文章编号: 1003-0077(2007)02-0035-11

自然语言处理在信息检索中的应用综述

王灿辉,张 敏,马少平

(清华大学 计算机科学与技术系,北京 100084)

摘 要: 在信息检索①发展的过程中,研究者们不断尝试着将自然语言处理应用到检索里,希望能够为检索效果提高带来帮助。然而这些尝试的结果大多和研究者们最初的设想相反,自然语言处理在大多数情况下没有改进信息检索效果,甚至反而起了负面作用。即便有一些帮助,也往往是微小的,远远不如自然语言处理所需要的计算消耗那么大。研究者们对这些现象进行了分析,认为: 自然语言处理更适合于应用在需要精确结果的任务中,例如问答系统、信息抽取等;自然语言处理需要针对信息检索进行优化才可能发挥积极作用。最新的一些进展(例如在语言模型中加入自然语言处理)在一定程度上印证了这一结论。

关键词: 人工智能;自然语言处理;综述;信息检索

中图分类号: TP391 **文献标识码**: A

A Survey of Natural Language Processing in Information Retrieval

WANG Can-hui, ZHANG Min, MA Shao-ping

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Natural language processing (NLP) has been used in information retrieval (IR) by researchers, in the hope of improving retrieval effect. But most of the results are in the opposite way hypothesized. In most cases, NLP didn't yield increases in IR precision but took a negative effect. Even if NLP helped IR under some circumstances, the improvements were much smaller than the processing cost needed by NLP. Researchers perform analysis on these phenomena and come to the conclusion that: IR-related tasks that acquire accurate results, such as question answering (QA) and information extraction (IE), are more suited for the use of NLP. NLP needs to be optimized for IR in order to be effective. Recent research, e.g. adding NLP factors to language model, has more or less confirmed the conclusion.

Key words: artificial intelligence; natural language processing; overview; information retrieval

1 引言

基于全文索引的信息检索发展至今已有十几年的历史。在这十几年里,研究者们不断尝试着将自然语言处理应用到信息检索中,试图提高信息检索的效果。自然语言处理包括自然语言处理技术和自然语言处理资源。在信息检索中使用自然语言处理技术的尝试大部分没有获得好的效果。尽管在小部分实验中信息检索效果有了一些提高,但改进的程度往往很小,为此而使用的复杂的自然语言处理技术则有着巨大的计算消耗,很难被认为是值得的[1]。在信息检索技术中结合自然语言处理资源,例如词典,实验结果也不能令人满意[2]。

信息检索中常常使用到的自然语言处理技术包括去除停止词、取词根、词性标注、词义消歧、句法分析、命名实体识别、指代消解等,自然语言处理资源

收稿日期: 2006-06-04 **定稿日期**: 2006-11-14

基金项目: 国家 973 重点基础研究资助项目(2004CB318108);国家自然科学基金资助项目(60621062,60503064);国家 863 计划资助项目(2006AA01Z141)

作者简介: 王灿辉(1981—),男,博士生,主要研究方向为信息检索、机器学习。

①如果不加特别说明,本文中的信息检索是狭义的,即文档检索(Document Retrieval)。

包括的则是 WordNet[3]和 HowNet[4]这样的词典。

自然语言处理技术被用来对自然语言进行处理,目的是让计算机"理解"自然语言的内容。而信息检索中所涉及的文档和查询都是用自然语言描述的,因此,在信息检索中使用自然语言处理以提高其效果的想法被寄予了厚望。信息检索可以看作是用查询和文档内容进行匹配的过程,匹配的单位通常是查询和文档中的词。基于词匹配的信息检索中存在着与自然语言特点相关的问题,同样促使研究者们求助于自然语言处理[5]:

· 不同的词可以表达同一个意思
· 同一个词可以表达多种意思
· 对一个概念的描述可以有不同的角度
· 同一个词在不同的领域也会有不同的意思

自然语言处理技术最大的难点在于自然语言中有各种级别的歧义难以消除,包括词汇级别、句法级别和语义级别[5]。歧义的存在使计算机在"理解"自然语言时发生了困难,并很可能出现了错误。这无疑为自然语言处理没能为信息检索带来较大帮助提供了一个解释。然而事实上这个解释并不全面。因为和信息检索的效果相比,自然语言处理的很多技术实际上已经有了很高的准确率——尽管直接用两者的准确率进行比较并不科学。

因此,本文对信息检索中使用自然语言处理的研究工作进行综合分析,总结出哪些自然语言处理技术和资源对信息检索有帮助,需要达到怎样的精度才能使信息检索的效果有较大提高,并试图对未来自然语言处理在信息检索中的使用方向进行归纳和展望。

本文按照如下方式组织:第二部分介绍自然语言处理在信息检索中的应用情况;第三部分对自然语言处理对信息检索帮助不大的原因进行分析;第四部分是对未来自然语言处理在信息检索中使用的归纳和展望;第五部分是总结。

2 自然语言处理在信息检索中的应用

自然语言处理包括自然语言处理技术和资源。技术又可分为基本和高级两种,这个分类一方面根据自然语言处理的深度和层次,另一方面则考虑了技术的复杂性和难度。自然语言处理资源主要指的是机器可读的词典。

2.1 基本自然语言处理技术的应用

基本自然语言处理技术包括去除停止词、分词、取词根和词性标注等。

2.1.1 去除停止词(Stopword)

停止词指的是在文档中出现次数很多而本身没有实际意义的词,例如英文中大部分的介词、冠词等。去除停止词常被用在信息检索系统中,作为文档预处理的一个步骤。通常使用一个停止词表来过滤,并可根据实际的文档集合选择合适的停止词表。

