载《中国科技术语》，2009 年，第 3 期，Vol.11，No.3

单词型术语的结构自动分析

冯志伟

摘要：本文根据计算术语学的原理，使用有限状态转移网络对单词型术语进行自动分析。首先以英语术语为例，介绍了有限状态转移网络的基本原理和分析过程，然后，分别讨论了德语、法语和汉语的单词型术语的自动分析问题，最后讨论了分析结果的形式表示方法。

关键词：计算术语学，有限状态转移网络，自动词法分析，

Automatic analysis of single-word term structure

FENG Zhiwei

Abstract: Based on fundamental principle of computational terminology, the author automatically analyzes the single-word terms by the Finite State Transition Network (FSTN). In this paper, the basic principles and analysis process of FSTN are introduced in examples of English terms, then the automatic analysis of German, French and Chinese single-word terms are discussed, lastly the formal expression of analysis result is also discussed.

Key words: computational terminology, Finite State Transition Network (FSTN), automatic morphological analysis.

1998 年的计算语言学国际会议 COLING-ACL’98 上，组织了世界上第一次计算术语学的讨论会（First Workshop on Computational Terminology），这次讨论会首次使用的“计算术语学”这个学科名称。从此在术语学的研究中，明确地引进了自然语言处理（Natural Language Processing，简称 NLP）的方法和技术，出现了“计算术语学”[①]（computational terminology）这样的学科。

冯志伟在 1997 年的术语学与知识转播国际会议上发表的《日语形态的有限状态转移网络分析》[②]一文，是我国学者最早的研究计算术语学的论文，可是当时并没有引起我国术语学界的关注，在我国术语学研究中，几乎还没有其他的文章专门讨论过计算术语学的问题，本文根据计算术语学近年来的新发展，介绍计算术语学中单词型术语的结构自动分析方法，希望我国术语学工作者能够关注计算术语学这个新兴领域的研究，以推动我国术语学研究现代化的进程。

单词型术语结构分析的目的是让计算机知道单词型术语的结构，并且把与该术语有关的

---

① D. Bourigault, Ch. Jacquemin, Marie-Claude L’Homme, Recent Advances in Computational Terminology, John Benjamins Publishing Company, 2001.

② 冯志伟，日语形态的有限状态转移网络分析，《术语学与知识转播国际会议论文集》，1997 年，北京。

语言学信息（主要是形态信息）自动地加在该术语上，为术语进一步的自动处理做好准备。这是计算术语学最为基础的工作[③]。

单词型术语是由一个单词构成的，其中仅仅包含一个单词。一般地说，单词可以由词根、词缀和词尾构成，词根和词缀可以组成词干，词根后面也可以没有后缀而单独成为词干，在这种情况下，为了表述上的方便，我们就直接简单地把它叫做词干。这样，我们就可以用如下的“有限状态转移网络”（Finite State Transition Network，简称 FSTN）来表示一个单词的词法分析过程[④]。

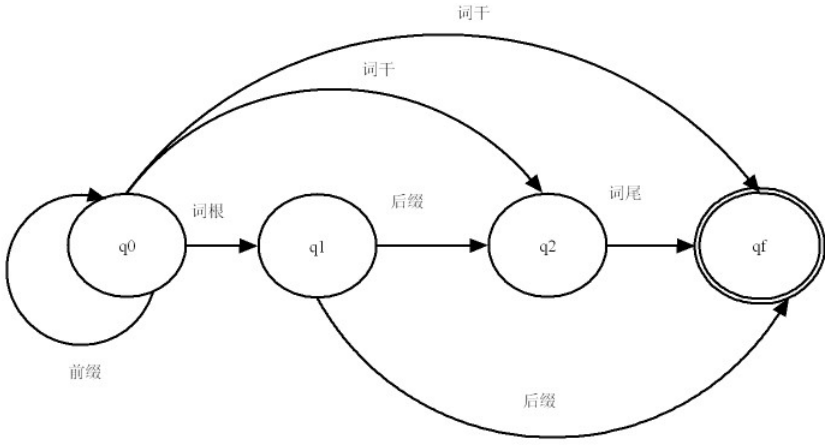

图 1　用限状态转移网络作词法分析

在图中，如果一个单词只包含词干（这时词干也就是词根），则其遍历过程是：q0→qf. 如英语的 form （“形式”）。

如果一个单词包含前缀、词干，则其遍历过程是：q0→q0→qf. 如英语的 reform （“改革”，re- 是前缀，form 是词干）。

如果一个单词包含词根、后缀，则其遍历过程是：q0→q1→qf. 如英语的 formation （“形成”，form 是词根，-ation 是后缀）。

如果一个单词包含前缀、词根、后缀，则其遍历过程是：q0→q0→q1→qf. 如英语的 reformation （“革新”，re- 是前缀，form 是词根，-ation 是后缀）。

