统一语义视图下的垂直领域跨语言检索模型*

孙晓玲　林鸿飞

大连理工大学计算机科学与工程系，大连 116024

E-mail: sunxiaoling1234@126.com , hflin@dlut.edu.cn

摘　要：随着 Internet 的快速发展和人们需求的不断提高，单语言的信息检索已经不能满足人们的需要，而网络语言的多样化和用户所掌握语言的差异性导致自由获取信息困难，因此跨语言检索受到了越来越多的关注。本文探讨了在生物医学领域的跨语言检索系统，利用医学本体 CMeSH 为检索语言和目标语言建立统一的语义视图。实验结果表明，统一语义视图下的垂直领域跨语言检索模型要比流行的机器翻译的方法效果有所提高。

关键词：跨语言检索，语义视图，语义标注，语义相似度，CMeSH

Unified Semantic View Based Cross-Language Retrieval Model in Vertical Field

Sun Xiaoling , Lin Hongfei

Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024

E-mail: sunxiaoling1234@126.com , hflin@dlut.edu.cn

Abstract: With the rapid development of the Internet and the improvement of user's needs, people are not satisfied with retrieving in the same kind of language, and Language diversity and differences led to the difficulty for users to retrieval information. So Cross-Language Information Retrieval (CLIR) receives people's more and more concerns. This paper discusses the cross-language retrieval model in biomedical field, and CMeSH is used to build a unified semantics view for original language and target language. The results show that the method of cross-language retrieval model in vertical field which is under the unified semantic view is better than the prevailing machine translation method.

Keywords: CLIR, Semantic View, Semantic Annotation, Semantic Similarity, CMeSH

1　前言

跨语言信息检索(简称 CLLR)是指让用户使用一种语言的“查询条件”在另外一种语言的“文档集”中进行检索。当查询条件不变，“文档集”扩展到多种语言时，通常我们把这个检索过程称为多语言信息检索[1]。CLIR 研究发展到今天，面向自由文本的方法成为主流技术。自由文本方法按照使用的翻译资源可分为：基于机读词典、机器翻译系统、本体或者基于语料库的方法。在生物医学领域，由于生物医学术语的特殊性使得建立一个高效的搜索引擎很具有挑战性[2]。但是生物医学领域拥有丰富且较完善的语义资源，譬如 MeSH、UMLS，这也使得基于本体的检索方法应用于该领域成为可能。

本文探讨了如何利用双语语义资源 CMeSH 为源语言和目标语言文本建立统一的语义表示，从而实现生物医学领域文本的跨语言语义检索。利用 CMeSH 为检索语言和目标语言建立一个统一的语义视图是本文的关键所在。统一语义视图具有以下两层含义：

1、 从文档中抽取能够反映文档语义特征的概念，采用概念对文档进行语义标注。同时将用户信息需求也转化为与之密切相关的概念，将查询和文档在表达上得到统一。

2、 建立基于概念的匹配机制。查询和文档映射到双语 CMeSH 中后，由于 CMeSH 严格规范的树状层次结构，可以用来进行语义相似度的计算，以达到更加精确的匹配。

*基金项目：国家自然科学基金资助项目（编号：60373095，60673039）和国家 863 高科技计划资助项目（编号：2006AA01Z151）。

本文组织如下：第二部分介绍了 CLIR 方面的相关工作；第三部分是系统采用的主要方法及细节；然后是系统的实验和结果分析，最后是结果与展望。

2 相关工作

Landauer 和 Littman 最早研究了基于自由文本的 CLIR，并利用一种自动技术来降低词汇差异对检索效果的影响。Radwan 和 Fluhr 在 1991 年提出一种使用了人工编码的翻译知识，通过提问式翻译策略实现 CLIR 的技术。尽管此后又取得很多进展，但上述两项研究中的基于语料库和基于知识的方法，仍占主导地位。 目前，面向自由文本的 CLIR 研究趋势是综合使用两种方式，以实现检索效率最大化。Twenty-One 系统就是一个综合利用了多种翻译知识的系统，它能支持荷兰语、法语、英语和德语，在词典进行提问式翻译的同时，使用基于语料库的歧义消解方法。

在特定领域中，领域词典已经被用来进行 CLIR[3]。[4]描述了使用 UMLS 进行德语和西班牙语检索 OHSUMED 语料的实验。但是这些方法都是用词典来进行查询翻译。而针对多语词典一种不同的使用方法则是融合文档分类技术，如潜在语义索引和广义向量空间模型，缺点在于依赖于双语对齐语料。