实际使用的信息检索系统例如 Web 搜索引擎中往往不采用去除停止词这一技术,因为它对于检索效果的提高并没有实质上的帮助,反而可能导致在处理一些查询时得不到好的结果。经典的例子就是"to be or not to be"这个查询。因此,在大多数实际检索系统中停止词也被作为索引项保留下来。

信息检索实验系统中则通常会去除文档中的停止词。尽管仍然不能处理实际系统中可能遇到的一些特殊查询,但完全可以通过实验设置来避免。去除停止词虽然对提高检索效果帮助很小,但可以提高检索效率,这对于实验系统来说已经很有价值了。

2.1.2 分词

分词是中文、日文等亚洲语言的信息检索中遇到的特殊问题,大多数欧洲语言并不需要分词。分词技术被广泛应用在中文信息检索系统中。

Peng 等在 TREC5 和 TREC6 的中文数据集上进行分词和检索实验[6]。该数据集包括《人民日报》的 164 768 个新闻报道,139 801 篇文章以及来自新华社的 24 998 篇报道。使用的 54 个查询同样来自 TREC5 和 TREC6。他们的实验表明,分词精度和检索效果并不是单调正比的关系。分词精度在 70 %左右时可获得最佳的检索效果,如果分词精度太高,反而可能导致检索效果下降。例如他们使用的分词精度最高的一组结果中,"农作物"被作为一个词,没有被分成"农"和"作物",从而无法和包含"作物"的查询相匹配,因此没能作为相关文档返回。事实上,正如文献[7]中指出的,Peng 等的实验结果是因为他们所用的索引方式不是现代信息检索系统常用的字词结合方式。

Foo 等使用 1998 年 3 月《人民日报》经济版的 266 个文件(共 532 KB)来测试分词精度和检索效果的关系。他们在实验中得出了以下结论[8]:

1. 尽管分词精度和检索结果没有直接的联系,但不同的分词方法对检索结果确实是有影响的;

2. 对查询和文档使用一致的分词方法时能获得较好的检索效果,一致性比分词精度对检索效果更为重要。只要保持一致性,即便使用最简单的二

元语法(Bigram)分词,也能获得与使用手工分词相当的检索效果;

3. 手工分词的检索效果并不比自动分词好。原因可能是手工分词结果尽管把词的含义表达得更精确,但缺乏灵活性,在查询和文档使用不同方式表达同一概念时往往无法将它们匹配上;

4. 使用简单的二元语法分词会得到大量意义不明的词,但在实验中这一分词方法并没有对检索效果产生明显的不好影响。

然而,正如 Foo 等自己指出的,他们的实验所用的数据集太小,得出的结论并不很可靠。特别是第 3 点中的手工分词缺乏灵活性,实际上只要采取恰当的索引策略就可以有很大程度的改进。而第 4 点中二元语法分词得到的意义不明的词,很有可能会在检索中返回许多不相关文档而降低检索精度。

金澎等利用从《人民日报》和中国新闻社获取的 8 万余网页数据对不同的分词算法和词典进行了一系列分词和检索实验[7]。检索系统同时使用字和词作为索引单位。查询集合包括从 TREC5 和 TREC6 中选出的 19 个查询和作者自行设计的 5 个查询。他们的实验结果表明,精度最高的分词方法对应的检索效果最好,但对于同一个分词算法,最好的检索效果未必是和分词性能最好的词典相对应。

2.1.3　取词根(Stemming)

取词根能够使具有相同词根而形态不同的词匹配上。常用的取词根方法包括基于规则(例如 Porter Stemmer)和基于词典(例如 KSTEM)两种。两种取词根技术都不完美,有时会将不该匹配上的词匹配上,例如使用 Porter Stemmer 后,"policy"和"police","organization"和"organ"都将具有相同的形态,而有时又没能把该匹配的词匹配上,例如 Porter Stemmer 不能将"European"和"Europe"变成相同的形态。这些问题可能会导致严重的检索错误。

Strzalkowski 和 Vauthey 在检索系统中使用了词典辅助的取词根方法,取词根结果中不合理的情况有所改进,检索精度也有 6 %到 8 %的提高[9]。Xu 和 Croft 提出基于语料库的取词根方法,使 Porter 和 KSTEM stemmer 对检索精度的提高都略有改进[10]。

实际上,尽管取词根技术的使用对信息检索效果只有较小的提高,但由于这种技术可用性很强,所以被广泛地使用在信息检索系统中。

2.1.4　词性标注

词性标注在信息检索中的用途并不明显,最大的问题在于即便词性标注已经有了很高的精度(文献[11]中提到的英文词性标注精度是 97 %),该怎么将它用在检索里仍在研究之中。

一种用法是只对某些词性的词进行索引。Kraaij 和 Pohlmann 研究了不同词性的词对检索的重要性[12]。他们的结论是文档中对检索有帮助的词中 58 %是名词,29 %是动词,13 %是形容词。而如果只关注那些排序最靠前的相关文档,则发现有帮助的词中 84 %都是名词。Arampatzis 等仅使用名词完成检索实验,结果比使用所有词有相对 4 %的提高[13]。当然这个提高的幅度很小,而且并不保证在别的环境中能够稳定。一种可能的方法是对不同词性的词赋以不同的权重,这种做法的可行性仍有待研究。

另一种用法是将不同词性的词分开,只让查询和文档中词性相同的词能够匹配上。然而事实上有时的确需要匹配不同词性的词。Voorhees 在对其实验结果进行分析时认为,具有相同词根的形容词、动词和名词没有匹配上是导致其实验失败的原因之一[15]。苏祺等使用 TREC7 和 TREC8 的数据集,考察了这一用法对 SMART 系统检索效果的影响[11]。实验结果表明: 对于具有同一词形不同词性的词,使用词性标注加以区分,有助于减少匹配的噪音,从而能提高检索精度;相同词义、相同词根、不同词性的词没有匹配上,导致检索召回率下降;词性标注对信息检索的影响同查询与文档中具体词汇分布密切相关;词性标注对某些查询的检索效果有所改进,但改进不明显,并受到索引项权重选择的影响。

