第32卷 第18期 计 算 机 工 程 2006年9月
Vol.32 № 18 Computer Engineering September 2006
·博士论文· 文章编号：1000—3428(2006)18—0009—03 文献标识码：A 中图分类号：TP391

基于语言概念空间的跨语种信息检索模型

吴 晨[1,2]，张 全[2]，缪建明[1,2]
(1. 中国科学院研究生院，北京 100039；2. 中国科学院声学研究所，北京 100080)

摘 要：提出了一种基于语言概念空间的跨语种信息检索模型，该模型以建立在语言概念空间中的形式化语境单元框架表示处理所需的中间信息，通过用以描述语境单元框架的语义符号间的匹配和生成机制来实现文本的跨语种检索，有助于避开用形式多样的具体语言作为处理中介存在的模糊消解问题。实验证明，这一模型显著改善了检索系统的性能。

关键词：跨语种信息检索；语言概念空间；语境单元框架

Model of Cross-language Information Retrieval Based on Language Concept Space

WU Chen[1,2], ZHANG Quan[2], MIAO Jianming[1,2]
(1. Graduate School of Chinese Academy of Sciences, Beijing 100039; 2. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080)

【Abstract】 This paper presents a model of cross-language information retrieval(CLIR) based on language concept space(LCS), the model takes formalized context framework which is formed in language concept space as mediated interface, the framework is three-coordinated and composed of domain, situation and background. The model can avoid the disadvantage of the word sense ambiguity which occurs in the traditional CLIR. The experiments indicate that the LCS based CLIR system has good performance.

【Key words】 Cross-language information retrieval; Language concept space; Sentences group unit framework

跨语种信息检索(CLIR)技术作为一项独立的技术进行研究已经有10多年的时间，在这期间，研究人员取得了相当大的研究成果。但同时，他们也发现，目前所采用的方法存在一些自身很难克服的弊端，Jian-Yun Nie[1]和Mayfield、McNamee[2]指出：当前CLIR技术的缺陷源于近似的解决问题的方法，比如，翻译和检索分离、从单语种检索的结果到多语种检索结果的界定存在问题等，Chen和Gey[3]提出了一种基于适量反馈和分解的跨语种信息检索机制。但是这种方法只能看作是对原有算法存在缺陷的弥补，而很难去解决它。

作者认为，要很好地解决跨语种信息检索现存的问题，必须把理解深入到语义层面，而在概念空间形成的语义是最理想的选择，它可以最大程度地减少语言歧义。本文设计的模型就是基于概念空间的跨语种检索模型，该模型以语境单元框架作为语义形式化表示的基础，试图从语义层面，通过概念匹配解决模型中的关键性问题。

1 语境单元

语境单元是文本的形式化表示单位，用来描述语言空间中的句群信息，它由中科院声学所黄曾阳提出[4]。语境单元由3个要素构成：即领域、情景和事件背景。

1.1 领域

领域是指该句群陈述内容应归属的概念领域，描述事件的类型，或者说性质和归类，比如思维活动、信仰活动等，依据是HNC[4]的两类复合基元的定义，同时在两类理论复合基元的基础上进行了浓缩和扩展，领域类型共分为10大类[5]，这10大类领域具有统一的编号，如(1)心理活动及精神状态，编号：71和72；(2)专业及追求活动(第2类劳动)，编号：a和b，这些类可扩展细分，也可组合。

1.2 情景

情景是指事件的动态表示，是对领域的具体说明，包含了句群的深层语义结构表示式(领域句类表示式)，句群内各语义要素的表示结构(单元框架)和将各表示式结合在一起的数据结构(情景框架)。

1.3 背景

背景是指事件发生的条件环境以及作者的立场背景，包括基本背景信息以及陈述中心的时间空间信息、作者的参照点、目的等等。它由语句中的辅语义块提供，包括方式、工具、途径、条件、参照、起因和目的。