如果一个单词包含词干、词尾，则其遍历过程是：q0→q2→qf, 如英语的 forms （form 是词干，-s 是词尾）。

如果一个单词包含前缀、词干、词尾，则其遍历过程是：q0→q0→q2→qf. 如英语的 formations （form 是词根，-ation 是后缀，-s 是词尾）。

如果一个单词包含前缀、词根、后缀、词尾，则其遍历过程是：q0→q0→q1→q2→qf. 如英语的 reformations （re- 是前缀，form 是词根，-ation 是后缀，-s 是词尾）。

由此可见，采用有限状态转移网络，可以非常清楚地描述屈折型语言单词的词法分析过程。

应该指出的是，在词根与后缀相连接时，有时会发生音变。例如，英语的词根 decide 与后缀 -ion 连接成 decision 时，-de- 变为 -s-, decide 中的元音 i 读为 [ai], 在 decision 中变为 [i]。但是，英语的词根 deny 与后缀-able 连接成 deniable 时，-y 在书写形式上变为-i，deny 中的 y 读为[ai]，在 deniable 中变为-i-之后，读音仍然为[ai]。对于这些复杂的音变问题，在用有限状态转移网络来进行单词的词法分析时，应该建立相应的音变规则来处理。

下面，我们进一步举例说明如何用有限状态转移网络来进行德语、法语单词型术语的结构分析。

[③]冯志伟，计算术语学，《术语标准准化与信息技术》，2008 年第 4 期，总第 52 期，4-9.

[④]冯志伟，形式语言理论，《计算机科学》，1979 年，第 1 期，p34-57。

德语屈折变化丰富，名词、形容词、冠词和指示词有性、数、格的变化，动词有变位形式。

德语中存在着大量的派生词，一个单词的词干加上前缀可构成许多新的单词。最常见的是由动词加前缀构成新的动词，由名词和形容词加后缀构成新的名词和形容词。

由动词加前缀构成的动词，如由 rufen （叫）加前缀 aus- 构成 ausrufen（呼喊），aus-是前缀，ruf 是词干，-en 是词尾，也可以用图 1 中的有限状态转移网络来进行词法分词，其遍历过程是：q0→q0→q2→qf.

由名词和形容词加后缀构成新的名词和形容词，如由名词 Kunst（艺术）加后缀 -ler 构成的名词 Kunstler（艺术家），由名词 Stern（星）加后缀-artig 构成的形容词 sternartig（星状的，stern 是词根，-artig 是后缀），由形容词 neu （新的）加后缀 -artig 构成的形容词 neuartig （新型的，neu 是词根，-artig 是后缀），也可以用图 1 中的有限状态转移网络来进行词法分析，其遍历过程是：q0→q1→qf.

在德语中还经常使用复合词，这种复合词由限定词加上基本词构成，基本词位于复合词的后部，，复合词的性和数由基本词决定，基本词还决定复合词的基本含义，限定词对基本词起修饰和限定的作用。例如，在 Intelligenztest（智力测验）这个复合词中，基本词是 Test（测验），限定词是 Intelligenz（智力），它进一步限定了基本词 Test 的确切含义。

图 1 中的有限状态转移网络不能分析这样的复合词，我们必须加以改进，使它在分析了复合词中的限定词之后，还能进一步分析复合词中的基本词。为此，我们从终极状态 qf 出发，再加一条指向初始状态 q0 的弧，并标以 #，使之从状态 qf 跳回 q0，再进一步分析复合词中的基本词。[⑤]如图 2 所示。