2.2　高级自然语言处理技术的应用

高级自然语言处理技术包括句法分析、短语识别、命名实体识别、概念抽取、指代消解和词义消歧等。由于短语识别、命名实体识别、指代消解等技术都需要用到句法分析,而句法分析技术并不直接用于信息检索,因此不对句法分析进行专门的讨论。

2.2.1　短语识别

识别查询和文档中的短语可以借助于自然语言处理中的句法分析技术,也可以采用统计的方法。短语识别技术在信息检索中使用的效果好坏不一,很大程度上取决于具体的识别技术、使用的短语类型以及使用的匹配策略[14,16~18]。近年来短语识别技术的使用有了一些新的进展。

Nie 和 Dufort 将短语作为附加的单元结合到

传统的基于词的索引中。他们将短语和词放在不同的向量中,分别计算出查询和文档的相似度后再进行加权[19]。在 TREC6 和 TREC7 数据集上的实验结果表明,这种短语使用方法大幅度提高了检索精度。

Liu 等考虑了查询和文档中存在的 4 种短语:专有名词、词典短语、简单短语和复杂短语[20]。他们采用了灵活的短语识别技术,只要组成短语的所有词在一定大小的窗口内出现就识别出这个短语。每一种短语对应的窗口大小都不同,通过决策树学习得到。计算组成短语的所有词之间的亲密度,只有亲密度大于一定阈值的短语才被选用。在 TREC9、TREC10 和 TREC12 数据上的实验结果表明,短语识别为检索精度带来了 3%到 16%的提高。

2.2.2 命名实体识别

命名实体是一种标识了某个概念或实体的特殊短语,例如专有名词、人名、地名、机构名等。显然,命名实体比词和一般短语表达了更加精确的信息。但在信息检索中使用命名实体并没有为效果提高带来多少帮助[21]。一方面因为命名实体识别技术自身存在的错误,另一方面,研究者们也困惑于如何对命名实体进行部分匹配,例如"Bill Clinton"和"Clinton",应该赋以怎样的权重呢?这一点类似于分词中分词精度很高时遇到的问题,因此可以尝试用近似的方法加以解决。

2.2.3 概念抽取

概念是比命名实体更为一般的一种特殊短语。命名实体标识了某种概念,因此可以认为都属于概念。但概念还包括了更多不属于命名实体的短语,例如"information retrieval"。然而概念抽取也没能提高信息检索效果[21]。研究者们提出了许多疑问:真的有必要在信息检索中使用概念么?该如何使用概念?如何对表达了同一含义的不同概念进行规范化?例如"95%","95 percent"和"0.95"。

2.2.4 指代消解

指代消解技术为文档中出现的代词或指代不明的短语找到它们实际所指代的事物。例如用来指代"Bill Clinton"的"Mr. President","He denied all responsibility"中的"he",都可以使用指代消解技术给出相应的具体解释。这个技术能够消除文档中不明确的表达方式,看上去应该可以对信息检索有所贡献,然而事实并非如此。指代消解对信息检索效果提高也没有帮助。一方面仍然因为指代消解的结果自身还有较多的错误,另一方面因为代词和指代不明的短语实际上并不怎么影响信息检索的结果[21]。

2.2.5 词义消歧

词义消歧是研究者们不断尝试着应用到信息检索中的一种自然语言处理技术,针对自然语言中存在的"同一个词可以表达多种意思"的问题,为每个词找到其在具体语境中实际表达的含义。

(1) 词义索引

Voorhees 在尝试将词义作为索引项时使用了词义消歧技术[15]。因为语言中广泛存在的一词多义和同义词现象,直接使用词义建立索引是很直接的一种想法。Voohees 使用词义消歧技术为文档中的词确定其在 WordNet 中的含义,而在 WordNet 中,词的每一种含义都对应着一个同义词集合(Synset),这样就可以用词对应的 Synset 来表示它的含义并建立索引。

Voohees 的实验结果表明,使用词义索引的检索效果不如直接使用 Stemming 后的词建立的索引,有时甚至差得多,下降的幅度为 6%～40%。对检索结果进行逐个查询的分析后发现,有一些查询的确能够受益于词义索引,然而更多的查询效果都下降了。几乎所有下降的情况都是因为查询和文档间本该匹配上的词由于使用了词义索引后没能匹配上。原因主要有三个:

第一,查询和文档中的同一个词,表达的意思也相同,但使用词义消歧技术后却选择了两种不同的含义;

第二,有些查询中的词由于缺少上下文而无法进行词义消歧,这通常是因为查询太短;

第三,由于实验中只对名词进行词义消歧,形容词和动词等在索引中则保留了原有的形态,导致了一些在 Stemming 后可以匹配上的形容词、动词和名词无法匹配。

可以发现,第一点是因为词义消歧发生了错误。那么,词义消歧需要到多高的精度才能够对信息检索有帮助呢?Sanderson 给出的答案是 90%[22]。他在实验中发现,如果词义消歧存在 20%到 30%的错误率,那么成功消歧所带来的检索效果改进都会被消歧失败导致的负面影响而抵消。

(2) 歧义对信息检索的影响:消歧是否有必要

词的歧义的存在对信息检索到底有多大影响呢?考虑这样一个查询"fly",它是一个带有歧义的查询,包含许多不同的含义:苍蝇?乘坐飞机?飞快地跑?还是在说拉链前的那块布?再看看这些查

