一种适用于语言模型的检索词扩展方法

张斌 周延泉
北京邮电大学
buptbin@gmail.com zhouyanquan@bupt.edu.cn

摘要： 近年来信息检索有了越来越广泛的应用，如何给用户提供准确的信息也显得尤为重要。对于给定的主题判断某一句话或某一篇文章是否与所给主题相关以及相关的程度是信息检索的重要基础。目前判断相关性的方法有很多，其中基于关键词的相关性识别是常用的方法，该方法的优势是实现容易且准确率较高。我们把这里的关键词称为检索词，检索词的好坏将直接影响最后的检索结果。一般情况，只从原始的查询语句提取检索词往往不能得到满意的检索结果，还需要采用一定方法对原始的检索词进行扩展。本文通过对比试验，比较了几种不同方法对检索词进行扩展的结果，并给出了一种有效的适合语言模型的扩展检索词的方法。

关键词： 信息检索 相关性判断 检索词 扩展词

One Method for Expanding Search Terms in Language Model

Bin Zhang Yanquan Zhou
Beijing University of Posts and Telecommunications
buptbin@gmail.com zhouyanquan@bupt.edu.cn

Abstract :Recent years, information retrieval has more and more applications. How to provide accurate information to users is especially important. For the given topic, detecting whether a sentence or a document is relevant to the topic is the base of information retrieval. There are kinds of methods. And key-words based is a normal one which achieves higher accuracy. We define key-words here as query terms which play an important role to the result. Generally speaking, only extracting query terms from the original search sentence often does not return satisfactory results. In order to improve the results we usually do query expansion. This paper has compared several query expansion methods and gives an effective one for the language model.

Keywords: Information Retrieval, Relevance Detection, Query Terms, Query Expansion

1.引言

检索的基本任务就是根据用户的输入，在给定的文本集合里返回一个与用户输入相关的子集。在信息检索领域，相关性反应的是检索结果与用户输入的查询在语义上的符合度，也可称作句子的相似度。基于关键词的相似度判断是一种常用的方法，通常把用户给定的主题进行分词处理，然后提取名词、动词等语义丰富的词作为关键词配合检索模型进行检索。

语言模型在自然语言处理与语音识别领域受到了广泛关注，成功的被应用于机器翻译、中文分词、语音识别等领域[1]， 1998年Ponte和Croft[2]将语言模型(Language Model，LM)应用到信息检索(Information Retrieval，IR)中，从此，语言模型成为信息检索领域常用的方法之一。本文采用语言模型作为相关性判别模型，由于句子较短，三元词串及大于三元的词串在句子中出现的概率较小，所以本文取三元语言模型。

本文接下来按以下几个部分组织：在第二部分是相关工作介绍，第三部分介绍了对给定的关键词进行扩展的方法，第四部分介绍主题相关性判断模型，第五部分是实验方法及流程，第六部分是实验结果并对结果进行比较分析，第七部分是实验结论。

2. 相关工作：

NTCIR-6 新增了从给定的文本中找出与主题相关的句子的任务[3]，该任务可以理解成句子级的检索任务，即从给定的文章里找出与所给主题相关的句子集合。Evans 使用了标准向量空间模型，采用 TFIDF 方法计算权重得到了较好的结果[4].Li et al 通过计算主题特征向量与句子特征向量的内积来判断句子的相关性[5]。国内有学者利用骨架依存的方法计算汉语句子间相似度 [6]. 李彬等学者采用了语义依存计算句子相关性 [7],哥伦比亚大学的 Goldsdein 等人通过最大边缘相关的方法（Maximal MarginalRelevance）进行相似度计算[8], 学者 Chris H.Q.Ding. 等采用了隐含语义索引（Latent Semantic Indexing）的方法[9] ,Younghc Kim 使用了语言模型方法来判断句子的相关性[10]，本人认为该方法的关键在于检索词的扩展，针对该语言模型方法，在提取扩展词时应注意既要有一元词串又要有二元及三元词串，同时要考虑各自的比重，本文通过 TFIDF 提取一元词串，通过多词表达提取了二元及三元词串，实验结果表明这样处理的结果比只有一元词串或只有二元词串更有效。