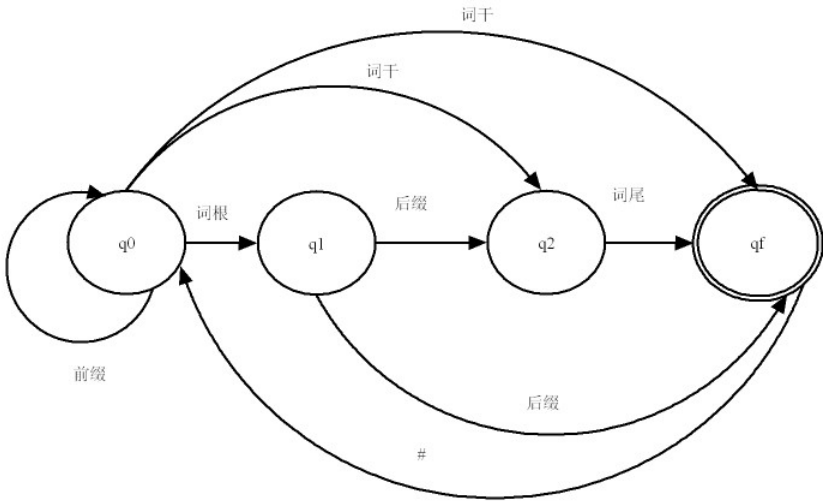

图 2　　可以分析复合词的有限状态转移网络

例如，Weltgeschichtlich （世界历史的）这个复合词，由名词 Welt（世界）加形容词 geschichtlich（历史的）复合而成。Welt 是限定词中的词干（这个限定词只有词干），geschicht 是基本词中的词根，-lich 是基本词中的形容词后缀。这个复合词可利用图 2 中的有限状态转移网络来进行词法分析，其遍历过程是：q0→qf→q0→q1→qf. 其中，在 qf 与 q0 之间，进行了一次返回初始状态的“跳跃”。

德语的术语很多是复合词，在许多复合词术语中，在组合成复合词的各个词之间，往往要加上 -s-, -es-, -en-, -n-, -er- 等字母，有的要去掉修饰词的词尾 -e-. 例如，术语 Lebenszeichen （生命象征）中，Leben （生命）与 Zeichen （象征）之间加上了 -s-; 在术语 Sinneszelle （感觉细胞）中， Sinn （感觉）与 Zelle （细胞）之间加上了 -es-; 在术

[⑤]冯志伟，德汉机器翻译 GCAT 系统，《语言现代化》，第 10 辑，1990 年，p139-162。

语 Nervenzelle （神经细胞）中，Nerv （神经）与 Zelle （细胞）之间加上了 -en-; 在 术语 Sonnenstrahl （阳光）中，Sonne （太阳）与 Strahl （光线）之间加上了 -n-; 在 术语 Kinderklinik （儿童诊所）中，Kind （儿童）与 Klinik （诊所）之间加上了 -er-; 在 术语 Erdgas （天然气）中，去掉了修饰词 Erde （地球）的词尾 -e. 这些问题，在词法分析时，要建立相应的音变规则来加以处理。

有时，德语的复合词术语可由两个以上的词组成，这只需在转移到终极状态 qf 之后，再往开始状态 q0 跳跃一次或几次就行了，仍然不难用图 2 中的有限状态转移网络来进行词法分析。但是，当复合词由若干个词组合而成的时候，切分时往往会出现莫棱两可、举棋不定的情况，这就需要在各种可能的切分情况中进行选择，确定一种正确的切分，排除不正确的切分。

例如，Bauerlaubnisse （建筑许可）这个复合词术语，在德语的机器词典中，存有 Bauer （das Bauer, 中性名词，鸟笼），Bau （动词 bauen 的词干，建筑），Bauer （der Bauer, 阳性名词，农民），Erlaub （动词 erlauben 的词干，许可），Erlaubnis （die Erlaubnis, 阴性名词，许可），Laub （das Laub, 中性名词，树叶），Nisse （die Nisse, 阴性名词，虱子卵），-se（名词词尾）等语素，因此，可能存在的切分情况有三种：

① Bau + erlaubnis + se

② Bauer + laub + nisse

③ Bau + erlaub + nisse

为了在这三种可能的切分中选择出正确的切分，我们可检查每种切分在语义上的相容性。

在①中，其语义的组合情况是：

建筑 + 许可 + 名词词尾

切分出来的三个部分的语义是相容的。

在②中，其语义的组合情况是：

鸟笼 + 树叶 + 虱子卵

或 农民 + 树叶 + 虱子卵

切分出来的三个部分在语义上不相容。

在③中，其语义的组合情况是：

建筑 + 许可 + 虱子卵

切分出来的三个部分在语义上也不相容。

所以，我们选择语义上相容的第①种切分，排除语义上不相容的第②③两种切分，并确定这个复合词的词义为“建筑许可”。

法语是从拉丁语演变而来的。与拉丁语相比，法语的词形屈折已大大简化，名词没有格的变化，性和数主要通过名词前的冠词、限定词来区别，动词有变位形式，形容词也有性与数的变化，少数形式还比较复杂；法语的词从结构上也可以分为前缀、词干、词根、后缀、词尾几部分，名词、形容词、动词都可以通过加前缀或后缀来派生。

由词干加前缀构成的词，如 contrevent （风窗，contre- 是前缀，vent 是词干），extrafin （纤细，extra- 是前缀，fin 是词干），可用图 1 中的有限状态转移网络来分析，其遍历过程是：q0→q0→qf.