询:“fly buzz”,“fly airplane”,“fly pants”。这样不就清楚了吗?事实上,只要在“fly”这个查询里再加入一个词,原本的歧义就自动消除了,而如果只有一个词,自然语言处理技术也不可能对它进行词义消歧。对于文档就更容易了。文档中有很多词,上下文就可以保证词是无歧义的。因此,只要查询的长度不是太短(例如只有一个词),词义消歧技术在信息检索中就没有用武之地。

(3) 在信息检索中使用词义消歧的新进展

Sanderson 的结论、对词义消歧是否有意义的怀疑并没有使研究者们停止探索词义消歧在信息检索中的可能应用,近年来有了一些新的成果。

Stokoe 等利用和 WordNet 一起发布的 Semcor 语料对词义消歧系统进行训练,并在消歧中结合了词性、共现关系等信息,对于消歧失败的词则直接赋以它在 WordNet 中出现频率最高的词义[23]。尽管他们的词义消歧精度只有 62 %,但在 TREC9 数据上完成的检索实验表明消歧为检索效果带来了相对 45 %的提高。然而,他们实验中所用的基准系统性能很差,所以即便检索精度有了很大提高,但仍比 TREC9 上的最好结果差不少。

Kim 等使用了一种特别的词义消歧技术。他们只考虑 WordNet 中最原始的 25 种词义(Root sense),对每个词赋以其中的一种,这样可以确保消歧的精度[24]。尽管他们没有给出消歧的准确率,但在 TREC7 和 TREC8 数据上的实验结果表明这种消歧方法对检索效果有 10 %以上的提高。他们在 BM25 公式中加入词义信息的尝试也获得了成功。

(4) 用户查询的消歧

Allan 和 Raghavan 提出一种消除单个词查询中的歧义的方法[25]。他们统计出查询词邻域内频繁出现的词性模版,每个模版都对应着人工构造的一个问题,在实际系统中可以将这些问题提供给用户选择,让用户明确地指定一个查询目的,从而消除查询中的歧义。实验中他们定义了一个清晰度公式,比较消歧前后查询清晰度的变化。结果表明,来自 TREC 的查询在消歧后清晰度提高了 41 %,来自 Web 检索的查询清晰度提高了 25 %。

Liu 等则从相反的角度来消除查询中的歧义,他们处理较长的多词查询,利用查询内的上下文信息为查询中的每个词找到精确的词义[26]。消歧的大体步骤是: 首先识别出查询中的短语,并利用 WordNet 中的信息为短语中的词消歧,然后仍利用 WordNet 为查询中的其他词消歧。对于没有成功消歧的词,如果它在 WordNet 中出现频率最高的词义大于其他所有词义的频率和,则赋以它这个词义,否则借助一次 Web 检索来帮助消歧。消歧实验在 TREC13 Robust 评测的 250 个查询上进行,结果表明,Liu 的消歧方法可以为查询中的 333 个歧义词全部成功消歧,精度为 90 %,并在 5 个 TREC 数据集上将检索实验结果提高了 10 %到 25 %。

2.3　在语言模型中加入自然语言处理

近年来研究者们尝试将自然语言处理加入到语言模型(Language Model)中,取得了不错的进展。语言模型本质上是通过计算从文档中生成查询的概率来判断文档和查询的相关程度。

Allan 和 Kumaran 在语言模型中加入取词根技术(Stemming),认为 Stemming 可以看作是一种平滑(Smoothing)[27]。文档 d 生成查询词 w 的概率,用 d 生成 w 的最大似然估计和 d 生成与 w 同词根的那些词的最大似然估计进行加权。假设 w 及其同词根的词组成集合 c,他们进一步提出一个生成模型(Generative Model),认为 d 中生成 w 可以分两步,首先由 d 生成 c,然后再由 c 生成 w。这个假设来源于对作家写作的模拟,作家选词的时候,往往先想到要表达某个含义,然后再根据语法句法等规则选用正确形式的那个词。实验结果表明新模型的检索平均精度有 10 %左右的提高。

Cao 等利用 WordNet 和共现关系获得文档中词的关系,并结合到语言模型中[28]。他们认为查询和文档匹配有两种方式。一种是直接匹配,即文档和查询具有相同的词;另一种是通过一些词的关系而匹配,即文档中的词通过 WordNet 或者共现关系和查询中的词发生联系。假设文档在生成查询的时候分两步,首先随机获得文档中的一个词,然后根据这个词和查询中的词的关系作进一步计算。他们将文档中的词和查询中的词通过 WordNet 发生匹配的模型称为链接模型,并根据 WordNet 中的同义(Synonym),上义(Hypernym)和下义(Hyponym)关系将链接模型进一步细化成 3 个模型。在 3 个数据集上的检索实验结果表明,加入了 WordNet 和共现关系的语言模型比一元语法(Unigram)语言模型在精度和召回率上都有明显的提高。他们还发现,在模型中加入 WordNet 所获得的效果改进是稳定的。

Gao 等则在语言模型中加入了语言学中的概念[29]。这里的概念指的是语义块(例如命名实体)

和句法块(例如名词短语,动词短语)等。他们考虑文档在生成查询时,首先生成文档中的某个概念,再根据这个概念生成查询。这样,文档中的每个概念都代表了一个概念模型。文档与查询的相关度是通过所有的概念模型和一元语法语言模型、二元语法语言模型一起决定的。在6个数据集(包括中文和英文两种)上的检索实验结果表明,他们的模型在5个查询集合的检索效果比概率模型(BM25)和别的语言模型有较为显著的改进。分析效果变差的那个集合,他们认为原因是自然语言处理工具所使用的训练语料和实际的测试语料存在着语言差异,导致概念不能正确地抽取,模型也不能正常地工作。Gao等的实验结果表明,语言学特征(命名实体、短语等)只要被恰当地使用,就能够为检索效果的提高提供有效帮助。

