一种基于信息抽取和文本生成的多语种信息检索模型[1]

姚天昉[2]
(上海交通大学计算机科学与工程系 上海 200030)

徐飞玉
(德国人工智能研究中心 D-66123 德国 萨尔布吕肯)

张冬茉 李芳 王纤 盛焕烨
(上海交通大学计算机科学与工程系 上海 200030)

摘要:
本文提出了一种多语种信息检索模型。它主要基于信息抽取和文本生成技术。这个模型既体现了信息抽取技术所提供信息的精炼性和准确性，又体现了文本生成技术所提供检索结果的连贯性和规范性。
发挥了这两种技术的综合优势。在本文中，主要介绍了模型中信息抽取部分所采用的技术，包括多语种信息抽取、基于概念的多语种义类词典、模板自动开发、多语种信息检索和索引等。这个模型为在因特网上进行多语种特定专业领域信息检索提供了一种准确、快速、方便的计算机自然语言多语种信息检索手段。
关键字: 多语种信息检索模型, 多语种信息抽取, 基于概念的多语种义类词典, 模板自动开发, 多语种信息检索和索引

A Multilingual Information Retrieval Model based on Information Extraction and Text Generation

YAO Tian-Fang
(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 100030)

XU Fei-Yu
(German Research Center of Artificial Intelligence (DFKI), D-66123 Saarbruecken, Germany)

ZHANG Dong-Mo, LI Fang, WANG Qian and SHENG Huan-Ye
(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 100030)

Abstract This paper proposes a multilingual information retrieval model which is principally based on information extraction and text generation techniques. The model embodies both the conciseness and accuracy of the retrieval results provided by information extraction technique, and the coherence and standardization of ones supplied by text generation technique. It synthesizes the advantages of both techniques. In this paper, we mainly present the information extraction techniques adopted in the model, including multilingual information extraction, concept based multilingual thesaurus, template automatic development, multilingual information retrieval and index etc. We build the model for providing the multilingual information retrieval means on Internet, which is accurate, quick and convenient and is used in a specific domain.
Key words multilingual information retrieval, multilingual information extraction, concept based multilingual thesaurus, template automatic development, multilingual information retrieval and index

1 引 言

由于在万维网（WWW）和其它领域中的信息呈“爆炸”之势。人们对智能信息检索引擎的需要大大增加了。智能信息检索引擎不仅应该帮助人们很快地找到所需的信息，而且应该使人们能容易地理解信息。所以，信息存取（Information
Access）包含了两层意思：一是信息管理能使人们方便地找到所需信息；二是信息表示能使人们容易地理解信息。

众所周知，自然语言是最常用的人类社会交流工具，也最容易使人们理解。虽然目前大多数信息是用英语来表达的，而且许多用户能够理解英语。但是，仍然有相当部分的信息是用非英语来表达的，而且也存在相当多的用户

[1] 本课题得到国家自然科学基金的资助（项目编号：60083003）
[2] 第一作者在德国人工智能研究中心的资助下完成此文。在此表示感谢。

，他们不能理解这些语言。例如：对大多数德国用户来说，他们不懂汉语；同样，对大多数中国用户来说，他们也不懂德语。近几年来，越来越多的汉语和德语信息出现在因特网上。针对这种情况，德国的用户希望以德语作为询问项和检索结果语言来检索网上汉语的信息。同样，中国的用户也有与此语言相反的要求。

笔者提出的模型能使多语种信息存取（Multilingual Information Access）成为可能。在这个模型中，笔者采用信息检索技术与其它高级的语言技术如基于概念的多语种义类词典（Concept Based Multilingual Thesaurus）、信息抽取（Information Extraction）和多语种文本生成（Multilingual Text Generation）等相结合来对汉语、英语和德语网上信息进行信息检索。

比较一下数据库中所存储的信息结构，人们就会发现大多数Web文档中的信息是半结构化而且是异质的（由不同成分组成的）。基于字符串或模式（Pattern）的传统信息检索技术对于建立索引和搜索是与语言无关的。它通过使用简单的模式匹配来帮助用户较快地找到信息。但是，当所存取的信息是用用户不能理解的语言来表示的话，这种传统的信息检索技术就不能帮助用户完成信息存取。此外，传统的信息检索系统不能以简洁的方式提供文档中的信息，这样用户要花费更多的时间来浏览整个文档的内容。在本模型中，笔者将组合信息抽取和文本生成技术，最终使用户能以他们所熟悉的自然语言而无须浏览整个文档快速地存取信息。