由词根加后缀构成的词，如 mouvement （运动，mouve 是词根，-ment 是后缀），durable （持久，dur 是词根，-able 是后缀），可用图 1 中的有限状态转移网络来分析，其遍历过程是：q0→q1→qf.

由词根加前缀和后缀构成的词，如 surproduction （生产过剩，sur- 是前缀，product 是词根，-ion 是后缀，telespectateur （电视观众，tele-是前缀，spectat 是词根，-eur 是后缀），

也可用图 1 中的有限状态转移网络来分析，其遍历过程是：q0→q0→q1→qf.

在具体的法语词法分析中，图 1 中的有限状态转移网络显得过于笼统和简单。

在法语中，当名词后缀是 -ance, -ation, -ade, -ment 时，其词根一般是动词词根。例如，名词 obeissance （服从）的词根是动词词根 obeiss-,名词 creation （创造）的词根是动词词根 cre-, 名词 promenade （散步）的词根是动词词根 promen-，名词 fabrication （生产）的词根是动词词根 fabric- （fabriqu- 的音变形式）。

当形容词后缀是 -able, -if 时，其词根一般也是动词词根。例如，形容词 navigable （可通航的）的词根是动词词根 navig-, 形容词 pensif （沉思的）的词根是动词词根 pens-。

当名词后缀是 -ité, -esse 时，其词根一般是形容词词根，例如，名词 fidelité （忠实）的词根是形容词词根 fidel-, 名词 souplesse （柔软）的词根是形容词词根 soupl-。

由形容词词根构成名词时，有时还会发生音变。例如，名词 sottise （笨拙）由形容词词根 sot-（愚笨）和后缀 -ise 构成，而在它们之间，要加辅音字母 -t-。

基于这些情况，在对法语的单词型术语进行结构分析时，我们有必要区分构成合成词的词根是动词词根还是形容词词根，从而更加细致地描述名词和形容词的词法分析过程。

另外，分析的方向也不一定总是从左到右，也可以从右到左，先分析词尾、后缀，再分析词根，最后才分析前缀。

为了处理法语中这些复杂的语言现象，我在法-汉机器翻译系统 FCAT 的研制中，曾经提出了如图 3 所示的有限状态转移网络[⑥]。

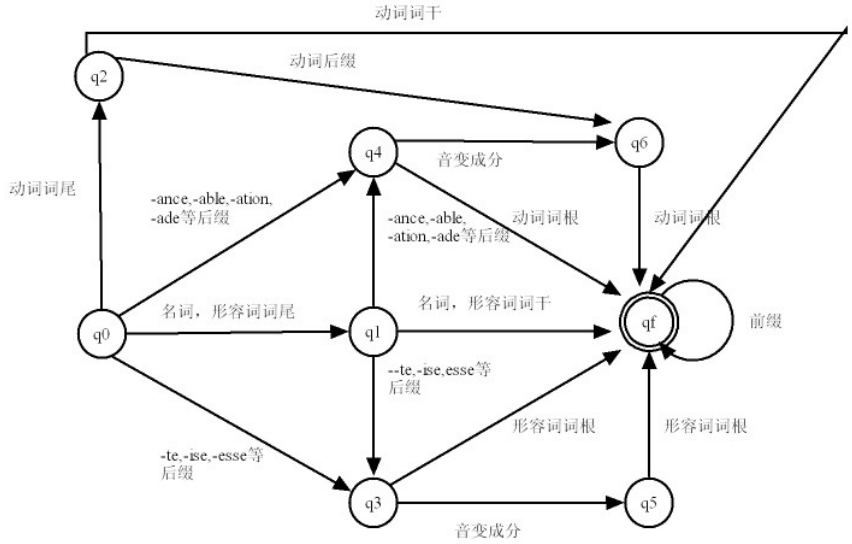

图 3　法语词法分析的有限状态转移网络

这样，词根为动词词根的名词，如果没有音变成分，则其遍历过程是 q0→q4→qf, 例如，法语的 creation, 先分析后缀 -ation, 后分析动词词根 cre-。如果有音变成分，则其遍历过程是 q0→q4→q6→qf. 例如，法语的 fabrication, 先分析后缀 -ation, 再把音变成分 -c- 变为 -qu-, 再分析动词词根 fabriqu-。