2.4 自然语言处理技术应用在信息检索中的效果

TREC5 NLP评测结果表明,查询扩展、短语识别、专有名词和取词根等自然语言处理技术应用在信息检索中,都比仅仅基于词的检索系统提高了效果,但仍无法超过基于统计的检索系统[30]。

Manning分析了自然语言处理技术在Web检索中的作用[31]。他认为通过句法分析得到的短语会有帮助,但利用统计方法得到的短语可以达到几乎同样的效果,并且已经被应用到系统中了。事实上,通过Web链接分析技术和锚文字(Anchor Text)的使用,Web检索已经获得了更大的进展。锚文字实际上是人工提供的同义词信息,而Web站点构建者所建立的Web链接以及Web用户的点击行为则将他们认为正确或者重要的信息提供给了搜索引擎。显然,人的智能总要比人工智能更胜一筹。

2.5 自然语言处理资源在信息检索中的应用

自然语言处理资源指的是类似WordNet和HowNet的词典。

Smeaton在检索实验中尝试使用句法分析等自然语言处理技术失败后,利用WordNet进行了实验[2]。他使用WordNet中的同义词集合计算词之间的语义相似度和语义距离。人工评测判定,计算出的语义相似度有近80%的精度。然而使用TREC数据完成的检索实验结果仍然不好,经分析他认为实验中用到的词义消歧工具精度不高以及TREC查询中出现了太多专有名词是导致失败的原因。

WordNet还经常被用于查询扩展,在许多情况下都能对检索效果有帮助[20,32]。Zhang等使用WordNet作查询扩展也取得了检索效果的提高,但提高的幅度不如基于统计方法完成的查询扩展[33]。赵军等的实验也有类似的结论[34]。

此外,如上文所述,WordNet是词义消歧中常用的工具。

自然语言处理资源是人工构造出来,或者是机器生成后经人工修订过的,准确性很高,很适合用在信息检索中,但需要针对不同的问题,精细和灵活地进行使用。

3 自然语言处理在信息检索中的应用启示

在信息检索中使用自然语言处理的尝试大多没有获得好的效果,即便有一些帮助也是很小的,远远没有达到令人满意的程度。研究者们对此进行了分析,并从中得出结论,认为自然语言处理需要针对信息检索任务进行优化。

3.1 为什么自然语言处理在信息检索中使用的效果不好

Lewis和Sparck-Jones认为基于统计方法的信息检索比使用自然语言处理取得了更好的效果,是因为统计方法"摘下了最低枝头上的果实"[35](即把容易的事情都做了),而剩下的都是难得多的问题,自然语言处理不能很好地解决是可以理解的。

自然语言处理的领域相关性很大,移植性差,因而使用在信息检索中消极影响大于积极影响。信息检索的效果取决于查询的性质,而实际的查询对于自然语言处理和信息检索而言都很有难度[36]。用户的查询大多不够清楚、专业和完整,而这一点是很难避免的,因为用户在检索之前对自己的需求可能都不太清楚[35],想用查询表达出来有时是很困难的,更别说清楚地描述了。

Strzalkowski等认为在信息检索中使用自然语言处理存在着以下障碍:自然语言处理的鲁棒性和效率还有待提高;自然语言处理的结果表示过于复杂;缺少一个能够很好地使用自然语言处理的信息检索模型[37]。

Voohees认为,自然语言处理技术必须没有错误,近乎完美才能够对信息检索效果提高有帮助[36],而目前显然还能达到。她还认为,没有使用自然语言处理的信息检索技术能够取得好的效果,

是因为这些技术中其实已经蕴含了语言学知识[36]。

以上文提到的词义消歧为例。实际中一词多义并不是导致信息检索失败的主要原因,除非查询非常短(只有一个词)。如果文档和查询之间有足够多的词匹配上了,相似度很高,那么两者中同一个词的上下文往往是类似的,歧义的词一般也表达了同一个含义。这样只要检索用户对召回率要求不高,只返回相似度很高的文档就基本可以保证这些文档是相关的,它们与查询中共有的词通常具有相同的含义。这样就自动实现了词义消歧。

3.2　自然语言处理需要针对信息检索任务进行优化

简单的自然语言处理往往能够对信息检索效果的提高有所帮助,尽管帮助不大,例如上文提到的去除停止词、取词根等。复杂的自然语言处理则基本上起不到什么积极作用,甚至对检索结果有害,例如上文提到的词义消歧、指代消解等。而复杂的方法往往大大增加了处理和存储消耗。因此自然语言处理需要针对信息检索任务进行优化,使其能够在降低复杂性的同时提高信息检索的效果[2,14]。

那些直接为提高信息检索效果而设计的自然语言处理技术往往行之有效,而独立于检索任务,纯粹从语言学角度出发的技术则通常不能够成功。例如 Porter Stemmer 是一种高效的取词根算法,专门用于信息检索,尽管它处理的结果中有一些语言学上的错误,但却为信息检索提供了帮助;统计方法得到的许多"短语"被认为违反了语言学知识,但这种方法是针对信息检索任务进行了优化的,同样获得了成功;词义消歧最初并不是为检索而设计的,它对检索没有帮助作用,甚至是有害的,然而如果针对某个领域进行优化,例如在医学文档中使用的 MeSH,则提高了检索效果[14]。

研究者们看到的结论是,有着大量计算消耗的自然语言处理对提高信息检索效果的帮助很小,而非自然语言处理的技术,例如统计方法,则带来了更大的帮助。然而看上去很小的积极影响往往是积极效果和消极效果叠加后的产物。因此,如果能够将自然语言处理带来的积极效果和消极效果自动分开,将对信息检索任务非常有效。这样的操作将需要综合考虑自然语言处理和信息检索,将两者有机结合起来,而不是简单地构建一个自然语言处理系统,然后将它像个黑盒一样应用到信息检索中。