使用信息抽取技术从文档（例如Web文档）中抽取特定专门领域的信息并用一个与此领域相关的模板来表示这些信息。模板可以被认为是关系数据库中的数据项的组合。在本模型中，笔者将采用与语言无关的模板表示。在此基础上，采用多语种文本生成技术来生成模板所表示的内容，用户可以选择他所熟悉的语言作为所生成的文本语言。使信息抽取和多语种文本生成技术相结合来完成多语种信息检索。这种模型有下列优点：所有的以三种语言表示的特定专业领域的信息能被以一种语言无关的方法来抽取；所有被抽取的信息能被任何熟悉三种语言之一的用户所存取；如果模板能够表达文档主要内容的话，那么信息抽取和多语种文本生成技术的结合实际上达到了基于知识的文摘系统的效果。

2 多语种信息检索模型

笔者建立本模型的目标是：

- 能够有效、快速、方便地对多语种文本进行信息抽取；
- 能够有效地利用信息抽取得到的结果进行文本生成以完成多语种信息检索任务；
- 通过对原型系统的性能评价，展示多语种信息检索模型具有较高的查准率（Precision）和查全率（Recall）。

由于笔者对多语种文本生成已进行过较为深入地研究，也发表过一定数量的论文。有兴趣的读者可参看[1, 2, 3, 4]。因此，笔者在本文中主要介绍本模型中被采用的多语种信息抽取、基于概念的多语种义类词典、模板自动开发、多语种信息检索和索引等技术方案。

2.1 多语种信息抽取

多语种信息抽取所带来的直接问题是：如果以某种语言信息加以抽取的话，是否可以得到另一种语言的信息作为抽取结果。其次，如果在系统中再增加一种可抽取的新的语言的话，如何最大程度地重用算法和与领域有关的单元，最小可能增加与语言有关的机制和数据源[5]。

笔者针对上述问题的对策是：采用与语言无关的模板作为抽取信息的中间结果，然后采用基于模板的多语种文本生成技术产生某一种语言的文本作为抽取结果。其次，采用前端词法/句法/语义分析与语言有关，而后端模板填充与语言无关的信息抽取策略。这样，如果增加一个语种的话，只要再增加该语种的词法/句法/语义分析单元即可。按照这样的考虑，笔者设计的多语种信息抽取过程如图1所示。