词根为形容词词根的名词，如果没有音变成分，则其遍历过程是 q0→q3→q5。例如，法语的 souplesse，先分析后缀 -esse，再分析形容词词根 soupl。如果有音变成分，遍历过程是 q0→q3→q5→qf. 例如，法语的 sottise, 先分析后缀 -ise，再分析音变成分 -t-，最后分析形容词词根 sot.

法语的名词、形容词、动词都有词尾屈折变化。如果名词、形容词有屈折变化词尾，则

[⑥]冯志伟，法汉机器翻译 FCAT 系统，《情报科学》，1987 年，第 4 期，p19-27。

首先还要分析词尾，再分析后缀和词根。无音变时，其遍历过程是 q0→q1→q3→qf 或 q0→q1→q4→qf, 有音变时，其遍历过程是 q0→q1→q3→q5→qf 或 q0→q1→q4→q6→qf. 如果动词有屈折变化词尾，则首先分析动词词尾，再分析动词词干，其遍历过程是 q0→q2→qf。

如果名词、形容词、动词还有前缀，则还需在终极状态 qf 分析了前缀之后，再回到这个终极状态 qf. 例如，法语的 prefabrication （预制），其遍历过程是 q0→q4→q6→qf→qf. 首先分析后缀 -ation, 再把音变成分 -c- 改变为 -qu-, 再分析动词词根 fabriqu-, 最后再分析前缀 pre-。

汉语单词型术语的结构比较简单，也可以使用图 1 中的有限状态转移网络来分析。

--只有词干的单词型术语：例如，“速度、能量”，遍历过程是：q0→qf。

--带前缀的单词型术语：例如，“超导体、非金属”，其中“超，非”是前缀，遍历过程是：q0→q0→qf。

--带后缀的单词型术语：例如，“电气化、绝缘体”，其中“化、体”是后缀，遍历过程是：q0→q1→qf。

--带前缀和后缀的单词型术语：例如，“非周期性，反铁氧体”，其中的“非、反”是前缀，“性、体”是后缀，遍历过程是：q0→q0→q1→qf。

汉语的语缀不仅可以附加在词根或单词上，还可以附加在词组上。例如，“非线性规划”，中的附加前缀“非”，“同素异形体”中的附加后缀“体”。对于这样的术语，我们可以按照德语单词型术语中复合词的结构分析方法来处理，使用图 2 中的有限状态转移网络来进行分析。使用这样的有限状态转移网络，“非线性规划”的遍历过程是：q0→q0→qf→q0→qf，“同素异形体”的遍历过程是：q0→qf→q0→q1→qf。

根据有限状态转移网络的原理，单词型术语经过自动分析之后，就可以输出与该单词型术语有关的形态信息。这些形态信息可以形式化地加以表示。例如，英语中以 beauty 为词干的单词型术语 beautified 经过形态分析之后，可以形式化地表示如下：

beatified:　　<<<*>N + ify>V + ed>A

其中，*表示 beauty，N 表示它是一个名词，加上-ify 之后，变成 beautify，是一个动词（V），再加上-ed 之后，变成 beautified，是一个形容词（A）。

同样，我们得到的其他单词型术语的分析结果可以形式化地表示如下[⑦]：

beautification:　　<<<*>N + ify>V + cation>N

beautifier:　　<<<*>N + ify>V + er>N

beautiful:　　<<*>N + ful>A

unbeautified　　<un# <<<*>N + ify>V + ed>A>A

unbeautiful　　<un# <<<*>N + ful>A>A

根据前面 beautified 的例子，读者不难理解到这些分析结果的含义。

单词型术语的自动分析是对于单词型术语中的各个组成成分进行自动分析，在自然语言处理中属于自动词法分析（automatic morphological analysis）的范围。而词组型术语的自动分析，就属于自动句法分析（syntactic parsing）的范畴了。我们将另文讨论词组型术语的自动分析问题。

---

⑦ 参看 Christian Jacquemin, Spotting and Discovering Terms through Natural Language Processing, p20, The MIT Press, 2001.