北京大学

博士研究生学位论文

汉英机器翻译若干关键技术研究

姓　　名：刘　　群

学　　号：19908835

院　　系：信息科学技术学院

专　　业：计算机软件与理论

研究方向：计算语言学

导　　师：俞 士 汶 教授

二〇〇四 年 五 月

Researches into Some Key Aspects of Chinese-English Machine Translation

Dissertation Submitted to

Peking University

in partial fulfillment of the requirement

for the degree of

Doctor of Natural Science

By

LIU Qun

(Computer Software and Theory)

School of Electronics Engineering and Computer Science

Dissertation Supervisor： Professor YU Shiwen

MAY, 2004

版权声明

汉英机器翻译中若干关键技术研究

摘　要

虽然机器翻译离人们的希望还有很大的距离，不过近年来统计机器翻译技术的一些进展使很多研究者相信，在现有的计算条件下通过研究方法的改进，机器翻译的水平还有较大的提高空间。作者认为，充分利用人类专家知识库、基于大规模语料库获取语言和翻译知识、建立反映语言深层结构对应关系的统计翻译模型是通向高质量机器翻译的有效途径。本文的研究工作就反映作者在这个方向上进行的一系列努力。

本文主要围绕汉英机器翻译中的一些关键技术展开研究。具体来说，本文在以下方面做出了有创新性的工作：

1．提出了一种基于层叠隐马尔可夫模型的汉语词法分析算法。这个算法由多个层叠的隐马尔可夫模型构成，粗切分采用基于 N 最短路径的算法，简单未定义词和复合未定义词采用基于角色的隐马尔可夫模型识别新词，并采用基于角色的词语生成模型估计未定义词的概率；细切分采用词汇化的隐马尔可夫模型；词性标注采用基于词性的隐马尔可夫模型；多种模型紧密结合，下层模型不仅提供多个最好的分析结果供高层模型使用，而且也给出了这些结果的概率。模型之间环环相扣，互为补充，最终达到整体结果的最优化，同时保持算法的高效率（线性时间复杂度）。

2．提出了一种基于《知网》的词汇语义相似度计算模型。这种方法充分利用了《知网》中所包含的丰富的人类语言学知识，直接计算两个词语的语义相似度，而无需通过大规模语料库的训练，方法简单有效。这种方法可广泛用于词义排歧、基于实例的机器翻译等多种领域。

3．提出了一种高效的双语短语对齐搜索算法。这种算法的主要优点是可以尽可能避免词语对齐错误给短语对齐带来的干扰，使得短语对齐的正确率和召回率比词语对齐的相应指标都要高出很多，效果很好。算法采用柱形搜索策略，时间消耗随着句子长度线性增长，效率也非常高。

4．定义了一种可以刻画两种语言深层句法结构对应关系的短语结构转换模板，并给出了从双语短语对齐的语料库中抽取这种模板的算法。对实验结果的初步分析表明，从一个八千句子对的短语对齐语料库中抽取出来的模板，已经可以覆盖各种常见的汉英句法结构的转换模式。

5．提出了一种微引擎流水线机器翻译系统结构。在这种结构下，整个机器翻译过程被分解成若干个串行的阶段，每个阶段可以有若干个功能相似的部件（微引擎）同时工作。通过添加和删除微引擎以及调整流水线的结构很容易实现各种机器翻译构件的协调工作，而无需修改系统的总体翻译算法和数据结构，有利于提高机器翻译系统的开发效率以及尝试新的机器翻译方法。文中介绍了一个基于这种结构实现的面向新闻领域的汉英机器翻译系统，并给出了实验结果。

关键词： 汉英机器翻译　汉语词法分析 词义相似度计算 双语短语对齐
翻译模板的自动抽取　多引擎机器翻译方法

Researches into Some Key Aspects of Chinese-English Machine Translation

Abstract

Current machine translation quality is far from user's expectation. However, recent advances on statistical machine translation make many researchers believe that there is a fairly big space to improve the quality of machine translation by improving the approaches we used. The author suggest that adequate utilization of human-made knowledge bases, language and translation knowledge acquisition from large scale corpus and construction of statistical translation model which can capture the correspondence between the deep structures of the source and target languages, are the proper way to achieve a high quality machine translation. This paper presents some researches that have been done by the author following this direction.

All the researches present in this paper are on some key technologies of Chinese-English machine translation. Specifically, these researches include:

1. An algorithm for Chinese morphological analysis based on Cascaded Hidden Markov Model (Cascaded HMM) is proposed. In this model, multiple layers of HMMs are used to resolve various morphological problems separately. The N-shortest paths approach is used in rough segmentation. Two layers of role-based HMMs are used to recognize simple unknown words and complex unknown words, and a role-based word generation model is used to estimate the probabilities of unknown words. A lexicalized HMM is used in fine segmentation. A POS HMM is used in part-of-speech tagging. These models in different layers are coupled tightly. Low level HMMs not only provide candidates to high level HMMs, but also provide the probabilities of these candidates. All the models are integrated in a whole framework to achieve a best result, while the time cost of the algorithm is still linear.
2. An algorithm for semantic similarity computing between Chinese words based on Hownet is proposed. This algorithm utilizes the rich human knowledge in Hownet to compute the semantic similarity between two Chinese words directly, without any data training from large corpus. This approach is efficient and simple, and can be used in word sense disambiguation, example-based machine translation, and some other areas.
3. A efficient search algorithm for phrase alignment in parallel corpus by tree-tree mapping is proposed. This algorithm can avoid the spreading of word alignment errors in to phrase alignment. So the precision and recall of phrase alignment are much better than those of word alignment. A beam search strategy is used in the algorithm, and the time cost is linear in terms of the length of sentence.
4. A Phrase Structure Transduction Template (PSTT) is defined, which can capture the correspondence of the syntax structures of source and target languages, and a extraction algorithm of PSTTs from phrase aligned parallel corpus is given. A primary analysis of the experiment results show that the PSTTs extracted from a phrase aligned corpus containing 8009 sentence pairs can cover most of the frequently used transform patterns between Chinese and English syntax structure.

5. A micro-engine pipeline machine translation architecture is proposed. In such an architecture several components with different algorithms are used in each phases of the system. New approach can be test under the architecture by adding a new micro-engine or adjusting the pipe structure, without changing the overall algorithm of the system. This architecture is propitious to the test of new machine translation approach and to improve the efficiency of development of machine translation system. An implementation of this architecture in a news-oriented Chinese-English machine translation system and the experiment results are given.