4　自然语言处理在信息检索中的应用展望

研究者们对未来自然语言处理在信息检索中的应用提出了展望,不仅包括狭义信息检索中检索效果的提高和检索结果显示的智能化,还包括对广义信息检索中问答系统、信息抽取等任务的帮助。

4.1　自然语言处理在狭义信息检索中的应用展望

所谓狭义信息检索,即本文讨论的文档检索。一直以来,自然语言处理对信息检索效果的提高帮助很有限,但未来可以着眼于信息检索的其他方面。而最新的一些进展也表明,自然语言处理与信息检索的有效融合能够改进检索效果。

4.1.1　自然语言处理用于信息检索系统结果显示

用户在文本框中输入一个查询,信息检索系统返回一个排序文档列表,这就是当前的信息检索界面。自然语言处理可以用来让系统返回的结果显得更加智能和人性化[38,39]:为检索结果中的文档作自动文摘并呈现给用户;将排在前面的若干篇文档作多文档自动摘要,或者可以由用户选择用哪几篇文档作摘要;将检索结果作聚类而不是把长长的文档列表返回给用户。

4.1.2　自然语言处理用于检测用户查询的上下文

Baeza-Yates 认为在信息检索中检测出用户查询的上下文信息对于提高检索效果是很关键的,而自然语言处理可以在这方面起到作用[1]。上下文信息为消除查询中的歧义提供了线索。这里所说的歧义指的是,不同的用户在使用同一个查询时往往有着不同的信息需求,想要的结果也不尽相同。假如能够确切地知道用户的详细个人信息和查询目的,也就是查询的上下文信息,检索系统就可以将适合该用户的最佳结果返回。上下文信息往往从用户的个人网页,用户浏览过的网页,点击行为,用户日志,IP 地址等资源中获得。自然语言处理可以被用来在这些资源中抽取出正确的上下文信息。

4.1.3　自然语言处理与信息检索统一模型

自然语言处理应该以恰当的方式与信息检索相结合,以取得检索效果的改进。这需要对自然语言处理和信息检索综合考虑,将它们有机结合起来建立一个统一模型[14,37]。

Zhou 和 Zhang 提出的 NLPIR 理论框架可以

说是这一方向的一点尝试[40]。NL PIR 框架的基本假设是查询和文档间存在着表示上的距离,如果能减小这个距离就可以获得更好的检索效果。达到这一目的的方法是在框架中结合不同层次的自然语言处理方法,包括直接方法、扩展方法、抽取方法、转换方法和统一方法。

直接方法包括去除停止词、取词根、分词、词性标注等;扩展方法包括使用词典进行查询扩展等;抽取方法包括命名实体、事实等的抽取;转换方法包括句法分析、指代消解、语义分析等;统一方法相对而言则是处于设想中的,实际可用的技术很少。

如图 1 所示,在 NL PIR 框架中,通过不同层次的自然语言处理方法,查询和文档间的距离越来越小,这样在两者间进行匹配就越来越容易。然而,这个框架只是一个初步的想法,距离真正实用的模型还差得很远。

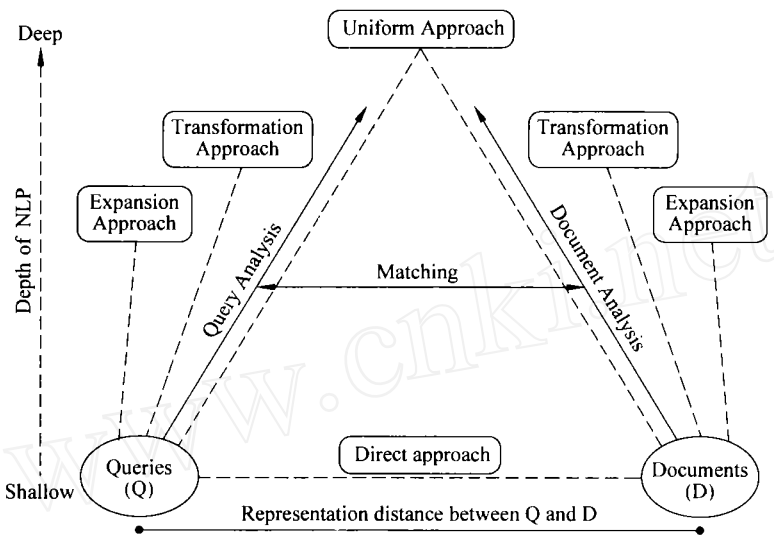

图 1 NL PIR 框架(引自文献[40])

4.2 自然语言处理在广义信息检索中的应用及展望

本文中提到的信息检索是狭义的,指的是文档检索。广义的信息检索除了文档检索外,还包括段落检索、问答系统和信息抽取等。长期以来,自然语言处理的发展是为了能够应用到像机器翻译这样需要精确结果的任务中[2],因此在问答系统、自动文摘和信息抽取等需要更精细的自然语言理解的任务中作用可能更大些[31,36]。事实上,在这些任务中,自然语言处理和信息检索间的交互作用已经取得了很好的成果[1,41,42],TREC 评测[43]的结果也表明自然语言处理可以有效地提高这些任务的效果。

如图 2 所示,按照查询长度和结果长度对这些任务进行划分[14]。

可以看到,查询长度较长的任务例如问答系统和信息抽取等更适合使用自然语言处理。同时可以想象的是,结果长度较短的任务需要对结果进行句法和语义的处理以保证其精确性,从而更需要使用自然语言处理。因此,和文档检索相比,问答系统和信息抽取中更适合使用自然语言处理。

