第 37 卷 第 2 期
Vol.37 No.2

计 算 机 工 程
Computer Engineering

2011 年 1 月
January 2011

• 人工智能及识别技术 •　　文章编号：1000—3428(2011)02—0157—03　　文献标识码：A　　中图分类号：TP391

基于文本分析的知识获取系统设计与实现

姚金国，代志龙

(复旦大学计算机科学与技术学院，上海 200433)

摘　要：知识获取一直以来都是构建专家系统的瓶颈问题。针对该问题，利用自然语言处理技术，设计并实现一个针对化学科技文献进行分析的知识获取系统，并对其关键技术进行分析。系统对输入文本进行分词及词性标注，在此基础上，使用 Tregex 在句法分析树上进行实体识别，同时利用依存关系进行搭配词识别。实验结果表明了该系统的有效性。

关键词：知识获取；句法分析；文本分析

Design and Implementation of Text Analysis Based Knowledge Acquisition System

YAO Jin-guo, DAI Zhi-long

(School of Computer Science, Fudan University, Shanghai 200433, China)

【Abstract】 Knowledge acquisition is a bottleneck to develop expert system. Aiming at this problem, it designs and implements a knowledge acquisition system to analyze the chemical literatures. Meanwhile, key techniques are expatiated. NLP approaches are used in this system to facilitate the automated extraction of chemical knowledge. The system segments the input text into words and assigns part of speech to each word. Tregex is used to identify the named entities on parsing trees and typed dependencies are utilized to extract qualifier for each entity. Experimental results indicate that the approaches for this task are highly efficient.

【Key words】 knowledge acquisition; syntax parsing; text analysis

DOI: 10.3969/j.issn.1000-3428.2011.02.054

1　概述

知识获取是一个与领域专家、专家系统建造者以及专家系统自身都密切相关的复杂问题，由于各方面的原因，知识获取至今仍然是一件相当困难的工作，被公认为是专家系统建造中的一个瓶颈问题。专家知识是人们在长期的生产实践中积累起来的财富，是领域知识与具体问题的解决方案相结合的产物，应该被人们共享。具体说来，使用专家知识有很多好处。首先，以专家知识构建的专家系统能够在无人监督的情况下，高效、准确、迅速地工作；其次，使人类专家的领域知识突破时间和空间的限制，专家系统程序可永久保存，并可复制任意多的副本或在网上供不同地区或不同部门的人使用。

知识获取的传统方式是不断地与领域专家交流，将得到的知识手工或者通过智能知识编辑器以预先定义好的方式存储。但由于知识工程师一般不具有该领域特定的知识背景，并且往往在沟通和思维方式上与领域专家存在偏差，这些问题使得知识获取过程变得无法控制，获取到的知识可信度也无法保障。因此，研究如何从其他的知识源(如科技文献、网络等)自动或半自动地获取知识已经显得尤为重要。研究这些技术，不仅能缩短知识获取的周期，还可以将知识源的领域从专家扩展到网页、技术文档[1]等。

在特定的专业领域中，领域专家将结论性的知识或经验用规范化的科学语言精确地表示出来，并以文本的形式对其进行存储和共享，这使得科学文献中有很多专门化的领域知识可以利用。而且从知识工程的角度来看，科技文献中的陈述都是经过验证的，概念化程度高，离专家系统中知识的形式化要求距离接近。在生物领域，文献[2]提出了基于文献的知识发现的理论，通过统计生物医学文献的标题、关键字中同时出现的专有名词，发现了大量非相关生物文献中隐藏着的不为人知的知识。

本文针对化学类(聚合物合成)科技文献的摘要及结论部分，采用自然语言处理技术，设计并实现一个化学类知识的自动抽取系统，以满足计算机辅助知识获取的需求。具体说来，系统抽取的每一条化学知识包括实验环境、实验措施、实验结果三部分，如表 1 所示。其中实验环境指聚合物合成时的外部环境，包括温度、压强、聚合物类型等；实验措施由调节参数和调节词组成；实验结果由目标性能及结果词组成。显然这类化学知识对于化学实验人员来讲，具有很强的指导意义，能够极大地减少重复实验，提高工作效率。