2 基于概念空间的跨语种信息检索模型

模型处理的基本思路是根据句子语义分析获取后得出句子的深层语义结构，通过语境单元生成阶段的处理获取以句群为单位的语境单元框架。所生成的语境单元框架根据处理阶段的不同被赋予两个名称：(1)索引语境单元框架，针对被检索的文档数据而言，所有被检索的文档数据经系统处理后得出的语境单元框架被组织成一个索引数据库供系统文本语义结构比较时用；(2)短时语境单元框架，针对用户输入的检索请求所提取的语境单元。这两项语义表示结构在文本语义结构比较模块中进行语义关系分析。最后，根据分析的结果

基金项目：国家“973”计划基金资助项目“自然语言理解的交互引擎研究”(2004CB318104)；中科院声学所知识创新工程基金资助项目“HNC语言知识处理理论及技术”

作者简介：吴 晨(1979—)，男，博士生，主研方向：自然语言理解；张 全，研究员、博导；缪建明，博士生

收稿日期：2005-10-08 **E-mail：**wuchen@mail.ioa.ac.cn

生成最终返回给用户的检索结果。模型处理框图如图 1 所示。

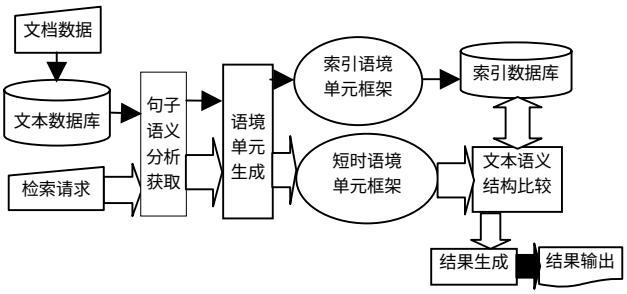

图 1 模型处理

2.1 句子语义分析获取

句子语义分析获取的任务是通过其来抽取句子的深层语义结构，它采用HNC句类分析技术[5]，句子语义分析获取输入是自然语言的句子，输出为该句子的语义表达符号，它是HNC定义的 57×56 个语义符号表达式(以下称句类)之一，以及各语义表示单元(以下称语义块)的符号表示。

2.2 语境单元生成

语境单元生成以句群为单位，句群由句子组成，各句子的句类混合生成语境单元框架结构，语境单元生成由以下几个步骤完成[5]：

(1)确定句群语义框架中的领域信息。领域信息由句子语义分析获取结果中各语义块的语义确定，语义块包括 Eg、El、C、B、A，分别表示主句的核心语义部分、子句的核心语义部分、各句子语义表示主体(以下称广义对象语义块)中的内容部分、B 和 A 表示各句子广义对象语义块的对象部分。从语义中优先抽取领域信息的顺序为 Eg > El > C > B > A。

(2)根据确定领域信息的语义块所属的句类，获取句群语义框架中的情景信息；也就是说情景信息由确定领域信息的句子的语义核心或者广义对象语义块确定，从包含该语义核心或者广义对象语义块的句子句类扩展而来。

(3)综合句类分析结果中的辅语义块和句中出现的综合类概念形成句群语义框架中的背景信息。

生成的语境单元将以框架的形式给出，下面给出语境单元的一种形式化表示方式，它由 3 个侧面组成，分别对应领域、情景和背景，每一个侧面表示都以“{}”括起，如下所示：

{DOM//[领域 1 : 出处 1(∨ 出处 *n*)] ∨ … ∨ [领域 *n* : 出处 1(∨ 出处 *n*)]}
{SIT// 要素名称 1[语义概念符号 1 ∨ … ∨ 语义概念符号 n] | …
| 要素名称 *n*[语义概念符号 1 ∨ … ∨ 语义概念符号 *n*]}
{BAC// 要素名称 1[要素内容] | … | 要素名称 n[要素内容]}

2.3 文本比较及结果生成

文本语义结构比较(简称文本比较)模块负责比较两个文本的语义特征结构，计算这两个结构之间的相似度，给出比较结果。结果生成模块负责生成文本检索的结果，这些结果根据可信度、精确度排序，无用的不一致文本将被剔除，根据用户的需求可以限定算法处理记录数目的空间，加快算法执行的速度。文本比较与结果生成两个步骤相互关联，该算法处理步骤如下：