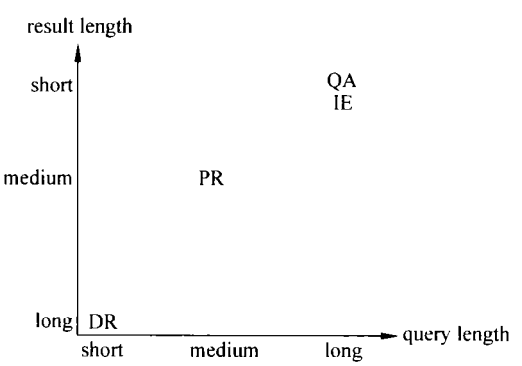

图 2 按照查询长度和结果长度对文档检索(DR)、段落检索(PR)、问答系统(QA)和信息抽取(IE)进行划分(引自文献[14])

4.3 未来自然语言处理在信息检索中的可能应用方向

迄今为止,自然语言处理在信息检索中发挥的作用不尽人意,但并不能因此而停止尝试和研究。随着自然语言处理技术自身的不断进步,信息检索模型的进一步完善,信息抽取、数据挖掘、语义网络(Semantic Web)等技术的不断进展,自然语言处理

与信息检索必将形成紧密的结合,达到相得益彰的效果。在此尝试列举一些未来自然语言处理在信息检索中可能的应用方向:

· 在信息检索模型中加入自然语言处理,建立有机的统一模型。正如 2.3 所述,在语言模型中引入自然语言处理已经获得了初步的成功。相信未来的信息检索模型中,自然语言处理将是不可分割的一部分。

· 自然语言处理用于语义网络的构建。自然语言处理用于对 Web 页面进行处理,形成语义理解,找出页面中的实体(Ontology),获得实体间的关系,从而构建语义网络。基于语义网络的信息检索将为用户提供更为精确的信息。

· 自然语言处理用于问答系统、信息抽取、数据挖掘等信息检索相关的领域。目前这些领域中已经大量使用了自然语言处理技术,相信未来会有更好的应用。

5　结论

长期以来,研究者们不断尝试着在信息检索中使用自然语言处理,然而结果并不能令人满意。复杂性较低的基本自然语言处理技术,包括去除停止词、分词、取词根等,计算消耗小,简单易行,对信息检索的帮助很小,但一些能够提高检索效率的技术例如去除停止词和取词根等,仍然是在信息检索实验平台中推荐使用的;复杂性高的高级自然语言处理技术,包括句法分析、短语识别、命名实体识别、概念抽取、指代消解和词义消歧等,计算消耗大,精度不高,对信息检索基本没有帮助,甚至可能有害。

研究者们对这一现象进行分析,认为自然语言处理在信息检索中不起作用的原因是: 一方面,自然语言处理技术的精度不够高,存在错误,即便有一些积极影响也会被消极影响所掩盖;另一方面,不使用自然语言处理的方法,例如统计方法,能够大幅度地提高检索效果的原因是它其实已经蕴含了语言学的知识,并且它解决的问题是相对容易的那些,剩下留给自然语言处理的都是难得多的。

自然语言处理在除狭义信息检索(文档检索)之外,例如问答系统、信息抽取中已经发挥了很大的作用。这些将是未来自然语言处理应用的发展方向之一。研究者们还建议将自然语言处理针对信息检索任务进行优化,例如用于信息检索结果的智能显示,用于获取用户查询的上下文信息从而把最佳结果返回给用户。

此外,自然语言处理与信息检索有效融合而成的统一模型,也将是研究者们关注的重点。在语言模型中加入自然语言处理获得的成功可以看作是这一方向取得的成果。

本文的最后,回顾一下 1995 年 Smeaton 在欧洲信息检索暑期学校时说的话[5]: "采用一些比仅仅数词的个数(指的是统计方法)更聪明的方法应当可以提高检索效果 ……因此我们尝试着将自然语言处理应用到信息检索中,但这并不容易。"

参考文献:

[1] Ricardo Baeza-Yates. Challenges in the Interaction of Information Retrieval and Natural Language Processing [A]. In: Proceedings of 5th International Conference on Intelligent Text Processing and Computational Linguistics [C], CICLing 2004, Seoul, Korea, February 15-21, 2004. 445-456.

[2] Alan F. Smeaton. Using NLP or NLP Resources for Information Retrieval Tasks [A]. In: Natural Language Information Retrieval [M]. T. Strzalkowski, editor, Kluwer, 1997. 99-111.

[3] http://wordnet.princeton.edu/.

[4] http://www.keenage.com/.

[5] Alan F. Smeaton. Natural Language Processing & Information Retrieval, a lecture presented at the European Summer School in Information Retrieval [Z]. Glasgow, 1995.

[6] Fuchun Peng, Xiangji Huang, Dale Schuurmans and Nick Cercone. Investigating the Relationship between Word Segmentation Performance and Retrieval Performance in Chinese IR [A]. In: Proceedings of 19th International Conference on Computational Linguistics [C], 2002. 72-78.

[7] 金澎,刘毅,王树梅. 汉语分词对中文搜索引擎检索性能的影响 [J]. 情报学报,2006,25(1):21-24.

[8] Schubert Foo, Hui Li. Chinese word segmentation and its effect on information retrieval [J]. Information Processing and Management, 2004, 40(1):161-191.

[9] Tomek Strzalkowski and Barbara Vauthey. Information retrieval using robust natural language processing [A]. In: Proceedings of the 30th annual meeting on Association for Computational Linguistics [C], 1992. 104-111.

[10] J Xu and W. B. Croft, Corpus-based stemming using cooccurrence of word variants [J]. ACM Transac-

tions on Information Systems (TOIS), 1998, 16(1): 61-81.

[11] 苏祺,咎红英,胡景贺,项锟. 词性标注对信息检索系统性能的影响 [J]. 中文信息学报, 2005, 19(2): 58-65.

[12] W. Kraaij and R. Pohlmann. Viewing stemming as recall enhancement [A]. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. ACM Press, 1996. 40-48.