Keywords: Chinese-English Machine Translation, Chinese Morphological Analysis, Lexical Semantic Similarity Computing, Phrase Alignment of Parallel Corpus, Translation Template Extraction, Multi-Engine Machine Translation

目 录

第1章 引言

1.1 研究的意义

自从 1949 年 Warren Weaver 发表《翻译》备忘录，正式提出机器翻译的思想以来，到现在已经经过了半个世纪的时间。虽然机器翻译的现状离人们的期望和市场的需求都还有相当大的距离，还远远不能满足人们的要求，不过人们对机器翻译研究的热情依然很高。这一方面是因为机器翻译的巨大需求和应用前景在不断激励着人们从事这方面的研究工作，另一方面，仅从学术角度看，机器翻译也是一个非常有意义的研究课题，其复杂性、挑战性和高难度的特点对研究者而言充满了魅力。机器翻译的研究，大大加深了人们对于语言、知识、智能等问题的了解，促进了相关学科的发展。作者认为，对全自动高质量机器翻译的不懈追求，正是计算语言学研究的终极目标之一和不竭动力的源泉。

汉语是我们的母语，是数千年中华文化的主要载体，同时又是一种非常独特的语言。目前，汉字的输入、输出等方面的问题已基本解决，而汉语更深层次的处理，如词法、句法、语义分析、机器翻译等，和世界上其他一些主要语种的处理技术相比，还有一定的差距。加强这方面的研究工作对中国的自然语言处理研究者来说更是责无旁贷的。

1.2 历史与现状

最早的机器翻译是建立在简单的单词对译、词频统计和词序变化的基础上。当人们认识到这种方法的局限性后，开始加强了对自然语言理解的研究。伴随着人工智能研究的发展和乔姆斯基语言学的大行其道，规则方法成为了机器翻译研究的主流。研究者们发现，在一些小规模应用或演示环境中表现出色的规则方法，在真正的大规模应用中却表现得非常糟糕。于是，从 1990 年代初开始，统计方法又被重新引入到自然语言处理研究中，在机器翻译方面，IBM 公司提出了著名的基于信源信道模型的统计机器翻译方法。在这以后的一段时间内，尽管统计方法在自然语言处理的很多领域都获得了成果，但对于机器翻译来说，统计方法并没有马上建立起优势地位。由于机器翻译问题本身的复杂性和计算机运行能力的限制，在很长一段时间内，很少有人能够重复 IBM 的统计机器翻译工作，以至于很多人对统计方法在机器翻译中的效果产生了怀疑。不过近年来，在一批研究者的不断努力下，也得益于计算能力的普遍提高，统计机器翻译终于开始表现出明显的优势并受到了普遍的重视。在最近的一些机器翻译评测中，基于统计方法的机器翻译系统都取得了很好的成绩。

不过，作者认为，目前的统计机器翻译方法也还存在一个明显的缺陷：现有的统计机器翻译基本上都是基于语言表层结构的，也就是说，现有的翻译模型所刻画的仅仅是两种语言的表层结构之间的对应关系。这种关系虽然可以面面俱到（通过大规模语料库的统计），但毕竟是非常肤浅的，不可能抓住两种语言之间本质性的对译关系。因此，从理论上说，在这种基于浅层关系的统计翻译模型下，翻译的质量必然受到根本性的局限。另外，由于生成的目标语言句子并不是从相应的深层结构中推导出来，而且也无法产生真正符合目标语言语法的句子。现在也有人对此提出了一些改进的办法，某些系统也开始引入了一些浅层或深层的的句法分析，不过总体效果并不理想，而且实验规模过小，不具有实质意义。

已有的实验显示，仅仅在表层或者浅层分析的基础上引入统计方法，就使得机器翻译的

质量赶上并超过了原有的基于规则的机器翻译系统（最近两次的 NIST 评测可以说明这一点）。可以想象，如果能够在反映语言本质特点的深层结构的基础上建立统计模型，完全可能把机器翻译的水平向前推进一大步。不过，要在语言的深层结构基础上建立统计模型，也会带来很多问题，如数据稀疏程度的增大和算法复杂度的增高。不过这正是值得研究者们努力的研究方向。

对于汉英机器翻译来说，汉语的分析问题也是一个无法回避的难题。由于汉语词缺乏形态变化，词与词之间没有间隔，汉语的字、词、短语、句子之间缺乏明确的界限，语法规则的限制较少，因此汉语分析的困难程度也较高。研究者们对汉语的分析问题，特别是词法分析问题，倾注了大量的精力，不过到目前为止，汉语词法分析的效果仍然很不理想。

1.3 研究的目标

基于目前汉英机器翻译的研究现状，作者希望，最终能够研制出一种基于大规模语料库的汉英机器翻译系统，这个系统的最重要的特点是能够建立反映汉语和英语的深层结构对应关系的统计模型。

考虑到汉英机器翻译问题的复杂性，这是一个很宏伟、也很长远的目标。具体来说难点主要体现在以下几个方面：

1. 汉语分析：由于汉语是孤立语，汉语的语素、词、短语之间没有明显的界限，汉语词语没有明显的屈折变化，也没有太多反映语法功能的词缀，这使得汉语的分析比英语、日语等语言的分析来说都更为困难。
2. 大规模语言资源的获取和加工：大规模的语言资源是机器翻译知识的源泉，没有丰富的语言资源和有效的加工技术，不可能有真正高质量的机器翻译。目前这方面的难点主要包括大规模高质量的双语词典的获取、双语语料库的词语对齐和短语对齐技术等。
3. 基于深层句法结构的统计翻译模型：现有的统计机器翻译技术大部分都是建立在句子的表层或者浅层结构基础上的，少量基于深层结构的统计模型，目前还不够成熟，效果也不理想。

围绕上述目标，本文在现有的一个完整的基于规则的机器翻译系统[刘群 1997][Liu 1998]的基础上，开展了一系列的研究工作。这些工作具体包括：

1. 基于统计的汉语词法分析；
2. 基于统计的汉语句法分析；
3. 汉语词语的相似度计算；
4. 汉英双语词典的获取；
5. 汉英双语语料库的获取与对齐；
6. 短语结构转换模板的抽取；
7. 基于短语结构转换模板的统计翻译模型；
8. 基于实例的机器翻译；
9. 多引擎机器翻译系统结构；
10. 机器翻译自动评价技术研究。

以上这些研究工作，有的已经基本完成，有的正在进行之中，有的刚刚启动。其中一些是作者独立进行的，还有一些是作者与课题组其他成员合作进行、或者完全由课题组其他成员进行的。目前我们已经完成的这些工作离上述的最终目标虽然还有一定距离，但已经取得了很大的进展，这些进展使我们向着这个最终目标跨进了一大步。

1.4 本文的工作和论文的组织

本文只涉及到上述工作中已经完成的、由作者独立进行或者与课题组其他成员合作进行的那部分工作。其中，与课题组其他成员合作进行的工作都在“致谢”中进行了说明。

具体来说，本文的工作包括：

1．基于层叠隐马尔可夫模型（Cascaded HMM）的汉语词法分析；

2．基于《知网》的汉语词汇语义相似度计算；

3．一种双语短语结构对齐的搜索算法

4．短语结构转换模板的提取与应用；

5．微引擎流水线机器翻译系统结构以及基于这种结构实现的一个面向新闻领域的汉英机器翻译系统。

本文第一章（本章）是引言，介绍了本文工作的意义、现状、目标等研究背景。第二章是综述，介绍了传统基于规则方法以外的一些机器翻译方法，以及机器翻译方法分类。第三章介绍了基于层叠隐马尔可夫模型的汉语词法分析方法。基于该模型开发的汉语词法分析系统取得了很好的效果。第四章介绍了一种基于《知网》的词汇语义相似度计算方法，这种方法仅使用《知网》作为知识库，不需要用语料库进行训练，既可以方便地计算出两个汉语词语之间的语义相似度，为基于实例的机器翻译中的相似例句查询提供了基础。第五章介绍了一种双语短语结构对齐的柱形搜索算法，这种算法不仅效率很高（运行时间与句子长度呈线形关系），而且可以在很大程度上避免错误的词语对齐对短语对齐造成的干扰。第六章介绍了一种短语结构转换模板的定义以及从双语短语结构对齐的语料库中自动抽取这种模板的算法，并通过实验对自动抽取得到的短语结构转换模板进行了详细的分析。第七章介绍了微引擎流水线机器翻译系统结构，以及基于这种结构开发的一个面向新闻领域的汉英机器翻译系统。微引擎流水线是一种细粒度的多引擎机器翻译技术，可以实现各种机器翻译算法部件级的并行与合作。文章的最后给出了总结和今后工作的一些设想。