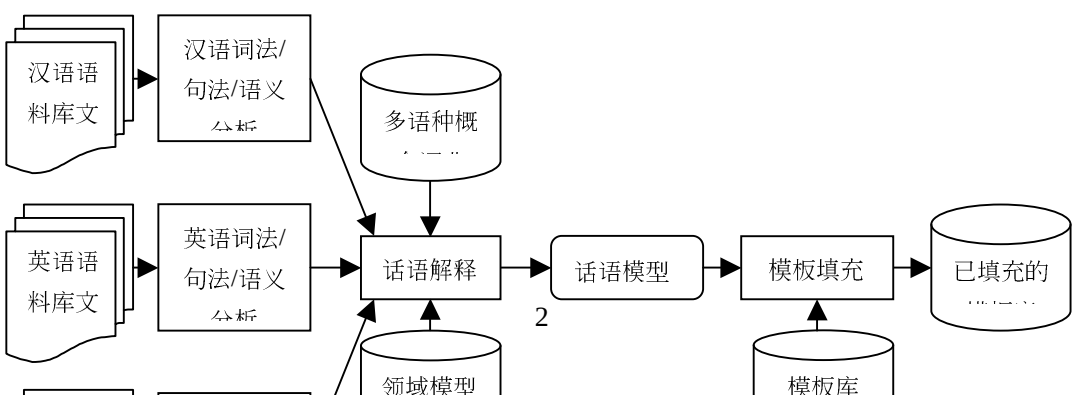

图**1** 多语种信息抽取过程

从对应语言的语料库的半结构化文本如Web网页中得到待处理的文本作为输入，然后对它进行相应的词法、句法和语义分析。语义分析的结果用谓词-变元结构表示。同时作为话语解释（discourse interpretation）单元的输入。话语解释单元的主要任务是集成多个语句的谓词-变元结构成为单一的话语模型。这个模型中的信息包括实体（entity）、对象（object）、事件（event）和属性（attribute）等。它将提供给后续模板填充单元使用。话语解释单元依赖于与语种无关的的领域模型。领域模型由本体（ontology）和有关的属性知识库组成。本体可用一个有向图表示。它的结点表示与语言无关的类或概念。这些类或概念与模板填充规则直接有关。同时，与本体结点有关的是属性-值结构。其中值可能是固定的，也可能依赖于与模型中其它信息有关的各种条件。根据这些条件对其进行赋值。属性-值结构的集合就形成了属性知识库。在话语解释单元的处理过程中，首先将每一个语句的谓词-变元结构以及与它有关的属性被加入到话语模型中。然后，通过多语种概念词典将谓词-变元结构映射到领域模型本体中的结点上。与领域模型中的类或概念以及属性知识库中的属性-值结构有关的信息同时被加入到话语模型中。在涉及句间谓词-变元结构的处理时，同时处理互指求解（coreference resolution）。对话语模型中实体结点间所涉及到的事实的因果关系进行适当的逻辑推理就可得出进一步的事实信息。从而可以得到新的模型片断（model fragment）。根据话语模型和与语种无关的模板就可以进行模板填充了。如果填充成功，则表示已获取到指定的信息。填充后的模板作为文本生成单元的输入。

对于具体某一语种的词法、句法、语义分析单元。笔者按照“由易到难，分层处理”的原则进行。词法分析是对文档进行预处理、划分词单元、确定词类。名称分析是对各种名词进行标识，通过名称匹配，标示出敏感的名称词组。部分句法分析是通过分析部分语法结构，例如名词短语、动词短语等，并确定相互之间的包容、依赖、执行等关系，符合条件的短语则予以合并。场景模式匹配是通过场景模式进行结构分析，取出正确的事实内容。最后，语义分析将事实内容转换成谓词-变元结构。

对于汉语处理来说，由于词法分析和句法分析技术在国内已有了深入地研究，并已取得了许多成果。笔者打算部分利用一些现成的工具，如[6]，以致于可以集中精力来进一步完成其它单元的开发。以探索一个与汉语信息抽取相适应的多语种信息抽取系统体系结构，而这种体系结构同时又广泛地适合英语和其它西方语言。

多语种信息抽取方法还存在着灵活性和适应多领域问题[7]。笔者在这方面将进行专业领域模型的研究，使它能应用于模板自动开发中。在构造专业领域模型时，将体现它的通用性，至少应该对领域移植来说是方便的。

2.2 基于概念的多语种义类词典

笔者在模型中采用基于概念的多语种义类词典是基于这样的考虑：一方面它可以解决多语种询问项之间翻译的问题；另一方面它可以帮助找到另一语言在意思或含义上的最相近的询问项。询问项翻译策略大部分是用在交叉语言(cross-lingual)信息检索方面的。语言A的询问项将被翻译成语言B的询问项。被翻译的询问项将被用来进行B语言文档的检索。为询问项翻译而需生成双语或多语种的字典，笔者将采用基于对齐方法的双语词汇自动获取方法（bilingual term alignment）[8]来创建这种字典。这种方法将利用不同语种并行文档作为语料来获取对应的将被翻译的双语词汇，从而建立一些特定专门领域的双语或多语种义类词典。它的优点在于可以动态地生成一些特定专门领域的义类词典，及时采集不同语种所出现的新词汇。这将比一般的双语词典更精确地提供对应翻译词汇，也可以解决不同语种词汇间多对多的关系问题。当然，除了并行文档以外，笔者还将研究不同语种的异构文档。

此外，由于基于概念的多语种义类词典是自动生成的。因此，如果增加一个语种的话，只要再产生包含该语种的基于概念的多语种义类词典即可。这就增加了系统的灵活性。另外，由于义类词典是基于概念的，对于交叉语言抽取内容的概念等效性是有利的。

2.3 模板自动开发

在完成信息抽取任务时，最耗时的工作之一是定义一些模板的填充规则。这些模板的填充规则不但决定信息抽取的内容和范围，而且还影响其后基于模板的多语种文本生成的质量。笔者决定在本模型中采用机器学习的方法来自动产生模板填充规则。