表 1　化学知识模板

实验环境	聚合物类型(如 NEPE 推进剂)
实验措施	调节参数(如 RDX 含量、氧化剂粒度等) 调节词(如增加、减少、加入等)
实验结果	目标性能(如比冲、密度等) 结果词(如增大、减小等)

2　系统体系结构

如上所述，系统主要可以通过以下 2 步来实现化学类知识的自动抽取：

(1)识别每一条化学类知识中的实体，包括聚合物类型、调节参数、目标性能等。

基金项目：国家自然科学基金资助项目(60303007)

作者简介：姚金国(1986－)，男，硕士研究生，主研方向：专家系统；代志龙，博士研究生

收稿日期：2010-07-17　　**E-mail：**kinguoyao@gmail.com

(2)分别为调节参数及目标性能寻找其搭配词，即调节词、结果词。

系统体系结构如图 1 所示，系统以中文文本为输入，首先对文本进行分词及词性标注。然后对一个输入语句进行句法分析，生成其句法分析树、依存关系树。最后系统采用 Tregex[3]在句法分析树上进行实体识别，同时利用依存关系进行搭配词识别。

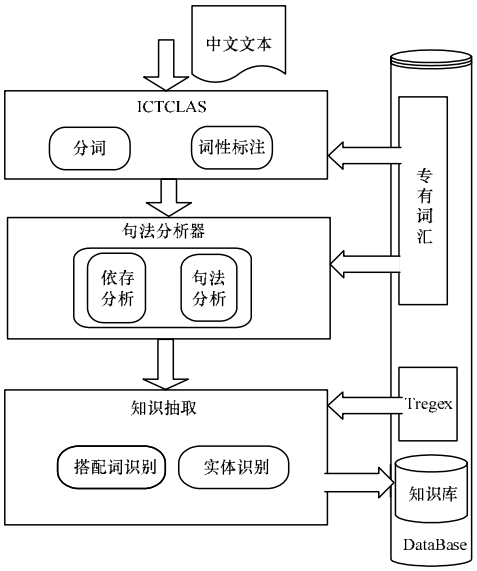

图 1 系统体系结构

3 关键技术分析

3.1 分词及词性标注

汉语的书写以汉字作为基本单位，词与词之间没有明显的形态界限，要进行汉语的计算机处理，必须首先将汉语的词与词分割开，即分词。自动分词后的文本是一个词串。对其中每个词，一般而言，在文本的特定语境下，每个词的词性和语义都是唯一确定的，这也正是人们能正确理解给定文本的基础。

本文使用 ICTCLAS 对输入的文本进行分词及词性标注。ICTCLAS 系统采用了层叠隐马尔可夫模型，将汉语词法分析的所有环节都统一到了一个完整的理论框架中，获得最好的总体效果。

与一般的文本相比，关于聚合物合成的科技文献具有很强的专业性，为了提高分词及词性标注的准确性，在外部定义一个专有词库，用于存储各种专有词汇，如氧化剂、比冲、铝粉等。专有词库的引入可以使得系统能够优先对词库内的词进行切分，从而提高分词及词性标注的准确性。例如，对语句“铝粉粒度级配的改变，可以使高能推进剂比冲效率由 0.88 提高到 0.92”。

在引入专有词库前，其分词结果为“铝粉粒度级配的改变，可以使高能推进剂比冲效率由 0.88 提高到 0.92”；在引入专有词库后，其分词结果为“铝粉粒度级配的改变，可以使高能推进剂比冲效率由 0.88 提高到 0.92”。

经过分词及词性标注后的文本在系统内的数据结构如图 2 所示。每个句子包括原始句子、分词后的句子(词与词之间用空格隔开)、词性标注后的句子以及各个词。一篇文章由句子组成，为了便于后续处理，同时存储了每个词在文章中出现的位置。