第2章　机器翻译方法综述

机器翻译经过 50 多年的发展，产生了很多不同的方法。比如人们常常提到基于规则的方法、基于统计的方法、同基于规则相结合的方法、基于实例的方法、中间语言方法、转换方法、基于知识的方法，诸如此类，等等等等。这些方法种类繁多，都有各自的优缺点。但这些方法往往是从不同角度、不同层面来说的，互相之间并不一定具备可比性。人们在初次接触机器翻译的时候，往往会被如此众多的方法所迷惑，如坠五里雾中，不容易理解这些方法之间内在的区别与联系。

在这一章中，本文将从范式（Paradigm）和分类这两个角度对机器翻译方法进行一个初步的梳理，不仅要对各种机器翻译方法作一个大致的介绍，而且试图刻画出它们之间的联系与区别。

所谓的范式，指的是对某些具体的机器翻译实现方法的一种抽象和归纳。范式往往要对机器翻译方法的某些方面做出明确的规定，而对另外一些方面可以没有明确的要求。但由于范式往往都有一些典型的实现方法或具体的系统，所以即使对那些没有明确要求的方面，人们往往也都有一些缺省的理解。比如说，基于转换的方法，作为一种范式，本身并没有规定采用规则方法还是统计方法，但人们谈到这种方法的时候，往往都把它理解成一种基于规则的方法。这是由于这种方法出现的时候，还没有出现现在意义上的统计机器翻译方法。而且一些典型的基于转换的系统，也都是采用规则方法实现的。另外，不同的范式往往对机器翻译方法的不同方面和不同层次做出规定，所以范式之间往往不具有可比性。可以这么说，通过范式对机器翻译方法进行总结，就是人们常说的抓典型的方法，或解剖麻雀的方法。通过对范式的研究，可以起到解剖麻雀的作用，有助于对机器翻译的实现技术进行比较全面和深入的了解。

但范式并不等同于分类。科学的分类往往要求执行统一的分类原则，分类的结果要求满足完整性和互斥性。而范式的定义并没有统一的标准和原则，因而也不满足完整性和互斥性。也就是说，可以有一些具体的机器翻译实现方法，同时满足两种或两种以上的范式，而有些机器翻译的实现方法可能不符合现有的任何一种范式。而分类则不然，一旦分类的原则确定下来，任何一种具体的机器翻译实现方法必定落入一种且只能落入一个类别当中。根据分类原则的不同，可以有各种不同的分类结果。通过对各种机器翻译方法进行分类研究，将各种机器翻译方法进行多角度多层次的比较，可以对机器翻译的各种方法之间的区别和联系有比较清楚的了解。

本章先介绍一些机器翻译中常见的一些范式，然后试图用分类的方法对这些方法进行总结和归纳。

2.1 机器翻译的范式

机器翻译的范式很多，常见的一些包括：

1. 直接翻译方法：早期的不经过句法分析直接进行词语翻译和词序调整的方法；
2. 基于转换的方法：基于源语言和目标语言的深层表示形式进行转换的方法，典型的转换方法要求独立分析，独立生成；注意，这里的深层表示既可以是句法表示，也可以是语义表示；
3. 基于中间语言的方法：利用独立于具体语言的某种中间表示形式（称为中间语言）

实现两种语言之间的翻译的方法；

4. 基于语言学的方法：以语言学知识为推理的基础，在对源语言的句法语义进行深入的分析和理解的基础上进行翻译的方法；
5. 基于知识的方法：利用人工智能中知识表示、知识推理技术进行机器翻译的方法；
6. 基于平行语法的方法：通过为源语言和目标语言构造一套平行的语法体系，在分析的同时完成翻译的方法；
7. 基于实例的方法：通过将源语言句子和实例库中已有的句子进行类比得到译文的方法；严格地说，基于实例的方法不是一种机器翻译的典型范式，因为其中的不确定性因素太多，各种具体的基于实例的机器翻译方法实施方案相差较大；
8. 基于信源信道模型的统计机器翻译方法：将翻译理解为信息传输的过程，通过对语言模型和翻译模型的估计来求解最佳译文的方法；
9. 基于最大熵模型的统计机器翻译方法：直接将翻译的概率分解为一组特征函数的乘积，通过最大熵模型进行参数估计的方法。

一些传统的机器翻译方法早已广为人知，如直接翻译方法、基于转换的方法、基于中间语言的方法等等，这里不再详细介绍。本文主要介绍一些比较新的或者与本文关系较为密切的几种方法，包括基于平行语法的方法、基于实例的方法和基于统计的方法等等。

2.2 基于平行语法的机器翻译方法

这种方法的基本思想是，用一套双语平行的语法模型，即两组相互对应的规则，同时生成两种语言的句子，在对源语言句子进行理解的同时，就可以得到对应的目标语言句子的生成过程。

这种方法的基本特点是：有明确的规则形式；源语言规则和目标语言规则一一对应；如果采用概率形式，那么源语言与目标语言服从相同的概率分布，即对应的规则在两种语言中出现的概率相同；对于两种语言的转换过程不使用概率模型进行描述。

以下分别介绍这种方法中几种具体的形式。

2.2.1 Alshawi 的基于加权中心词转录机的统计机器翻译方法

有限状态转录机（Finite-State Transducer）和有限状态识别器（Finite-State Recognizer）是有限状态自动机（Finite-State Automata）的两种基本形式。其主要区别在于有限状态转录机在识别的过程中同时可以产生一个输出，其每一条边上面同时有输入符号和输出符号两个标记，而有限状态识别器只能识别，不能输出，其每一条边上只有一个输入符号标记。

中心词转录机（Head Transducer）是对有限状态转录机的一种改进。对于中心词转录机，识别的过程不是自左向右进行，而是从中心词开始向两边执行。所以在每条边上，除了输入输出信息外，还有语序调整的信息，用两个整数表示。下图是一个能够将任意 a、b 组成的串逆向输出的一个 HT 的示意图：