3. 几种对主题词进行扩展的方法

本文主要讨论句子级的主题相关性判断，由于句子相对于文章来说比较短，存在稀缺性问题，要是语言模型能很好的发挥作用需要对关键词做一定扩展。

3.1 利用维基百科与 PMI 算法

首先利用维基百科找出与所给对象的相关词条，具体实现为将给定对象输入维基百科，从返回结果中提取出词条作为候选扩展词，然后利用 Goggle 计算每个候选扩展词与原始查询词之间的 PMI 值，公式如下：

$$PMI=\frac{C(exp,topic)}{C(exp)*C(topic)} \qquad (1)$$

其中 $C(\exp,topic)$ 是扩展候选词与原始 topic 同时输入 Goggle 搜索引擎返回的结果数，$C(\exp)$ 与 $C(topic)$ 分别表示只将候选扩展词、topic 输入 Google 搜索引擎返回的结果数，PMI 值越大说明候选扩展词与给定 topic 的相关度越大。

该方法的优点是不需要收集训练语料，实现简单，扩展词可能包含一元、二元及多元词串，而不全是一元词语。缺点是对某些主题维基百科可能没有收入，对这样的主题该方法不

适用，而对于维基百科已收入的主题，可能存在候选词条较少且质量差的问题。

3.2 利用搜索引擎与 TF-IDF 算法

TF-IDF（Term Frequency–Inverse Document Frequency）是一种基于统计的方法，用来衡量在文本集或语料库中，某个字词对给定文本的重要程度，字词的重要性与其在给定文本中出现的次数成正比，与其在文本集或语料库中出现的频率成反比。TF-IDF 基本原理如下：

TF（Term Frequency）词频指某个词语在给定文本中出现的频率。对于给定文件 d 里的词语 t 来说，它的重要性可以表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

其中 n_{ij} 是该词在文本 d 中出现的次数，分母是文本 d 中所有词出现的次数之和。

IDF（Inverse Document Frequency）逆向文件频率是一个词语普遍重要性的度量。其计算方法如下：

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{3}$$

其中分子|D|表示文本集或语料库中的文本总数，分子是文本集或语料库中包含词语 t 的文本总数。

最后的权重表示为：

$$w_{i,j} = tf_{i,j} \cdot idf_i \tag{4}$$

采用该方法对主题词进行扩展需要大量的文本作为文本集或语料库，本文取 NTCIR-7 的测试语料作为语料库，然后通过 TFIDF 算法计算每个主题的扩展词。由于权重越大表明该词对于给定文本的重要越高，所以，对于每一个主题都选权重较大的词作为扩展词即能保证该扩展词与给定主题的相关性较高。

该方法的优点是得到的扩展词与给定主题相关性较大，比较准确，算法简单易实现。缺点是需要大量训练语料，且只能得到一元词串的扩展词。

3.3 采用多词表达方法

目前多词表达还没有一个统一的定义，很多学者都有不同的意见。Timothy Baldwin对英文多词表达定义为 1. Decomposable into multiple simplex words 2. Lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic。一般中文多词表达可以指具有语法、语义或语用特性的，语义完整的多个词的组合。多词表达在自然语言处理中已有很多应用，比如机翻译习，词义消歧，信息检索等。本文采用LLR-based Hierarchical Reducing Algorithm (HRA for short)[11]提取多词表达。该算法把句子中的每个词定义为一个单元，通过计算相邻单元之间的对数似然(Log Likelihood Ratio)进行单元消减，若该值大于阈值，则这两个单元可以构成一个多词表达，该算法按以下循环进行：1. 计算句子中所有相连单元的对数似然，并选出最大的；2. 将对数似然最大的两个相邻单元合并为一个单

