所设计的10个查询进行检索，由于google检索返回量巨大，只取返回的前100个人工统计结果，计算其查全率和查准率。



图四



图五

从图四可以看到，基本查询的查全率要高于词典扩展和词典加潜在语义扩展，这是因为词典扩展和词典加潜在语义扩展都额外的引入了扩展词，词典扩展的查全率要高于词典加潜在语义扩

展这是因为词典扩展主要扩展的是同义近义词或上下位词，而词典加潜在语义扩展还引入了相关词，而相关词和原查询的相关度一般要小于同义近义词。

本方法看似查全率下降了，但其实本实验在google查询中，查询词之间使用的是与的关系，这与一般的实验采用或的评测是不相同的，这样做的目的是为了控制查询结果的数量从而更准确的预测查准率，而查准率对于本项目来说更有意义。从图五中可以看到词典加潜在语义扩展的方法的查准率要明显好于基本查询和基于词典扩展的方法，这是因为词典加上潜在语义扩展的相关词后，能比较好的消除歧义，提高查准率。

不过实验中也看到了查询6的词典加潜在语义扩展的方法的查准率并没有取得很好的结果，分析原因主要是由于语料库数据稀疏导致没有获得很好的相关词的。

此论文的扩展方法是要应用于跨语言检索中，查准率的提高也证明了只要词典和语料库足够丰富是能够为查询词翻译提供充足的排歧信息，从而能比较好的提高跨语言检索的性能。

参考文献

1. 吴　丹，李瑞芬．跨语言信息检索技术应用与进展研究．情报科学 Vol．24，No. 9

2. 黄名选等．查询扩展技术进展与展望．计算机应用与软件，Vol124 No. 11 Nov. 2007

3. 聂建云．实现一体化的跨语言和多语言信息检索．数字图书馆论坛 2006年第9 期

4. 李　莉　高庆狮．一种基于语义单元的查询扩展方法．计算机科学2008Vol135 №12

5. Donald Metzler, W. Bruce Croft. A Markov Random Field Model for Term Dependencies. SIGIR'05, August 15.19, 2005

6. D. Metzler, T. Strohman, H. Turtle, and W. B. Croft.Indri at terabyte track 2004. In Text Retrieval Conference (TREC 2004), 2004.

7. C. Zhai and J. La_erty. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proc. 24th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 334{342, 2001.

8. D. Metzler. Direct maximization of rank-based metrics. Technical report, University of Massachusetts,Amherst, 2005.

9. Donald Metzler, W. Bruce Croft. Latent Concept Expansion Using Markov Random Fields. SIGIR'07, July 23–27, 2007