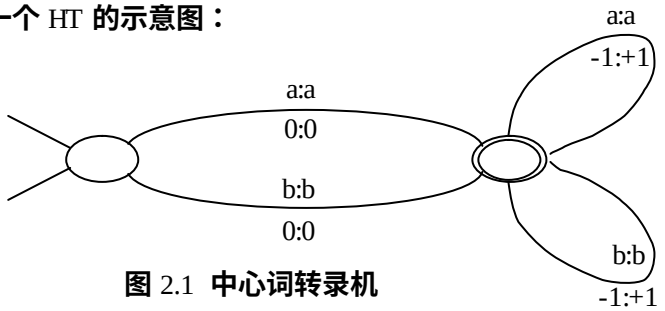


图 2.1 中心词转录机

基于加权中心词转录机（Weighted Head Transducer）的统计机器翻译方法是由 AT&T 实验室的 Alshawi 等人提出的，用于 AT&T 的语音机器翻译系统。该系统由语音识别、机器翻译、语音合成三部分组成。其中机器翻译系统的总体工作流程如下图所示：

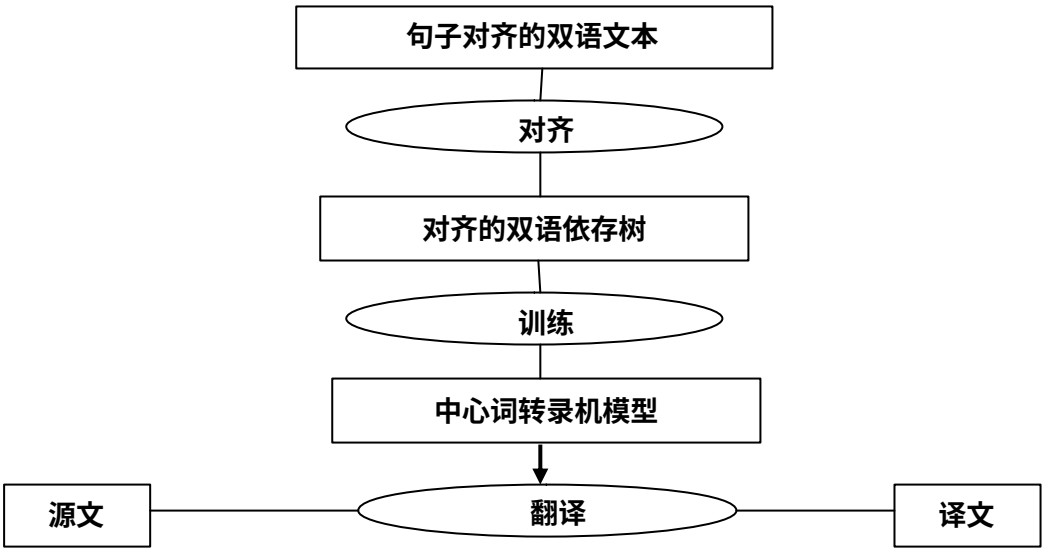


图 2.2 基于中心词转录机的机器翻译系统工作流程

在加权中心词转录机模型中，中心词转录机是唯一的知识表示方法，所有的机器翻译知识，包括词典，都表示为一个带概率的 Head Transducer 的集合。知识获取的过程是全自动的，从语料库中训练得到，但获取的结果（就是中心词转录机）很直观，可以由人进行调整。中心词转录机的表示是完全基于词的，不采用任何词法、句法或语义标记。

整个知识获取的过程实际上就是一个双语语料库结构对齐的过程。句子的结构用依存树表示（但依存关系不作任何标记）。他们经过一番公式推导，把一个完整的双语语料库的分析树构造并对齐的过程转化成了一个数学问题的求解过程。这个过程可用一个算法高效实现。得到对齐的依存树后，很容易就训练出一组带概率的中心词转录机，也就得到了一个机器翻译系统。不过要说明的是，通过这种纯统计方法得到的依存树，与语言学意义上的依存树并不符合，而且相差甚远。

这种方法的主要特点是：1.训练可以全自动进行，效率很高，由一个双语句子对齐的语料库可以很快训练出一个机器翻译系统；2.不使用任何人为定义的语言学标记（如词性、短语类、语义类等等），无需任何语言学知识；3.训练得到的参数包含了句子的深层结构信息，这一点比 IBM 的统计语言模型更好。

这种方法比较适合于语音翻译这种领域比较受限，词汇集较小的场合。

2.2.2 吴德恺的反向转录语法

反向转录语法（Inversion Transduction Grammar，ITG)是香港科技大学吴德恺（Dekai Wu）提出的一种供机器翻译使用的语法形式[Wu 1997]。

这种语法的特点是，源语言和目标语言共用一套规则系统。

具体来说，ITG 规则有三种形式：

A ？ [B C]

A ？ < B C >

A ? x/y

其中 A，B，C **都是非终结符，**x，y **是终结符。而且** B，C，x，y **都可以是空（用** e **表示）。**

对于源语言来说，这三条规则产生的串分别是：

BC BC x

对于目标语言来说，这三条规则产生的串分别是：

BC CB y

可以看到，第三条规则主要用于产生两种语言的词语，第一条规则和第二条规则的区别在于，前者产生两个串语序相同，后者产生的串语序相反。例如，两个互为翻译的汉语和英语句子分别是：

比赛星期三开始。

The game will start on Wednesday.

采用 ITG **分析后得到的句法树就是：**

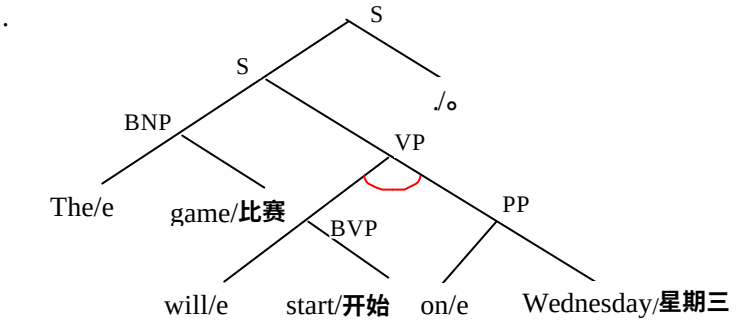


图 2.3 **反向转录语法（**ITG**）产生的双语句法树**

其中，VP **结点上的弧线标记表示该结点对应的汉语句子中两个子结点的顺序需要交换。**

通过双语对齐的语料库对这种形式的规则进行训练就可以直接用来做机器翻译。

吕雅娟[Lü 2001, 2002]**提出了一种基于** ITG **建立统计模型的方法，并实现了一个小规模（**2000 **个例句）的英汉机器翻译系统，取得了较好的实验结果。这个系统利用的英语的单语分析器和英汉双语词对齐的结果来获取** ITG**。系统结构如下图所示：**