元。该循环停止的条件是选出的最大对数似然值小于阈值或者该句中只剩一个单元。

该算法的优点在于可以提取任意长度的多词表达，即可以提取出二元、三元词串作为扩展词。

4. 主题相关性判断模型

在信息检索领域，语言模型方法是一种常用的方法，实验表明，该方法同样适用于句子级的相关性判断，不过需要一些扩展。

本文采用了引入扩展的语言模型方法判断句子与主题的相关程度，某句话 S 与给定主题的相关性 $P(Q \mid s)$ 可以用如下公式表示[10]：

$$P(Q \mid s) = \sum_{\forall Q'} P(Q \mid Q', s) P(Q' \mid s) \approx \sum_{\forall Q'} P(Q \mid s) P(Q' \mid s) \tag{5}$$

其中 Q' 表示每个主题的扩展检索词集合，Q 表示给定的主题，S 表示待判断的语句。这里只选一个扩展词集合 Q'，对上式两边同时取对数，有

$$\begin{aligned}
&\log P(Q \mid s) \approx \log p(Q \mid s) p(Q' \mid s) \\
&\approx \log p(Q \mid s) + \log p(Q' \mid s) \\
&\approx \log \prod_{t \in Q} p(t \mid s) + \log \prod_{t' \in Q'} p(t' \mid s) \\
&\approx \lambda \log \prod_{t \in Q} p(t \mid s) + (1-\lambda) \log \prod_{t' \in Q'} p(t' \mid s)
\end{aligned} \tag{6}$$

其中 t 为给定检索词中的词条，t' 为扩展词集合 Q' 中的每个词条，λ 为权重系数，可以通过调整 λ 的值来调整原始主题词与扩展词之间的权重，λ 的大小，取决于原始主题词于扩展词所占的比重，通过实验表明 λ 为 0.5 时结果较好。通过上式的变形可以引入权重调节原始主题词与扩展词对结果的影响。例如，主题为“911 恐怖袭击以后，美国的经济情况”，先对主题进行分词等处理，保留名词、动词等具有实际意义的词作为原始检索词，例如 911，恐怖袭击等，由这些词组成集合 Q，其中每一个词为一个 t，然后再采取一定算法对该主题进行扩展，扩展得到的词串集合为 Q'，其中的每个词串为一个 t'。假设一句话 S 与给定主题的相关度与该主题在这句话里出现的频率成正比，所以

$$p(t \mid s) = \frac{c(t)}{n} \tag{7}$$

其中 $c(t)$ 为词串 t 在句子 S 中出现的次数，n 为句子的长度。该假设存在实际意义，且该式考虑了句子长度对结果的影响。若 c(t)为零，则给 c(t)赋予一个极小值，这样可以保留句子长度对相关性的影响。若给 $p(t \mid s)$ 赋予一个极小值则忽略了句子长度对相关性的影响。我们选取三元语言模型，所以这里的 t' 包括一元、二元、三元词串，则

$$\mathrm{P}(Q \mid s) = \alpha * P_{uni}(Q' \mid s) + \beta * P_{bi}(Q' \mid s) + \gamma * P_{tri}(Q' \mid s) \tag{8}$$