(1)假设语境单元生成处理后生成的检索请求短时语境单元框架用c_r表示，将c_r与索引数据库中的各句群的语境单元框架c_x进行匹配，获取领域信息相同的句群，设句群总数为m，一次初选结果集为$C0$，取

$$X=\begin{cases}\text{int}(5/\log N) & N>1\\ 10 & N=1\end{cases}$$

其中，N 为预期要返回给用户的结果记录条数，N>0；int 表示取整。如果 Count(C0)>=X(Count(Cn)为集合 Cn 中的记录数)，则转(2)；否则，把 m 条句群放入一次初选结果集 C0，转(4)。

(2)假设记录 c 在近 n 个单位时间段内被用户选中的累积次数为 *dfn*，则它的优先选取因子

$$cp=\sum_{n\in[1,+\infty)}\frac{1}{2^n}cfn$$

取 $C1=\{c_m\mid c_m\in C0, cp_m\in Max_X(cp)\}$，其中，$Max_x(cp)$表示集合 $C0$ 中cp之最大的前X项。转(4)。

(3)假设句群c_n和句群c_m间的领域相似度用$rdom(c_n,c_m)$表示，取

$$C1'=\{c_m\in(C-C1)\mid rdom(c_r,c_m)>=rdom(c_r,c_l), c_l\in(C-C1), c_l\neq c_m\}$$

其中，c_r表示检索请求形成的短时语境框架。取 $C1=C1+C1'$，如果Count(C1)>=X，则取

$$C1=\{c_m\mid c_m\in C1, rdom(c_r,c_m)\in Max_X(rdom(c_r,c_m))\}$$

其中，$Max_x(rdom(c_r,c_m))$表示c_m与c_r产生的rdom(c_r,c_m)最大的前X项，转(4)；否则继续执行(3)。其中，$rdom(c_n,c_m)$为c_n和c_m所具有的领域字符串sc_n、sc_m之间的比较函数

$$rdom(c_r,c_l)=\frac{1}{\max(len(SCr),len(SCl))}\sum_{i=1}^{\max(len(SCr),len(SCl))}\frac{1}{2^{(|sc_{r[i]}-sc_{l[i]}|)}}$$

字符不存在时数值取 0。

(4)假设$gbk_m(c_n)$表示c_n中包含的情景信息所对应的第m个广义对象语义块语义符号，$objb_m(c_n)$表示c_n中第m个广义对象语义块内包含的对象语义符号，取

$$C2=\{c_m\mid c_m\in C1, (objb_1(c_m)\text{-}\!\!<\!\!>gbk_x(c_r)\text{或}\ gbk_1(c_m)\text{-}\!\!<\!\!>gbk_x(c_r)), x\in\{n\mid gbk_n(c_r)\}\}$$

其中，$\text{-}\!\!<\!\!>$ 表示语义符号等价，比较符号两边的字符串相似度，比较公式同$rdom(c_n,c_m)$，当所得数值＞λ时，$\text{-}\!\!<\!\!>$ 成立，否则不成立，即 $\succ\!\!\prec$。λ通常取经验值 0.7。

(5)检验 $C2$ 中各记录的第 2 个广义对象语义块，取

$$C2'=\{c_m\mid c_m\in C1, (objb_2(c_m)\text{-}\!\!<\!\!>gbk_x(c_r)\text{或}\ gbk_2(c_m)\text{-}\!\!<\!\!>gbk_x(c_r)), x\in\{n\mid gbk_n(c_r)\}\}$$

取 $C2=C2+C2'$。

(6)假设$bact(c_n)$表示c_n中包含的背景信息语义，取

$$C2''=\{c_m\mid c_m\in C2, bact(c_m)\succ\!\!\prec bact(c_r)\}$$

其中 $\succ\!\!\prec$ 表示语义符号相对。取 $C2=C2-C2''$。

(7)假设c_n和c_m间的语义核心相似度用rek(c_n,c_m)表示，对 C2 中的记录c_m，依照它与c_r间产生的rek(c_n,c_m)值，进行排序。如果Count(C2)>=N或者本次是从(8)返回，取

$$C2=\{c_m\mid c_m\in C2, \forall c_m\forall c_x(\text{rek}(c_m,c_r)>=\text{rek}(c_x,c_r)), m\in[0,\ldots,N-1], x\in[N,\ldots,Count(C2)]\}$$