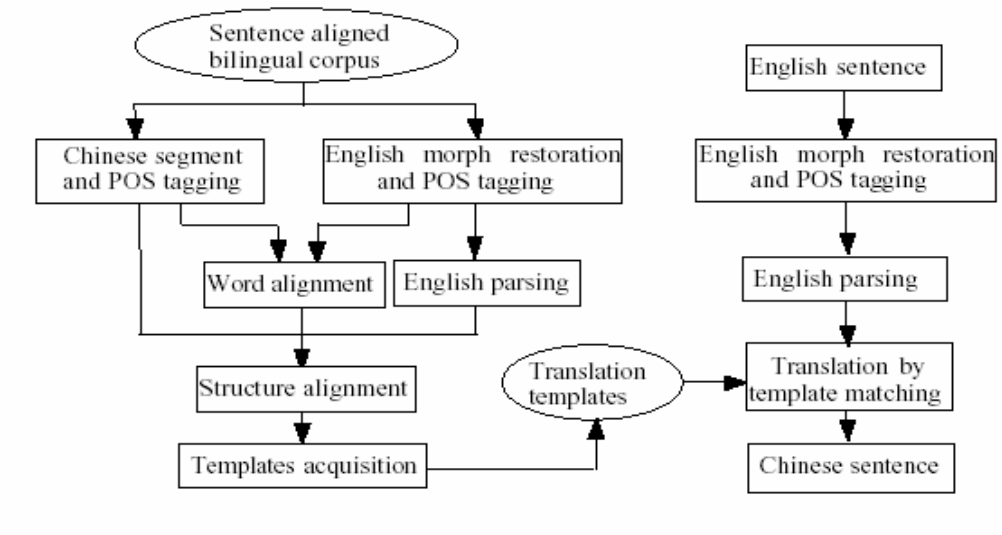


图 2.4 **基于** ITG **的机器翻译系统工作流程**

2.2.3 Takeda 的基于模式的机器翻译上下文无关语法

[Takeda 1996]提出了基于模式的机器翻译上下文无关语法（Pattern-based CFG for MT）。该模型对于翻译模板定义如下：

1. 每个翻译模板由一个源语言上下文无关规则和一个目标语言上下文无关规则（这两个规则称为翻译模板的骨架），以及对这两个规则的中心词约束和链接约束构成；
2. 中心词约束：对于上下文无关语法规则中右部（子结点）的每个非终结符，可以指定其中心词；对于规则左部（父结点）的非终结符，可以直接指定其中心词，也可以通过使用相同的序号规定其中心词等于其右部的某个非终结符的中心词；
3. 链接约束：源语言骨架和目标语言骨架的非终结符子结点通过使用相同的序号建立对应关系，具有对应关系的非终结符互为翻译。

举例来说，一个汉英机器翻译模板可以表示如下：

S:2 ？ NP:1 岁:MP:2 了

S:be ？ NP:1 be year:NP:2 old

可以看到，这种规则比上下文无关规则表达上更为细腻。例如上述模板中如果去掉中心词约束，考虑一般的情况，显然这两条规则不能互为翻译。与实例相比，这个模板又具有更强的表达能力，因为这两个句子的主语（NP:1）和具体的岁数值都是可替换的。

该文还证明了这种模板的识别能力等价于 CFG，提出了使用这种模板进行翻译的算法，讨论了如何将属性运算引入翻译模板当中，并研究了如何从实例库中提取翻译模板的算法。该文作者在小规模范围内进行了实验，取得了较好的效果。

2.3 基于实例的机器翻译方法

2.3.1 起源与发展

长尾真在《采用类比原则进行日-英机器翻译的一个框架》[Nagao 1984]一文中最早提出了基于实例的机器翻译的思想。长尾真这篇论文的原文发表在一次不太著名的会议上，作者没有找到这篇论文的原文。从其他一些文献[Sato & Naogo 1990] [冯志伟 2001] [Turcato 1999] [Hutchins 1994]的综述来看，长尾真的基本思想是：

- 初学英语的日本人总是记住一些最基本的英语句子以及一些相对应的日语句子，他们要对比不同的英语句子和相对应的日语句子，并由此推论出句子的结构。参照这个学习过程，在机器翻译中，如果我们给出一些英语句子的实例以及相对应的日语句子，机器翻译系统来识别和比较这些实例及其译文的相似之处和相差之处，从而挑选出正确的译文。
- 人类并不通过做深层的语言学分析来进行翻译，人类的翻译过程是：首先把输入的句子正确地分解为一些短语碎片，接着把这些短语碎片翻译成其它语言的短语碎片，最后再把这些短语碎片构成完整的句子，每个短语碎片的翻译是通过类比的原则来实现的。
- 长尾真认为语言学的数据比语言学的理论更为持久（durable），因此可以作为一个机器翻译系统的更坚实的基础。

- **长尾真提出的基于实例的机器翻译方法采用原始的、未经分析和标注的双语数据，以及一个由源语言和目标语言词语等价物（对于动词表现为两个对应的格框架，对于其他类型的词语表现为词对）作为翻译过程的语言学知识支持。翻译过程主要是一个匹配的过程，目标是在数据库中寻找与输入句子在语义上最相似的实例。然后将被翻译句子和实例中的源语言句子互相比较产生一个翻译模板。这个模板可以用词对词的翻译来进行填充，并得到译文。**

当时长尾真在这篇论文中只是提出了一个设想，并没有真正的实验。不过这样一种想法引起了人们普遍的兴趣，并得到了广泛的采用，形成了一大类机器翻译方法。这些方法各有不同的特点，也有很多不同的名称，比如说：基于实例的机器翻译(Example-Based Machine Translation, EBMT, [Sumita & Iida 1991]), **基于记忆的翻译**(Memory-Based Translation, MBT, [Sato & Nagao 1990]), **转换驱动的机器翻译**(Transfer-Driven Machine Translation, TDMT, [Furuse & Iida 1992]), **基于个例的机器翻译**(Case-Based Machine Translation, CBMT, [Kitano 1993])**等等。在具体的名称上大家并没有形成统一的标准，比如说有些研究者把** EBMT **和** MBT **混为一谈，而有些研究者则对他们进行严格的区分。有些方法与长尾真最先提出的具体设想已相差甚远，不过这些方法总体上都采用了类别推理这种基本的思路。因此也可以被笼统地称为基于实例的机器翻译方法。**

下面主要介绍几种基于实例的机器翻译的具体做法。

2.3.2 Sato 和 Nagao 的方法

[Sato & Nagao 1990]**实现了一个基于实例的机器翻译系统，其基本设想是：将实例按照词语依存树配对的形式进行存储，同时保存对应关系链接的集合。**

举例来说，一对英语和日语句子：

- He eats vegetables. ⇔
  Kare ha yasai wo taberu.

可以表现为以下的 Prolog **事实语句：**

```
ewd_e([e1,[eat,v],
        [e2,[he,[pron]],
        [e3,[vegetable,n]]]).
jwd_e([j1,[taberu,v],
        [j2,[ha,p],
            [j3,[kare,pron]]],
        [j4,[wo,p],
            [j5,[yasai,n]]]]).
clinks([[e1,j1],[e2,j3],[e3,j5]]).
```

[Sato & Nagao 1990]**演示了如何利用一个以上的翻译实例并将其片断进行组合来获得一个完整的句子的译文。**

- He buys a book on international politics ⇔
  Kare ha kokusaiseiji nitsuite kakareta hon wo kau

这个翻译可以使用了以下实例中的片断组合而成：

- He buys a notebook ⇔
  Kare ha nouto wo kau
- I read a book on international politics ⇔
  Watashi ha kokusaiseiji nitsuite kakareta hon wo yomu

在翻译的过程中，每一个输入句子都被表示为一个或多个匹配表达式。每一个匹配表达式表示在实例库中找到的某个依存子树的特定结点上所进行的某种操作（即插入、删除和替换）。利用这些操作，可以通过数据库中找到的实例片段来组合得到输入的句子。[Turcato et al. 1999]给出了一个英语句子“ He eats mashed potatoes. ”的匹配表达式：