这里 α，β，γ 分别表示一元词串、二元词串、三元词串的权重，$\alpha + \beta + \gamma = 1$，$\alpha$，

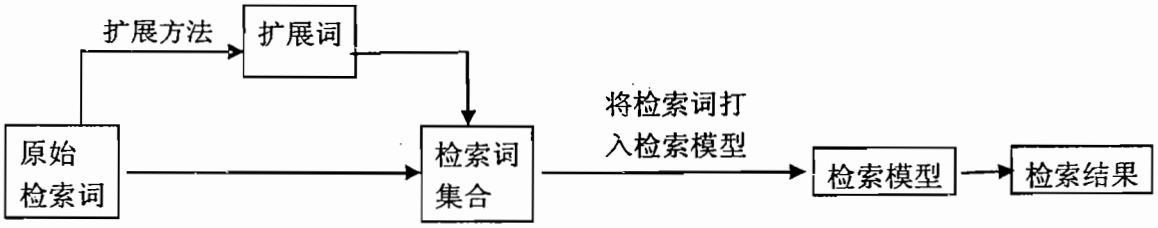
β，γ 的取值取决于一元、二元、三元词串在 Q 及 Q' 中的比重。

由该模型我们可以看出若一句话包含的检索词及扩展出来的检索词越多，则 $P(Q \mid s)$ 的值越大，也就证明该语句与所给主题的相关性越大。通过训练语料，可以训练模型参数及阈值，使该模型最优化。

5. 实验方法及流程

由上面的分析可以看出，由于句子较短，存在词语稀缺性问题，所以本文分别采用了 PMI，TFIDF 及多词表达的方法对检索词做了扩展，采用三元语言模型，经分析我认为要使语言模型发挥较好的作用，需要均衡一元，二元及三元检索词的比重，一元检索词出现在句子中的概率较大，但一元检索词包含的信息相对较少，无法保证准确性，例如“经济”这个词，既可以指中国经济，也可以指美国经济，但二元及三元检索词就可以很好的避免此问题。所以适当选取二元、三元检索词可以提高检索的准确率，但二元及三元检索词在句子中出现的概率比一元小很多，所以若二元、三元检索词过多可能会出现大多数检索词没有在句子中出现的情况，这样该语言模型也无法发挥作用，所以一元、二元、三元检索词的比重很重要。本文意在通过实验给出对语言模型的训练以及对检索词进行扩展的有效方法。

实验流程如下：



为了很好的进行结果比较，本文设计了两组试验，第一组实验分别采用 PMI, TFIDF, 多词表达及 TFIDF 与多词表达融合的方法扩展检索词，以比较不同的扩展方法对结果的影响；第二组实验对检索词的数量进行了调整，以探求检索词数量不同对结果的影响。

6. 实验结果及分析

由上面介绍的判别模型可以看出，扩展词的是否跟原始的查询词紧密相关对实验结果影响较大，所以扩展词也是该方法的关键，本文分别采用上面介绍的不同扩展方法进行了实验。实验的训练集采用的是 NTCIR-7 的标注答案，共 16 个主题，每个主题的文章平均为十几篇。选取联合早报的评论文章作为测试集。

实验分为二组，No.1 为采用不同的扩展方法的实验结果，扩展词取 20 个；N0.2 为取不同数量扩展词得到的实验结果,扩展词的获取方法采用 TFIDF 算法与多词表达方法结合。

同样取“911 恐怖袭击以后，美国的经济情况”这个主题为例，训练集为 NTCIR-7 关于该事件的 20 篇评论文章，测试集为联合早报的评论文章，共 204 个句子，其中相关的句子为 105 句，不相关的句子 99 句。

实验结果如下：

No.1:

	P@10	precision	Recall	F
PMI_expan	0.8	0.8217	0.5436	0.6543
TFIDF_expan	0.9	0.8593	0.5949	0.7031
MWE_expan	0.8	0.8515	0.5590	0.6749
TF_MWE	0.9	0.8923	0.5949	0.7139
No_expan	0.7	0.7801	0.5436	0.6407

Table.1 不同扩展方法的实验结果

其中 PMI_expan 表示采用维基百科与 PMI 算法进行扩展的结果；TFIDF_expan 表示使用 TFIDF 算法得到的结果；MWE_expan 表示采用多词表达方法得到的结果；TF_MWE 表示采用 TFIDF 算法与多词表达方法结合进行扩展的结果；No_expan 表示不进行扩展得到的结果。