基于实例的学习方法对产生模板填充规则是适用的[9, 10]。因为笔者用于本模型的模板最终是与语种无关的。所以，所采用的填充规则不但要能够体现各语种的个性，而且要能够体现它们的共性。要从句法和语义，特别是概念层面上获取信息。在构造学习算法时，要充分注意同义词的联系。借助于概念词典[11, 12]，扩大填充规则的适用范围。为了提高机器学习的正确率,笔者将训练的语料先进行人工加工标注。这样，在保证输入文本附加信息正确的前提下，提高模板自动开发的效率和正确率。

2.4 多语种信息检索和索引

在确定了多语种信息抽取方法、多语种词典构造以及基于模板的多语种文本生成方法之后，就可以着手建立多语种信息检索模型。这个模型意味着虽然在系统处理信息的过程中会涉及到其它用户不熟悉的语言，但这些处理对用户来说都是透明的。用户只要考虑怎样输入他所熟悉语言的询问，就可以得到他所熟悉语言的信息检索结果。图2表示了系统的搜索过程：

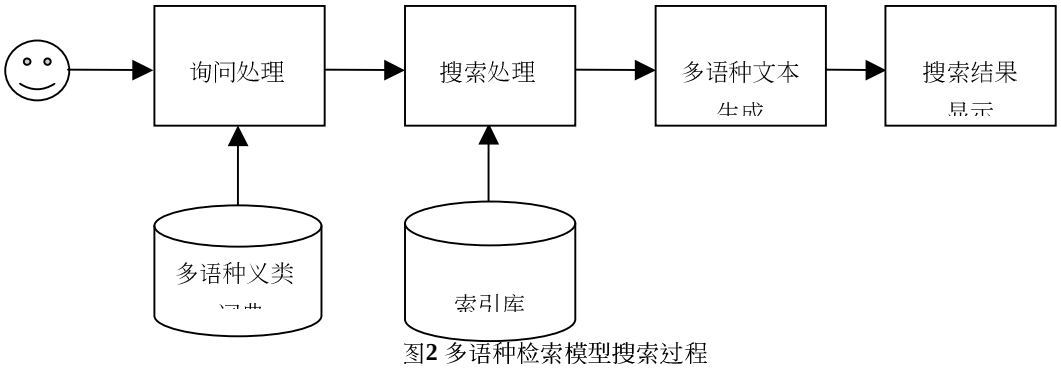

图2 多语种检索模型搜索过程

- 某一语言的询问项（可以是单词、词组或一个句子）将在基于概念的多语种义类词典的帮助下翻译成另一语言的询问项；
- 翻译好的询问项将被用来作为搜索项，从而搜索与其对应语言的文档。当文档被找到后，根据索引确定其对应的模板。然后对文档进行分析和信息抽取。使用文档的信息对相应的模板槽进行填充；
- 采用已填充成功的模板内容选择用户熟悉的语言进行自然语言文本生成。所生成的文本作为最终的检索结果提供给用户。

为了提高搜索效率，笔者将为各语种Web文档和有关的模板建立索引。在建立上述索引时，将采用标准的向量空间模型方法，使模板能更精确地匹配Web文档。模板索引和文档标识之间的对应联系被存储在数据库中。

例如：存在三个模板，分别标记为temp1, temp2和temp3，它们对应于一个文档，标记为doc_1。为这些模板与文档doc_1的联系建立一个索引：doc_1: templ, temp2, temp3

2.5 其 它

笔者将在多语种信息检索系统开发中协调语言工程和软件工程的关系[13]。充分注意按照软件工程的规范进行系统开发。笔者将选择面向对象的系统设计技术和开发工具，以便于系统集成和模块修改。如采用Java语言作为系统开发工具，便于移植和推广；采用XML（Extensible Markup

Language）语言作为数据交换形式，保证所处理文本格式的适应性；采用灵活的可视界面开发语言XSL（Extensible Stylesheet Language），保证界面的友好性等。

3 结 论

基于信息抽取和文本生成技术的多语种（包括汉语）信息检索模型的研究无论国内还是国外目前未见成果报道。加强这方面的研究，使之早出成果，这在学术上有着十分重要的意义。同时，所研制的成果有着广阔的应用前景。它具有明显的社会效益和经济效益，可以在各个专业领域应用。由于与网络技术联系紧密，易于获得推广性成果。