[e1, [r, e3, [e^x]]]

这里 r 表示替换，整个表达式的意思是“ 在实例 e1 中，用结点 e^x 替换结点 e3 ”。

对于每一个输入句子可能会产生多个有效的匹配表达式，系统使用一个评分系统来帮助选择最好的匹配。选择最好匹配的两个准则是：

- 被匹配的翻译单位越长越好
- 被匹配的翻译单位的上下文越相似越好

2.3.3 Kaji 的方法

[Kaji 1992]提出了一种两阶段基于实例的机器翻译实现方法，其系统结构图如下所示：

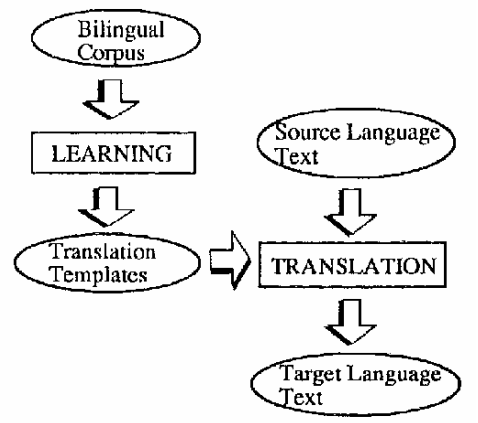


图 2.5 Kaji 提出的两阶段 EBMT 工作流程

可以看到，这里翻译过程中所使用的已经不是翻译实例本身，而是从翻译实例中抽取出来的翻译模板。一个翻译模板是一个双语句子对，其中对应的单元（词或短语）被配对并替换成变量。翻译模板的例子如下所示：

ADVP(X[NP] ？省略？？？)

/ ADVP (if X[NP] is omitted),

这个翻译模板是从下面的实例中学习得到的：

？？名？省略？？？？？？？？？？？定？？？

/ If the path name is omitted, the current path is assumed.

可以看到，这里的翻译模板可以是针对一个句子片段（短语）的，而且可以利用短语类信息。翻译模板的学习是在对齐的双语句子基础上同步进行句法分析和结构对齐得到的。系统中还引入了语义词典来对翻译模板进行细化，以解决翻译模板之间的冲突。

2.3.4 CMU 的泛化的基于实例的机器翻译方法

Pangloss 是一个由 CMU 和其他多所大学联合开发的机器翻译系统，这个系统中采用了多引擎的设想，一共采用了三个机器翻译引擎，其中一个使用的是一种范化的基于实例的机

器翻译方法（Generalized EBMT）[Brown 1996,2000]。

前面介绍的两种基于实例的机器翻译方法都要依赖于对实例进行句法分析或者依存关系分析，这与长尾真最初仅使用原始的未经分析和标记的语料库的想法实际上已相差甚远。相对来说，CMU 在 Pangloss 系统中采用的泛化的基于实例的机器翻译方法跟长尾真的最初的想法比较接近一些。实例的匹配完全采用字面匹配，不使用任何句法分析，只匹配尽可能长的单词序列。双语实例的对齐在翻译过程中实时进行。对齐算法也很简单，只采用了一部概率词典。

在这种方法中，对翻译实例的唯一标记操作就是所谓的泛化（Generalization），也就是说，把实例中一些具体词范化为一些类别。这里给出一对经过泛化的双语句子实例：

```
<PERSON> was in <CITY> on <DATE>.
<PERSON> war am <DATE> in <CITY>.
```

实例的泛化大大提高了实例的匹配率，可以减少实际翻译中所需要的实例库的规模。这种方法不需要深层的分析，简单实用，匹配率也较高。

2.3.5 基于实例的机器翻译方法的优缺点

基于实例的机器翻译方法的优点主要有：

- 译文直接从语料库中的实例变换而来，真实可靠，翻译质量高；
- 不需要花费大量人工去调试规则库，翻译的效果随着语料库的增大逐步提高；
- 翻译算法不需要进行深层的分析，速度快，效率高；
- 能够解决一些传统的基于规则的机器翻译系统处理不好的问题，特别是对一些不符合语法的语言现象处理较好。

基于实例的机器翻译的主要问题有：

- 覆盖率问题：基于实例的机器翻译往往很难达到很高的覆盖率，因而在很多情况下都是作为其他系统的补充；
- 泛化问题：为了提高覆盖率，往往要对翻译实例进行泛化，但泛化也会导致错误的匹配，特别是引入深层的句法分析或者依存关系分析以后。如何在泛化和匹配的正确率方面取得平衡也是一个问题；
- 对齐问题：对齐正确率直接影响到翻译的正确率，而词语和短语的对齐问题现在都没有达到非常高的程度；
- 实例的匹配问题：如何通过实例的组合得到被翻译的句子有很多种做法，如基于字符的匹配、基于短语结构的匹配、基于依存关系的匹配等等。

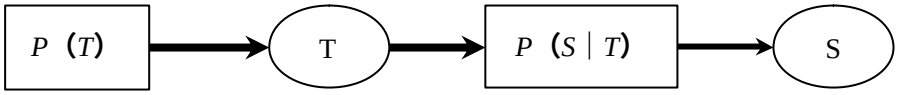
2.4 基于信源信道模型的统计机器翻译方法

基于信源信道模型的统计机器翻译方法源于 Weaver 在 1947 年提出的把翻译看成是一种解码的过程。其正式的数学框架是由 IBM 公司的 Brown 等人建立的[Brown 1990,1993]。这一类方法的影响非常大，甚至成了统计机器翻译方法的同义词。不过本文只把它作为统计机器翻译方法中的一类。

2.4.1 IBM 的统计机器翻译方法

2.4.1.1 基本原理

基于信源信道模型的统计机器翻译方法的基本思想是，把机器翻译看成是一个信息传输的过程，用一种信源信道模型对机器翻译进行解释。假设一段目标语言文本 T，经过某一噪声信道后变成源语言 S，也就是说，假设源语言文本 S 是由一段目标语言文本 T 经过某种奇怪的编码得到的，那么翻译的目标就是要将 S 还原成 T，这也就是就是一个解码的过程。



根据 Bayes 公式可推导得到：

$$T = \arg\max_{T} P(T)P(S \mid T)$$

这个公式在 Brown 等人的文章[5]中称为统计机器翻译的基本方程式（Fundamental Equation of Statistical Machine Translation）。在这个公式中，$P(T)$是目标语言的文本 T 出现的概率，称为语言模型。$P(S|T)$是由目标语言文本 T 翻译成源语言文本 S 的概率，称为翻译模型。语言模型只与目标语言相关，与源语言无关，反映的是一个句子在目标语言中出现的可能性，实际上就是该句子在句法语义等方面的合理程度；翻译模型与源语言和目标语言都有关系，反映的是两个句子互为翻译的可能性。

也许有人会问，为什么不直接使用 $P(T|S)$，而要使用 $P(T)P(S|T)$这样一个更加复杂的公式来估计译文的概率呢？其原因在于，如果直接使用 $P(T|S)$来选择合适的 T，那么得到的 T 很可能是不符合译文语法的（ill-formed），而语言模型 $P(T)$就可以保证得到的译文尽可能的符合语法。