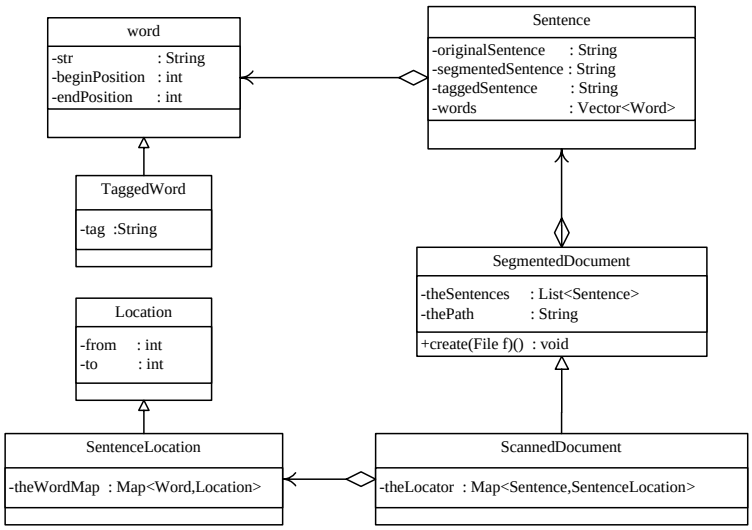

图 2 类图设计

3.2 句法分析

句法分析将分词及词性标注器的输出作为输入，为输入的语句生成其句法分析树及依存关系树。人们说话的方式都存在一些结构和规则，语法分析的目标就是努力分离出这些语法结构。句法分析树描述的是如何聚集，以及中心词和附属词之间是如何相互联系的，依存关系描述的是词与词之间的关系。

在依存关系中，某个词是句子的中心词，其他词或者依赖于这个词，或者依赖于那些经过一系列依存关系与中心词有联系的其他词。

本文采用 Stanford Parser[4]进行句法分析，为每一个输入语句生产其句法分析树及依存关系。例如，对于语句“HMX 粒度降低，NEPE 推进剂燃速降低，压力指数降低”，其句法分析树、依存关系对分别如图 3 和表 2 所示。

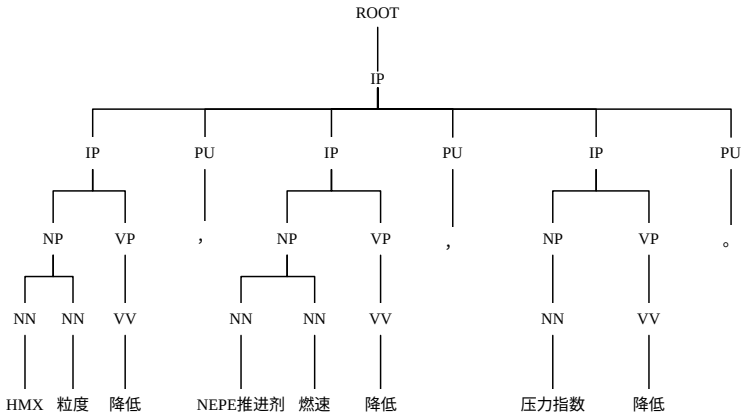

图 3 句法分析树

表 2 依存关系对

依存关系	依赖项	修饰项
nmod	粒度-2	HMX-1
nsubj	降低-3	粒度-2
nmod	燃速-6	NEPE 推进剂-5
nsubj	降低-7	燃速-6
ccomp	降低-3	降低-7
nsubj	降低-10	压力指数-9
ccomp	降低-3	降低-10

在上述依存关系对中，每个词后面的数字表示词在句子中的位置，如粒度-2 表示粒度是句子中的第 2 个词。其中，nmod 表示名词修饰关系[5]。

3.3　知识抽取

在系统中，首先使用 Tregex 以实现对调节参数、目标性能等实体进行有效的识别。Tregex 由于其操作的方便性以及跨平台性，目前在自然语言处理领域也得到了广泛的应用。Tregex 类似于对字符串进行处理的正则表达式，它能够对句法分析树上的特定模式进行识别。例如，“ NP<<#含量 ”用于在句法分析树上查找 “ 以含量作为中心词的 NP 节点 ”。

在实现化学知识抽取的过程中，词间关系的识别是重要的一步，如调节参数与调节词、目标性能与结果词等词之间的搭配关系。对于这个任务一种比较直观的做法就是最短距离法。其主要思想就是首先定义出一个调节词的词典 d1、一个结果词的词典 d2，然后对每个调节参数或目标性能在句子中寻找与其距离最近的调节词或结果词。这一方法没能对待处理的句子进行任何有效分析，其缺陷显而易见。如对这样的句子“ 燃速随 RDX 粒度的减小而降低 ”关于燃速的搭配就是<燃速，减小>而非<燃速，降低>。而本文依据依存关系，可以直接找出燃速的搭配词即为降低。