[13] A. T. Arampatzis, Th. P. van der Weide, C. H. A. Koster and P. van Bommel, Text Filtering using Linguistically-motivated Indexing Terms [R]. Technical Report CSI-R9901, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands, 1999.

[14] Thorsten Brants. Natural Language Processing in Information Retrieval [A]. In: Proceedings of 20th International Conference on Computational Linguistics [C]. Antwerp, Belgium, 2004. 1-13.

[15] Ellen M. Voorhees. Using WordNet to disambiguate word senses for text retrieval [A]. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. ACM Press, 1993. 171-180.

[16] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases [A]. In: Proceedings of the RIAO97 [C]. 1997. 200-216.

[17] Tomek Strzalkowski. Natural language information retrieval [J]. Information Processing & Management, 1995. 31 (3) :397-417.

[18] S. E. Robertson and S. Walker. Okapi/ Keenbow at TREC-8 [A]. In: Proceedings of the 8th Text Retrieval Conference [C]. NIST Special Publications 500-246, Gaithersburg, 1999. 151-162.

[19] Jian-Yun Nie and Jean-Francois Dufort. Combining words and compound terms for monolingual and cross-language information retrieval [A]. In: Proceedings of Information [C]. Beijing: 2002. 453-458.

[20] Shuang Liu, Fang Liu, Clement Yu and Weiyi Meng. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases [A]. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. ACM Press, 2004. 266-272.

[21] James Allan. Natural Language Processing for Information Retrieval, tutorial presented at the NAACL/ ANLP language technology joint conference in Seattle [Z]. Washington, April 29, 2000.

[22] M. Sanderson. Word Sense Disambiguation and Information Retrieval [A]. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. ACM Press, 1994. 49-57.

[23] Christopher Stokoe, Michael P. Oakes and John Tait. Word Sense Disambiguation in Information Retrieval Revisited [A]. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press [C]. 2003. 159-166.

[24] Sang-Bum Kim, Hee-Cheol Seo and Hae-Chang Rim. Information Retrieval using Word Senses: Root Sense Tagging Approach [A]. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. ACM Press, 2004. 258-265.

[25] James Allan and Hema Raghavan. Using Part-of-speech Patterns to Reduce Query Ambiguity [A]. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. ACM Press, 2002. 307-314.

[26] Shuang Liu, Clement Yu and Weiyi Meng. Word Sense Disambiguation in Queries [A]. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management [C]. ACM Press, 2005. 525-532.

[27] James Allan and Giridhar Kumaran. Stemming in the Language Modeling Framework [A]. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (poster) [C]. ACM Press, 2003. 455-456.

[28] Guihong Cao, Jian-Yun Nie and Jing Bai. Integrating Word Relationships into Language Models [A]. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. ACM Press, 2005. 298-305.

[29] Jianfeng Gao, Haoliang Qi, Xinsong Xia and Jian-Yun Nie. Linear Discriminant Model for Information Retrieval [A]. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. ACM Press, 2005. 290-297.

[30] Tomek Strzalkowski and Karen Sparck Jones. NLP Track at TREC-5 [A]. In: Proceedings of the 5th Text Retrieval Conference [C]. NIST Special Publications 500-238, Gaithersburg, 1996. 97-100.

[31] Christopher Manning. Opportunities in Natural Language Processing [Z]. presentation given at Oracle, 2002.

[32] Zhiguo Gong, Chan Wa Cheang and Leong Hou U. Web Query Expansion by WordNet [A]. In: Proceedings of 16th International Conference of Database and Expert Systems Applications [C]. Copenhagen, Denmark, August 22-26, LNCS 3588, 2005. 166-175.

[33] Min Zhang, Ruihua Song, Chuan Lin, Shaoping Ma, et al. Expansion-Based Technologies in Finding Relevant and New Information: THU TREC2002 Novelty Track Experiments [A]. In: Proceedings of the 11th Text Retrieval Conference [C]. NIST Special Publication, Gait hersburg, MD, USA: 2002. 591-595.

[34] 赵军,金千里,徐波. 面向文本检索的语义计算 [J]. 计算机学报,2005,28(12): 2068-2078.

[35] David D. Lewis and Karen Sparck-Jones. Natural Language Processing for Information Retrieval [J]. Communications of the ACM, 1996, 39(1): 92-101.

[36] Ellen M. Voorhees. Natural Language Processing and Information Retrieval [A]. Information Extraction: Towards Scalable, Adaptable Systems [M]. LNCS 1714, 1999. 32-48.

[37] Tomek Strzalkowski, Fang Lin, Jin Wang and Jose Perez-Carballo. Evaluating Natural Language Processing Techniques in Information Retrieval: A TREC perspective [A]. In: Strzalkowski, Tomek (Ed). Natural Language Information Retrieval [M]. Kluwer, 1999.

[38] Alan F. Smeaton. Information Retrieval: Still Butting Heads with Natural Language Processing [A]. In: Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology [M]. Frascati, Italy, July 1997. 115-138.

[39] Karen Sparck Jones. What is the role of NLP in text retrieval [A]. In: Natural Language Information Retrieval [M]. T. Strzalkowski, editor, Kluwer, 1999.

[40] Lina Zhou and Dongsong Zhang. NLPIR: A Theoretical Framework for Applying Natural Language Processing to Information Retrieval [J]. Journal of the American Society for Information Science and Technology, 2003, 54(2): 115-123.

[41] Hui Yang and Tat-Seng Chua. QUALIFIER: Question Answering by Lexical Fabric and External Resources [A]. In: the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) [C]. 2003. 363-370.

[42] Margaret Connell, Ao Feng, Giridhar Kumaran, Hema Raghavan, Chirag Shah and James Allan. UMass at TDT 2004 [A]. TDT2004 Workshop [C]. 2004.

[43] http://trec.nist.gov/.