这样，机器翻译问题被分解为三个问题：

1. 语言模型 $Pr(t)$的参数估计；
2. 翻译模型 $Pr(s|t)$的参数估计；
3. 搜索问题：寻找最优的译文；

从 1980 年代末开始到 1990 年代中期，IBM 的机器翻译研究小组在统计机器翻译的思想指导下进行了一系列的研究工作[5,6,3]并实现了一个法语到英语统计机器翻译系统。

对于语言模型 $Pr(t)$，他们尝试了采用 n 语法、链语法等语法模型。链语法模型比 n 元语法模型的优点在于可以考虑长距离的依赖关系。下面着重介绍翻译模型。

2.4.1.2 IBM 统计翻译模型

对于翻译模型 $Pr(t|s)$，IBM 公司提出了 5 种复杂程度递增的数学模型，简称为 IBM Model 1~5。模型 1 仅考虑词与词互译的概率 $t(f_j|e_i)$。模型 2 考虑了单词在翻译过程中位置的变化，引入了参数 $Pr(a_j|j,m,l)$，m 和 l 分别是目标语和源语句子的长度，j 是目标语单词的位置，a_j 是其对应的源语单词的位置。模型 3 考虑了一个单词翻译成多个单词的情形，引入了产出概率 $\phi(n|e_i)$，表示单词 e_i 翻译成 n 个目标语单词的概率。模型 4 在对齐时不仅仅考虑词的位置变化，同时考虑了该位置上的单词（基于类的模型，自动将源语言和目标语言单词划分到

50 个类中)。模型 5 是对模型 4 的修正，消除了模型 4 中的缺陷（deficiency)，避免对一些不可能出现的对齐给出非零的概率。

在模型 1 和 2 中，首先预测源语言句子长度，假设所有长度都具有相同的可能性。然后，对于源语言句子中的每个位置，猜测其与目标语言单词的对应关系，以及该位置上的源语言单词。在模型 3,4,5 中，首先，对于每个目标语言单词，选择对应的源语言单词个数，然后再确定这些单词，最后，判断这些源语言单词的具体位置。

这些模型的主要区别在于计算源语言单词和目标语言单词之间的连接（Connection）的概率的方式不同。模型 1 最简单，只考虑词与词之间互译的概率，不考虑词的位置信息，也就是说，与词序无关。好在模型 1 的参数估计具有全局最优的特点，也就是说最后总可以收敛于一个与初始值无关的点。模型 2 到 5 都只能收敛到局部最优，但在 IBM 的实验中，每一种模型的参数估计都依次把上一种模型得到的结果作为初始值，可以看到最后的结果实际上也是与初始值无关的。

下面以模型 3 为例，说明一下从源语言（英语）文本产生目标语言（法语）文本的过程：

1. 对于句子中每一个英语单词 e，选择一个产出率 ϕ，？概率为 $n(\phi|e)$；
2. 对于所有单词的产出率求和得到 *m-prime*；
3. 按照下面的方式构造一个新的英语单词串：删除产出率为 *0* 的单词，复制产出率为 *1* 的单词，复制两遍产出率为 *2* 的单词，依此类推；
4. 在这 *m-prime* 个单词的每一个后面，决定是否插入一个空单词 NULL，插入和不插入的概率分别为 *p1* 和 *p0*；
5. 设 ϕ_0 为插入的空单词 NULL 的个数。
6. 设 *m* 为目前的总单词数：$m\text{-}prime+\phi_0$；
7. 根据概率表 $t(f|e)$，将每一个单词 e 替换为外文单词 f；
8. 对于不是由空单词 NULL 产生的每一个外语单词，根据概率表 $d(j|i,l,m)$，赋予一个位置。这里 j 是法语单词在法语串中的位置，i 是产生当前这个法语单词的对应英语单词在英语句子中的位置，l 是英语串的长度，m 是法语串的长度；
9. 如果任何一个目标语言位置被多重登录（含有一个以上单词)，则返回失败；
10. 给空单词 NULL 产生的单词赋予一个目标语言位置。这些位置必须是空位置（没有被占用)。任何一个赋值都被认为是等概率的，概率值为 $1/\phi_0$。
11. 最后，读出法语串，其概率为上述每一步概率的乘积。

2.4.1.3 搜索算法

从上述 IBM Model 3 的介绍中可以看出，对于统计机器翻译而言，搜索算法是一个严重的问题。因为搜索空间一般都是随着源语言句子的大小呈指数增长的，要在多项式时间内找到全局最优解是不可能的。为了在尽可能短的时间内找到一个可接受的译文，必须采用各种启发式搜索策略。

对于搜索问题，IBM 采用一种在语音识别取得广泛成功的搜索算法，称为堆栈搜索（Stack Search)，这里不做详细介绍。其他的搜索算法还有柱形搜索（Beam Search)、A*搜索等等。

虽然搜索问题很严重，不过 IBM 的实验表明，搜索问题并不是统计机器翻译的瓶颈问题。实际上，统计机器翻译的错误只有两种类型：

1. 模型错误：即根据模型计算出概率最高的译文不是正确译文；
2. 搜索错误：虽然据模型计算出概率最高的译文是正确译文，但搜索算法没有找到这

个译文。

根据 IBM 的实验，后一类错误只占所有翻译错误的 5%。

2.4.1.4 Candide 系统

与传统的基于转换的机器翻译方法相比，我们可以看到 IBM 的统计机器翻译方法中没有使用任何的非终结符（词性、短语类等）。所有的参数训练都是在词的基础上直接进行的。

IBM 的研究者基于上述统计机器翻译的思想，以英法双语对照加拿大议会辩论记录作为双语语料库，开发了一个法英机器翻译系统 Candide。

表 2.1 Candide 和 Systran 的 ARPA 测试结果

	Fluency		Adequacy		Time Ratio	
	1992	1993	1992	1993	1992	1993
Systran	.466	.540	.686	.743		
Candide	.511	.580	.575	.670		
Transman	.819	.838	.837	.850	.688	.625
Manual		.833		.840		

上表是 ARPA 测试的结果，其中第一行是著名的 Systran 系统的翻译结果，第二行是 Candide 的翻译结果，第三行是 Candide 加人工校对的结果，第四行是纯人工翻译的结果。评价指标有两个：Fluency（流利程度）和 Adequacy（合适程度）。（Transman 是 IBM 研制的一个译后编辑工具。Time Ratio 显示的是用 Candide 加 Transman 人工校对所用的时间和纯手工翻译所用的时间的比例。）

从指标上看，Candide 已经和采用传统方法的商品系统 Systran 不相上下，译文流利程度甚至已经超过了 Systran。

不过，Candide 采用的并不是纯粹的统计模型。实际上，Candide 采用的是也是一种“分析－转换－生成”的结构。分析阶段使用了形态分析和简单的词序调整，生成阶段也使用了词序调整和形态生成，分析和生成这两个过程都是可逆的。只有在转换阶段使用了完全的统计机器翻译方法。这种做法可以达到三个目的：使隐藏在词语变形之后的英法语对应规则性显示出来；减少了双语的词汇量；减轻了对齐的负担。不过，也正因为这个原因，有人抨击统计机器翻译是“石头汤（Stone Soup）”，并认为在这个系统中真正起作用的还是规则方法，因为英法两种语言词序本身相差就不是太大。通过预先的词序调整，两种语言的词序更为接近，这实际上避开了 IBM 统计机器翻译方法的最大问题。