转(10)；否则，转(8)。

(8)取

$$C2'''=\{c_m\mid c_m\in C1, (objb_x(c_m)\text{-}\!\!<\!\!>gbk_x(c_r)\text{或}\ gbk_x(c_m)\text{-}\!\!<\!\!>gbk_x(c_r)), x\in\{n\mid gbk_n(c_r)\}, x\neq 1,2\}$$

对 $C2'''$ 中的记录c_m，依它与c_r间产生的rek(c_r,c_m)值排序。

(9)取 $C2=C2+C2'''$

如果 Count(C2)>=N，取 $C2=\{c_m\mid c_m\in C2, m\in[0,\ldots,N-1]\}$。

(10)文本语义结构比较及结果生成结束，依次返回C2 中各c_m所指向的句群，作为最终结果。

3 算法示例

为了更好地说明模型处理流程，这里给出处理的一个示例。示例实现的是英—汉跨语种信息检索。示例中包括一条

查询请求(英语形式)和 3 条预先维护好的待检索语句(汉语形式)。查询请求用(1)标示，被检索语句用(2)、(3)、(4)标示。示例中，我们不描述处理细节，直接给出处理的各阶段结果，在每个关键词后面给出一定的概念解释，用“()”括起，概念解释包括该词语的语义符号表示和对该符号的语义解释。

(1)get the concept of nature language processing

根据算法“句子语义分析获取”的处理，可以得出该句子的句类知识，此句子的句类为!31XT19*21J。表示作用句和针对性接受句的混合。

该句类的句类代码表示为!31XT19*21J=A+XT19+TBC。

句类中的各语义块含义分别是：A：(省略)；

XT19：get(9219，v93808)；

TBC 包含两个方面：1)TBCB：nature language processing(fga68*030506，特指自然语言理解学科)，表示 TBC 语义块中所含的对象；2)TBCC：concept(r80)，表示 TBC 语义块中对象所指的内容。

通过“语境单元生成”的处理，可以得到该句群的“短时记忆要点框架”为：

领域：368，来源 TBCB，表示一类学科研究。

情景：领域句类： XT19*21J。领域句类与句子句类相同。

背景：无。

综上所述，形式化的“语境单元框架”可以表示为：

{DOM//368 ：TBC}

{SIT//XT19 [(9219, v93808)] | A[省略] | TBCB [fga68] | TBCC [r80]}

(BAC// [无])

(2)HNC 理论是“hierarchical network of concepts(概念层次网络)”的简称，是关于自然语言处理的一个理论体系。

根据算法“句子语义分析获取”的处理，可以得出此句子的句类为 jDJ+jDJ。表示这是一个迭句，由两个是否判断句堆迭而成，后面一个语句共享前面一个语句的第 1 个广义对象语义块，即“HNC 理论”。

处理迭句时将其拆分成两个独立的句子，即把主语“HNC 理论”赋予第 2 个小句组成独立句子。

这两句语句的句类代码都为 jDJ=DB+jD+DC。限于篇幅，直接给出所得出的形式化的“语境单元框架”：

{DOM//220 ：DB}

{SIT// jD [jlv111] | DB[fr820 *1] | DC[fv36 (r820, jr40-)]

DCB[fr820 *1]DCC[fv36]}

(BAC// [无])

其中，220 表示关于某一探索效应；jlv111 表示基本逻辑概念的肯定；fr820*1 特指某种探索的效应，f 表示专有名词；fv36 表示一种语习概念；(r820,jr40-)表示根据探索效应产生的一套理论系统。

(3)自然语言处理就是研究如何能让计算机理解并生成人们日常所使用的(如汉语、英语)语言，使得计算机懂得自然语言的含义，对人给计算机提出的问题，通过对话的方式，用自然语言进行回答。

根据算法，我们直接给出该句群形式化的：“语境单元框架”，有

{DOM//368 ：A}

{SIT//XD01[va60] | A[fga68 * 030506] | DBCXY[v000 8, vr810 ∨ g311 ∨ v802]DBCB[pw + jv30]DBCC[r65232 ∨ jgwa30/jgr 50 ∨ gr53821] }