对于部分不能直接通过依存关系找到搭配词的调节参数或目标性能，考虑一下场景，如果有 2 个参数 *P*1 和 *P*2，它们在依存树上的关系为 *conj*(*P*1,*P*2)，*conj* 表示并列关系。其中 *P*1 与词 *A* 搭配，*P*2 没有搭配词，那么可以考虑把 *P*2 也与 *A* 搭配。例如：“ NEPE 推进剂的抗拉强度、伸长率均随环境压强的增加而增大 ”，增大同时修饰抗拉强度、伸长率。具体算法如下：

输入 依存关系 *G*
　　已经找到搭配词的实体集合{*PA*}
　　尚未找到搭配词的实体集合{*A*}

```
do{
    size1 = Size of {PA};
    For each entity p1 in {A}
        If P2 ∈ {PA}and conj(P1,P2) ∈ G
        {   Modifier = Modifier of P2 in {PA};
            Add (P1, Modifier) to {PA};
            Remove p1 from A; }
    size2 = Size of {PA};
}while(size1 != size2)
```

4　系统实现

基于上述分析，本文采用 Java 语言编写实现了一个化学类知识获取系统。系统使用了 ICTCLAS、Stanford Parser、Tregex 等开源工具，采用 Java JNI 实现对 ICTCLAS 系统的调用。系统主界面如图 4 所示。界面的最右边是用 Tregex 定义的各种实体的识别规则，中间是经过系统处理过后的文本。文本中被识别出的实体及实体所属类型都已用特殊颜色及文字标记。同时，可以查看各个语句的中间处理结果，如词性标注、句法分析树、知识抽取结果等。如 “ HMX 粒度降低，NEPE 推进剂燃速降低，压力指数降低。” 的最终处理结果如图 5 所示。

为了证明系统有效性，使用该系统对 200 条语句进行了实验验证，采用准确率、召回率对抽取的化学知识的各个部分进行评测，其结果如表 3 所示。

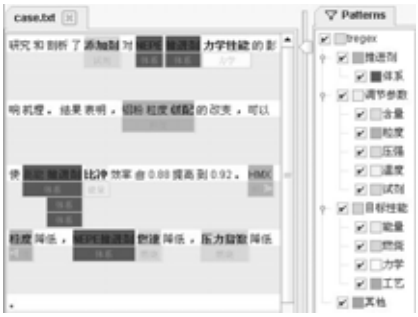

图 4　系统主界面

图 5　知识抽取结果

表 3　实验结果　(%)

模板项	准确率	召回率
实验条件	87.3	86.4
实验措施	80.0	76.5
实验结果	84.9	87.3
平均	84.0	83.4

5　结束语

本文利用自然语言处理技术，设计并实现了一个针对化学科技文献进行分析的知识获取系统，并对其关键技术进行了分析，系统集成了 ICTCLAS、Stanford Parser、Tregex 等开源工具，取得了良好的实际效果。同时，研究如何从科技文献中抽取知识，可以有效地缩短知识获取的周期，提高知识系统构造效率，具有很强的实际意义。

参考文献

[1] 蒋宏潮，王大亮，张德政. 基于领域本体的中医知识获取方法[J]. 计算机工程, 2008, 34(12): 16-18.

[2] Swanson D R. Online Search for Logically Related Non-interactive Medical Literatures: A Systematic Trial-and-Error Strategy[J]. Journal of American Society for Information Science, 1989, 40(5): 356-358.

[3] Levy R, Andrew G. Tregex and Tsurgeon: Tools for Querying and Manipulating Tree Data Structures[C]//Proc. of International Conference on Language Resources and Evaluation. Genoa, Italy: [s. n.], 2006.

[4] Klein D, Manning C D. Fast Exact Inference with a Factored Model for Natural Language Parsing[C]//Proc. of Advances in Neural Information Processing Systems. Vancouver, Canada: [s. n.], 2002: 3-10.

[5] Marneffe M C, Manning C D. The Stanford Typed Dependencies Representation[C]//Proc. of the Workshop on Cross-framework and Cross-domain Parser Evaluation. Manchester, UK: [s. n.], 2008: 1-8.

编辑　任吉慧