本文的方法类似于基于中间语言的 CLIR 方法。王进等[5]提出一种基于语义的跨语言信息检索模型 Onto-CL IR ,利用本体(Ontology) 在知识表示和知识描述方面的优势,解决查询请求从查询语言到检索语言之间转换过程中出现的语义损失和曲解等问题,从而保证在检索过程中能够有效地遵循用户的查询意图。MNIS-Text Wise 实验室的“概念中间语文献检索”(Conceptual Interlingua Document Retrieval)小组开发的 CINDOR 系统使用了较为独特的语间转换技术来实现 CLIR。该系统以 WordNet 的同义词群“synsets”为基础，通过将几种语言的同义词都链接到对应概念的“synset” 号上，建立了一个名为“概念中间语”的概念表示知识库。系统就可以将文献标引词和提问词都转换为“synset 号”，从而跨越语言的障碍。德国人工智能研究中心语言技术实验室(DFKI-LT)实现了 MuchMore 系统[6]，利用 MeSH 标注，实现了德英的跨语言检索。本文与之比较类似，将中英文文本都映射到 CMeSH 本体中，建立统一的语义视图，改进的是利用 MeSH 中的树状层次语义关系增加了语义相似度的计算，实现语义检索，使得检索效果更加精确。

3 主要方法

以 CMeSH 领域本体为参照，对中英文文档进行语义化处理，以本体中的概念对文档进行抽象，提取文档的语义向量，采用 XML 进行描述或存储到索引数据库中。用户检索时，对用户的查询语句进行同样的语义处理，即参照本体对查询语句进行抽象，抽取出查询语句对应的语义查询向量。最后计算查询向量与文档索引库中的语义向量之间的语义相似度，按照相似度的大小顺序向用户返回检索结果。系统的架构如下图 1 所示。

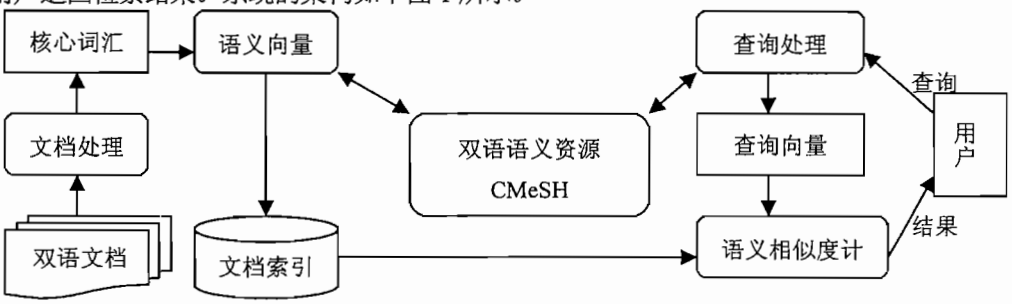

图 1 系统架构

本文的重点在于如何抽取语义向量和查询向量，以及用户查询时如何进行语义相似度的计算。下面主要介绍这两部分。

3.1 基于 CMeSH 的语义标注

《医学主题词表》(Medical Subject Headings,简称 MeSH)是美国医学图书馆编纂的一部大型医学专业叙词表，是手工检索 IM(Index Medicus)和计算机检索 Medline 的主题词文本，也是医学领域使用最广泛最具权威的词表。MeSH 中的术语经过严格规范，以树状结构表示。中文医学主题词表(CMeSH)，是中国医科院医学信息研究所出版的《医学主题词表》中文本，与 MeSH 中的概念一一对应，用于医学文献的标引、编目和检索。

生物医学术语的特性对 IR 系统的效率提出了挑战[7]。首先，生物医学词汇同义词和多义词非常多。其次，多个词组成的术语经常使用，这样使得传统的基于 bag-of-words 方法不太适用。第三，新的术语、缩略语很多，还有很多术语的变形。综上这些使我们认识到使用一种统一语义视图的重要性。通过将双语文本的关键字映射到双语 CMeSH 体系，把双语文本统一在相同的语义视图之下，实现从中文查询到英语的检索和挖掘。

用户的查询可能不规范，语义标注[9]就是用规范化的领域本体 CMeSH 的概念对文档和查询进行标注，从而达到一致，提高检索的性能，同时，将不同语言映射到统一语义空间，可以减少查询翻译等技术对跨语言检索效果的影响。