笔者提出的多语种信息检索模型充分考虑到目前国内外此类研究中的不足之处，以人工智能和计算语言学理论为基础，确定符合知识工程的专业领域规则库和信息库，并能结合网上信息查询的技术特点，利用NLP技术优势，进行智能化的基于概念层次的多语种信息搜索，而不同于传统意义上的基于词语匹配层次的信息搜索。它的主要特色是：

- 在国外的多语种信息检索研究中，所处理的语种主要是西方语言，对东方语言（如汉语）还不能完全适用。在本模型中，既参考了国外的多语种信息检索模型，又保持中文信息处理的特色，建立新型的、适合于东西方多语种的多语种信息检索模型，体现了多语种文本处理中的个性和共性；
- 本模型采用与语种无关的核心结构，既解决了模型的模块化问题，又可以结合各语种的语言知识库对多语种进行分析处理，发挥了所有语言的表达潜力；
- 能够通过多语种信息抽取、基于概念的多语种义类词典和多语种生成技术对多语种和异质的信息进行多语种和交叉语言信息存取；
- 采用机器学习自动开发模板以及建立具有多领域适应性的领域模型，使得尽可能适应其它专业领域的信息存取；
- 避免由于使用机器翻译系统而带来的翻译质量和语种限制问题。

参 考 文 献

1 Tianfang Yao, Dongmo Zhang, Qian Wang. MLWFA: A Multilingual Weather Forecast Text Generation System. In Proc. of 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000): Tutorial Abstracts and Demonstration Notes. Hong Kong, China, Oct., 2000.

2 Feiyu Xu , Klaus Netter and Holger Stenzhorn. MIETTA-A Framework for Uniform and Multilingual Access to Structured Database and Web Information. In Proceedings of IRAL 2000, Hong Kong.

3 王纤， 姚天昉.汉语天气预报文本内容规划器的设计与实现. 见：陈力为、袁琦主编《语言工程》. 清华大学出版社. 中国，北京. 1997年8月.

4 张冬茉, 姚天昉, 王纤.句子规划器的设计与实现. 上海交通大学学报. 第32卷第10期.中国，上海. 1998年10月.

5 Tianfang Yao, Qingzhong Gao. A Multilingual Surface Generator with FB-LTAG. In Proc. of the Natural Language Processing Pacific Rim Symposium 1999. Beijing, China, Nov., 1999.

6 Robert Gaizauskas, Kevin Humphreys, Saliha Azzam,Yorick Wilks. Conceptions vs. Lexicons: An Architecture for Multilingual Information Extraction. Lecture Notes in Computer Science; Vol.1299; Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, Germany. Nov., 1997.

7 刘开瑛. 中文文本自动分词和标注. 商务印书馆. 北京. 2000年5月.

8 Yorick Wilks and Roberta Catizone. Can We Make Information Extraction More Adaptive. Lecture notes in computer science; Vol.1714; Lecture notes in artificial intelligence. Springer-Verlag, Berlin, Germany. Oct., 1999.

9 Fang Li and Wilhelm Weisweber. Bilingual Lexicon Extraction From Internet. In Workshop Proceedings "Terminology Resources and Computation" in LREC 2000 Second International Conference on Language Resources and Evaluation. Athens, Greece. May 2000.

10 Scott B. Huffman. Learning information extraction patterns from examples. In S. Wermter, E. Riloff, and G. Sheler, eds., Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing. Springer-Verlag, Berlin, Germany. 1996.

11 Mary Elaine Califf and Raymond J. Mooney. Relational Learning of Pattern-Match Rules for Information Extraction. In Proc. of the 16th National Conference on Artificial Intelligence (AAAI-99). Orlando, Florida, U.S.A. July, 1999.

12 URL: http://www.cogsci.princeton.edu/~wn/

13 URL: http://www.keenage.com/

14 Roberto Basili, Massimo Di Nanni,Maria Teresa Pazienza. Engineering of IE Systems: An Object-Oriented Approach. Lecture notes in computer science; Vol.1714; Lecture notes in artificial intelligence. Springer-Verlag, Berlin, Germany. Oct., 1999.