由结果可以看出，没有进行扩展的结果最差，说明采用引入扩展的方法对句子级的主题相关性判断是有效的；PMI_expan 的结果较差，主要原因是维基百科的候选扩展词数量较少且质量较差；TFIDF_expan 比 MWE_expan 的结果好一些，主要原因是 TFIDF 算法获得的扩展词主要为一元词串，而多词表达方法获得的扩展词主要为二元词串，在被检索的句子中，一元词串出现的频率要高于二元词串，所以采用一元词串进行检索得到的结果要好于二元词串；TF_MWE 的结果最好，主要因为 TFIDF 算法与多词表达方法结合平衡了一元词汇与二元词汇的数量（由于三元词串较少且权重最低影响较小不予考虑），使主题相关性判断模型真正发挥了优势。

No.2:

	P@10	precision	Recall	F
15(8,6,1)	0.9	0.8647	0.5897	0.7012
20(10,9,1)	0.9	0.8923	0.5949	0.7139
25(12,12,1)	0.9	0.8976	0.5846	0.7080
30(17,13,1)	0.9	0.8667	0.6000	0.7091

Table.2 扩展词数量不同的实验结果

15(8，6，1) 表示扩展词总数为 15，其中 8，6，1 分别表示一元词串、二元词串与三元词串的数量。由实验结果可以看出，当扩展词数量在 15-30 之间变化时对结果的影响较小，在扩展词数量在 20 左右时取得最好结果。

P@10 最高值在 0.9 的主要原因是部分语句，比如描述 911 恐怖袭击以后中国经济的情况，由于这样的语句会含有较多扩展词，对于这样的句子上述主题相关性判断模型会将其判为相关，且得分较高，这是该模型的缺陷。

7. 实验结论

通过对比实验我们可以得出这样的结论：对主题词进行扩展会明显提高实验结果；对检索词的扩展可以降低由于句子较短而引起的稀缺性影响；上述相关性判断模型的关键在于对

检索词的准确扩展以及扩展词中一元、二元、三元词串的均衡；扩展词数量在 15-30 之间变化时对结果的影响较小，当扩展词取 20 个时结果相对较好。

参考文献

[1] 楼炉群，牛军钰. 信息检索中语言模型的研究. 软件技术与数据库，第 33 卷，第 4 期，2007 年 2 月。

[2] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval[C]//Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998: 275-281.

[3] Y. Seki, D. Evans, L. Ku, H. Chen, N. Kando, and C. Lin. Overview of Opinion Analysis Pilot Task at NTCIR-6. In Proceedings of the 6th NTCIR Workshop Meeting, Japan, 2007.

[4] D. Evans. A low-ressources approach to Opinion Analysis: Machine Learning and Simple Approaches. In Proceedings of the 6th NTCIR Workshop Meeting, Japan, 2007.

[5] Y. Li, K. Bontcheva and H. Cunningham. Experiments of Opinion Analysis on the Corpora MPQA and NTCIR-6. In Proceedings of the 6th NTCIR Workshop Meeting, Japan, 2007.

[6] 穗志方，俞士汶. 基于骨架依存树的语句相似度计算模型. 中文信息处理国际会议（ICCIP'98）. 1998

[7] 李彬，刘挺，秦兵，李生. 基于语义依存的汉语句子相似度计算. 哈尔滨工业大学信息检索研究室论文集，第一卷，1003.

[8] Jaime Carbonell and Jade Goldsdein. The use of MMR, dirversity-based reranking for recording documents and producing summaries. In Proceedings of ACM-SIGIR'98, Melbourne Australia, August 1998.

[9] Chris H. Q. Ding. A Similarity-based Probability Model for Latent Semantic Indexing. Proceedings of 22nd ACM SIGIR Conference, pp. 59-65. August 1999.

[10]Youngho Kim, Seongchan Kim, Sung-Hyon Myaeng. Extracting Topic-related Opinions and their Targets in NTCIR-7. In Proceedings of the 7th NTCIR Workshop Meeting, Japan, 2008.

[11] Zhixiang Ren, Yajuan Lv, Jie Cao, Yun Huang. Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions. 2009 Workshop on Multiword Expressions.