3.1.1 CMeSH 术语提取

一个 CMeSH 术语通常包含了两到三个单词。比如英文的“abdominal pain”，包含了“abdominal”和“pain”两个单词。因此，本文在从查询或文档中提取 CMeSH 词的时候，采取了一种序列化方法，其过程类似于前向最大匹配。扫描器首先扫描一个单词，若是 CMeSH 词，继续扫描下一个，直到找到一个最大匹配的 CMeSH 术语。然后继续从下一个单词开始扫描，依此类推，直至结束为止。

3.1.2 语义向量

对于一个术语，本文采用下面的语义信息对文档和查询进行标注：

1) Concept Unique Identifier (CUI)

2) Medical Subject Headings ID (CMeSH codes)

CUI 是一个概念的唯一标识，如果中英文的概念映射到一个 CUI 上，那么就达到完全匹配。但是只用 CUI 是不准确的，CMeSH 词之间的关系是通过语义联系起来的树状结构， CMeSH codes 即 CMeSH 术语在树状结构中的编码信息具有很明显的结构性[8]，能够代表它所属的类别和层次深度，可以用来进行语义相似度计算，从而达到中英文词语义上的匹配，提高检索精度。

语义标注为如图 2 所示的 XML 文档形式[9]，可能产生几种歧义情况。

1) 一个 Term 可能被分配了多个 CUI

2) 一个 CUI 可能映射到多个 CMeSH code

第一种情况被认为是一种歧义，把每一个 CUI 都当成独立的元素，分别进行标注。而对于第二种情况，一个 CUI 映射到多个 CMeSH code 时，则认为这些 CMeSH code 都从属于这个 CUI，是这个 CUI 的一种可能表示形式。

```
- <MeshTerm term="postoperative nausea">
  - <Concept concept="Postoperative Nausea" CUI="M0029913">
      <Mesh code="C23.550.767.859" />
      <Mesh code="C23.888.821.712.700" />
      <Mesh code="C23.888.821.937.059" />
    </Concept>
  </MeshTerm>
```

图 2 基于 CMeSH 的语义标注的形式

3.2 语义相似度计算

本文中语义相似度的计算是利用 CMeSH codes 进行的计算。综合使用本体类层次关系中的

多种影响因子，如语义距离，语义层次深度，信息内容及相应的调节因子等，来计算基于 CMeSH 的概念之间的语义相似度。

语义距离是指在本体树中连接两个节点的通路中的最短路径所经历的边数。语义距离是决定语义相似度的一个重要因素。通常来说，两个概念的语义距离越大，其相似度越低。语义深度是指两个概念的 LCS(Lowest Common Subsumer)的深度。一般来说，语义相似度不仅仅与语义距离有关，还与概念层次有关，相同距离的层次关系，语义相似度随着概念层次的递减而增高。因为层次越低，概念间密度越小，带来的差异的可能性也就越小。

将语义距离(l)和语义深度(d)结合起来，本文采用类似[10]中的公式，如公式 1 所示：

$$Sim_{dis}(C_1,C_2)=F(l,d)=\frac{f(d)}{f(d)+f(l)} \tag{1}$$

f 函数既可以是线性的也可以是非线性的。本文中 f 取非线性函数: $f(x)=e^{x}-1$,相似度计算公式如 2 所示：

$$Sim_{dis}(C_1,C_2)=\frac{e^{\alpha d}-1}{e^{\alpha d}+e^{\beta l}-2} \quad \alpha,\ \beta \text{ 是平滑因子} \tag{2}$$

信息内容方法也是节点间计算概念相似度的方法。相似性的值定义为概念节点的信息含量。一个节点的信息含量的值通过估计这个节点在大量文本语料库中出现的频率来获取。但是这样需要大量文本语料库。本文采用了 Nuno Seco[11]提出的一种直接基于本体结构来计算概念的信息内容的方法。假设如下：本体中的概念，它的同义词越多，信息内容就越少，即叶子节点的概念具有最多的信息量，形式如公式 3：