(BAC//Cn[r 65232])

(4)自然语言处理领域中使用的理论和技术，包括句法处理、语义概念解释和上下文与世界知识 3 大部分。

根据算法，我们直接给出该句群形式化的“语境单元框架”，有

{DOM//371 ：DC}

{SIT// jD [jlv111] R011T0[v97 1# v653710] | DB[fgwa34 *1] | DC[gwa71] | TC[r820 ∨ ga62] }

(BAC//Cn[f ga68 * 030506])

这 4 个句群中，(1)与(3)的领域信息相同，都是 368，表示一类具体学科研究，而(2)的领域信息为 220，表示科学探索的效应，(4)的领域信息为 371，表示教育中的教授。从领域信息来看，(1)与(3)具有最大的相似性，其次为(2)句，(2)句的领域信息 220，提取于探索类概念，与 368 有较强的关联性。通过文本比较及结果生成算法的计算，可以得出理想的结果 0.99。进一步考察各句的广义对象语义块关系，可以看出(3)句中的对象与(1)句中具有 B(对象)性质的概念完全相同，都为自然语言处理，通过算法相似性计算可以得到相似性值 0.99，虽然(3)句中的 EK 与(1)句中的 EK 不一致，通过计算为 0.392，但是，与其他两句相比，(3)句已经具有最大的可信度、精确度。于是，返回结果的次序应该是(3)、(2)、(4)，这与用户要求相符。

与基于关键词的信息检索相比，此类检索方式具有比较明显的优势。Nature language processing 和 concept 在目标语言中对应的关键词“自然语言处理”和“概念”在(2)句和(4)中出现的概率要远远高于(3)句，根据关键词检索的可信度算法必然返回(2)句和(4)句作为最佳结果，而非(3)句。

4 相关试验及讨论

算法测试数据来源为“人民网”和“新华网”2004 年 1 月～2005 年 4 月的标准语料，所有测试语料由 57 401 个句子(22 550 个句群)组成，检索请求则根据语料中的相关内容预先设置，设置方法参照 TREC9 的 Cross-language Information Retrieval (English-Chinese) Track(参见 http://trec.nist.gov/)所提供的标准主题(CH55-Ch79)。同时，对应该主题，对答案进行了人工统计。

为了更好地验证算法的有效性，我们将根据算法实现的多语种信息检索系统(CCLIR)与基于聚类的单语种 Jelinek-Mercer 平滑模型(JMIM)，基于字典第一解释的 CLIR(DCFI)，基于语义的单语种信息检索(CMIR)进行了对比。实验数据如表 1 所示。

表 1 实验结果

Presision Recall	JMIM	DCFI	CMIR	CCLIR
0	0.720 1	0.642 1	0.816 1	0.807 1
0.1	0.577 9	0.522 7	0.741	0.713 7
0.2	0.454 5	0.434 2	0.663 6	0.662 2
0.3	0.373	0.353	0.582 1	0.561 5
0.4	0.344 7	0.332	0.543 1	0.512 1
0.5	0.272 1	0.210 1	0.434 1	0.420 4
0.6	0.232 2	0.155	0.412 4	0.412 3
0.7	0.179 9	0.166 7	0.322 1	0.322 3
0.8	0.151 4	0.141 2	0.201 5	0.200 9
0.9	0.120 4	0.116 1	0.141 2	0.124 1
1	0.038	0.030 1	0.123 2	0.112 2

各模型试验对比如图 2 所示。

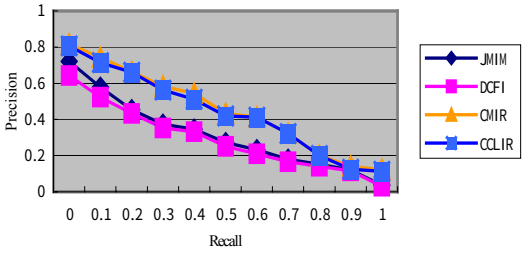

图 2 查准率-召回率对比

(下转第 19 页)