$$p(c)=\frac{\log\frac{hypo(c)+1}{\max_c}}{\log\frac{1}{\max_c}} \tag{3}$$

hypo 是一个概念的同义词的个数，$\max_c$ 是一个常量，代表本体中的概念的最大数目。在 CMeSH 中，$\max_c$=24767。

概念的信息内容计算出来之后，利用两个概念的共同父节点的信息内容来度量相似度的大小，如公式 4，$S(C_1,C_2)$ 是两个概念的共同父节点。一个概念可能有不止一个父节点，两个概念之间可能有多条路径，我们选取 p(c)最小的值。

$$p_{mis}(C_1,C_7)=\min_{c\in S(C_1,C_2)}\{p(c)\} \tag{4}$$

两个概念间的相似度如 5 所示：

$$Sim_{IC}(C_1,C_2)=1-p_{mis} \tag{5}$$

本文将上述(1)和(5)两种相似度结合起来，通过赋予一定的权值进行线性差值得到公式 6。这样可使概念相似度的计算更加全面，计算结果更加准确。

$$Sim(C_1,C_2)=\beta * Sim_{dis}(C_1,C_2)+(1-\beta)* Sim_{IC}(C_1,C_2) \tag{6}$$

β 为调节因子，由于语义相似度是一个主观性很强的概念，不同的应用，相似度依赖于各种影响因子也是各不相同，调节因子可以根据不同的要求来进行调整。

4 实验结果及分析

本文是生物医学领域的跨语言检索，但是目前缺少这方面权威的中英文双语语料。本文使用 TREC Genomic2004 的语料及 Topics。首先，进行单语检索实验，根据以往实验[9]，如果没有双

语测试集，单语中语义标注的效果同样可以用来预测跨语言检索的效果。然后进行了中文到英文的单方向的跨语言检索实验，请专家对 Genomic 的 Topics 进行翻译和校验，用于检索实验。

4.1 单语检索

表 1 单语检索

	MAP	R-prec	P@5	P@10
Baseline	**0.2794**	**0.3173**	**0.4880**	**0.4720**
Token-CUI	**0.2797**	**0.3204**	**0.4960**	**0.4700**
Token-CUI-CMeSHCodes	**0.2620**	**0.2988**	**0.4840**	**0.4440**
Token-CUI-CMeSHCodesSim	**0.2855**	**0.3293**	**0.4976**	**0.4730**

单语检索的实验结果如表 1 所示：

1. Baseline：使用原始的查询词进行检索。
2. Token-CUI：关键词和 CUI 同时建立索引。Query 的形式如下：

#combine(#syn(ATPase M0493578) and #syn(apoptosis M0026116))

3. Token-CUI-CMeSHCodes：关键词、CUI 和 CMeSHCodes 建立索引。
4. Token-CUI-CMeSHCodesSim：关键词、CUI 建立索引。查询的时候加入基于 CMeSHCodes 的语义相似度计算，对结果进行 Re-Rank。Re-Rank 的公式如 7 所示：

$$Sim(Q,D) = D(\theta_{Q+CUI} \| \theta_D) + \beta \cdot D(\theta_{Meshcodes} \| \theta_D) \quad (7)$$

检索结果的评价采用的是 TREC 评测中使用的评价方法。在 Baseline 的基础上，各种语义信息逐步加入到检索系统中，从结果可以看出，加入 CUI 的结果比 Baseline 提高一些，但是若仅仅将 CMeSHCodes 作为索引项来处理，反而使效果降低。因为 CMeSHCodes 在 CMeSH 中具有一种树状的层次关系，不能仅仅因为两个 CMeSHCode 不匹配，就认为不相关，这样会影响检索的效果。加入基于 CMeSHCodes 的语义相似度计算模型之后，原本不能匹配的项之间可以计算相似度，达到了语义概念上的匹配，效果得到提高，证明了语义检索具有一定的可行性。

4.2 跨语言检索

将英文 Topic 经专家翻译成中文检索英文文档，中文 Topic 和英文文档都映射到 CMeSH 中，利用 CUI 和 CMeSHCodes 进行标注，在统一语义视图下实现单方向的跨语言检索。

表 2 跨语言检索

	MAP	R-prec	P@5	P@10
MT	**0.1330**	**0.1581**	**0.2250**	**0.2312**
CUI	**0.0847**	**0.0907**	**0.1784**	**0.1629**
CUI-CMeSHcodesSim	**0.1492**	**0.1724**	**0.2300**	**0.2319**

跨语言检索的实验结果如表 2 所示:

1. MT：用机器翻译方法将翻译好的中文 Query 翻译成英文。采用的是 Google 的翻译系统。
2. CUI：利用标注的 CUI 进行检索。
3. CUI-CMeSHcodesSim：在方法 2 的基础上加入语义相似度计算模型。

机器翻译的方法在本文中作为一个 Baseline，结果如表 2 所示。机器翻译方法的缺点是对于较短的提问式,由于没有上下文和相关领域知识信息源,使得机器翻译系统在消除词语的歧义性方面存在较多困难。第二种方法则是仅仅使用了 CUI 标注信息，可以看出结果很低，很大程度上是因为查询词中的术语不能对应到 CMeSH 中，比如“DO1 抗体”这样的查询词，只有抗体可以在 CMeSH 中找到对应的概念，而 DO1 则找不到，这样检索的效果就要降低很多，还有一方

面的原因是提取 CMeSH 词时准确度的影响，提高 CMeSH 词抽取的精度和粒度是提高本文检索精度的关键所在。和单语检索效果类似，加入基于 CMeSHCodes 的语义相似度计算模型之后，效果明显提高，证明了加入语义信息的重要性，但是由于前面提到的抽取 CMeSH 对应概念的精度等因素影响，效果相对于 MT 方法的提高不是很明显。系统还有很大提升的空间。

5 总结与展望

本文实现了基于 CMeSH 的生物医学领域的跨语言检索，将中英文文本映射到 CMeSH 中，建立了统一的语义视图，并进行了语义相似度的计算，实验结果证明方法是有效的。当然本文的方法还有很多需要改进的地方，如：CMeSH 词提取的粒度方面，CMeSH 词往往由多个单词组成，本文中采用最大匹配，粒度较大，很可能丢失了很多有用的词汇；语义标注方面使用的语义信息还比较少，以后还可以逐渐增加。

跨语言信息检索已成为世界范围内研究的热门课题，但是跨语言检索系统的性能整体上和单语言检索仍有很大差距。今后的工作将主要针对上述的问题开展，同时，基于查询扩展的检索模式在实践中得到验证，可以大大提高检索的精度，因此还可以在用户的语义概念查询扩展方面进行研究。我们将继续努力，希望为这个领域的研究发展贡献一份力量。

参 考 文 献

[1] 闵金明,孙乐,张俊林.重新审视跨语言信息检索[J].中文信息学报,2006(4):33-40.

[2] G. Nenadic, I. Spasic, and S. Ananiadou. Mining biomedical abstracts: What's in a term? IJCNLP, 2004:797–806.

[3] Gey F. C., and H. Jiang. English-German Cross-Language Retrieval for the GIRT Collection-Exploiting a Multilingual Thesaurus. In: Proc. of the Eighth Text Retrieval Conference (TREC-8), National Institute of Standards Technology (NIST), 1999.

[4] Eichmann D.,M. Ruiz, and P. Srinivasan. Cross-Language Informaiton Retrieval with the UMLS Metathesaurus. In: Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

[5] 王进, 陈恩红, 张振亚, 王煦法. 基于本体的跨语言信息检索模型 [J]. 中文信息学报, 2004(3): 1-8.

[6] M. Volk, B. Ripplinger, S. Vintar, P. Buitelaar, D. Raileanu, and B. Sacaleanu. Semantic annotation for concept-based cross-language medical information retrieval. International Journal of Medical Informatics, 2002, 67(1/3):79—112.

[7] Dolf Trieschnigg. Biomedical Cross-Language Information Retrieval. SIGIR'08, 2008: 897.

[8] Von-Wun Soo, Chen-Yu Lee, Chung-Cheng Lin, Shu Lei Chen and Ching-chih Chen. Automated Semantic Annotation and Retrieval Based on Sharable Ontology and Case-based Learning Techniques. In the Proceedings of JCDL, 2003.

[9] Vintar Špela, Buitelaar Paul, Ripplinger Bärbel, Sacaleanu Bogdan, Raileanu Diana, Prescher Detlef. An Efficient and Flexible Format for Linguistic and Semantic Annotation. In Proceedings of LREC, 2002.

[10] Xiaoying Liu, Yiming Zhou, Ruoshi Zheng. Measuring Semantic Similarity in WordNet. In Proc. of the Sixth International Conference on Machine Learning and Cybernetics, 2007: 3431-3435.

[11] N. Seco, T. Veale, and J. Hayes. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. ECAI, 2004: 1089-1090.