2.4.1.5 IBM 统计机器翻译方法小结

IBM 提出的统计机器翻译基本方程式具有非常重要的意义。而 IBM 的其他工作只是对这个基本方程式的一种理解。从理论上说，IBM 的模型只考虑了词与词之间的线性关系，没有考虑句子的结构。这在两种语言的语序相差比较大时效果可能会不太好。如果在考虑语言模型和翻译模型时将句法结构或语义结构考虑进来，应该会得到更好的结果。

IBM 提出的统计机器翻译方法在研究者中引起了相当大的兴趣，很多研究者都开展了相关的工作，并取得了一些进展。下面简要介绍其中的一些改进。

2.4.2 王野翊（Yeyi Wang）在 CMU（卡内基梅隆大学）的工作

王野翊在他的博士论文[Wang 1998a,1998b]中提出了一种对于 IBM 统计翻译模型的一种改进方法。

由于 IBM 的模型完全没有考虑句子的结构信息，这使得人们怀疑 IBM 模型能否在句法结构相差较大的语言对中获得成功。王野翊在他的口语机器翻译实验中也发现，对于德语和英语这两种句法结构相差较大的语言来说，IBM 的词对齐模型是翻译错误的一个重要来源。为此，王野翊提出了一种改进的统计翻译模型，称为基于结构的翻译模型，这种模型与 IBM 五种翻译模型的关系如下图所示。

这个模型分为两个层次：粗（Rough Alignment）对齐模型和细对齐（Detailed Alignment）模型。首先，源语言和目标语言的短语通过一个粗对齐模型进行对齐，然后短语内的单词再通过一个细对齐模型进行对齐。

为了在粗对齐阶段实现双语短语的对齐，王野翊引入了一种双语的文法推导算法。在训练语料库上，通过基于互信息的双语词语聚类和短语归并反复迭代，得到一组基于词语聚类的短语规则。再用这组规则进行句子的短语分析。

王野翊的实验表明，结构的引入不仅使统计机器翻译的正确率有所提高（错误率降低了11%），同时还提高了整个系统的效率，也缓解了由于口语数据的严重缺乏导致的数据稀疏问题。

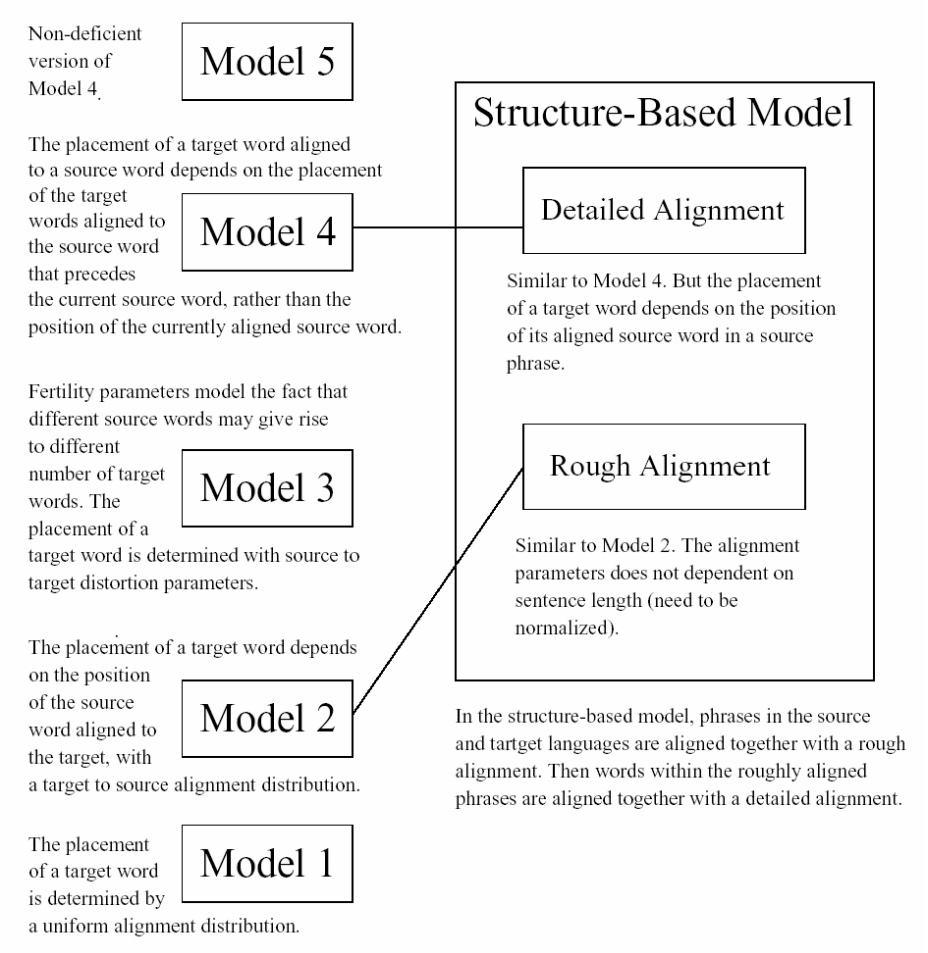


图 2.6 王野翊双层统计翻译模型

2.4.3 约翰·霍普金斯大学（JHU）的统计机器翻译夏季研讨班

IBM 提出统计机器翻译方法引起了研究者广泛的兴趣。不过，由于其他人无法得到 IBM 的源代码，而要进行统计机器翻译的研究，首先需要重复 IBM 的统计机器翻译试验，然后才谈得上对它进行改进。这将面临着编码方面巨大的工作量。于是，在 1999 年夏天，很多相关的研究者会聚在约翰·霍普金斯大学（JHU）举行了一个夏季研讨班，大家共同合作，重复了 IBM 的统计机器翻译试验，并开发了一个源代码公开的统计机器翻译工具包——Egypt。在这以后，这些研究者回到各自的研究机构，继续开展相关的研究工作，并提出了各种改进的模型，使得统计机器翻译的研究又出现了一个新的高潮。

在约翰·霍普金斯大学的 1999 年统计机器翻译夏季研讨班上，研究者们构造了一个基本的统计机器翻译工具集 Egypt，并将该工具集在感兴趣的研究者中间自由散发。在研讨班上，他们使用这个工具集作为试验的平台进行了一系列的实验[Al-Onaizan 1999]。

研讨班开始时预期达到的目标如下：

1. 构造一个统计机器翻译工具并使它对于研究者来说是可用的。这个工具集应该包含语料库准备软件、双语文本训练软件和进行实际翻译的实时解码软件。它主要瞄准两类用户：
   - 一类用户的问题是："我有一个双语语料库——我能用它做什么？"
   - 另一类用户的问题是："我有一个统计机器翻译的新想法——我如何测试它"
2. 在研讨班上用这个工具集构造一个捷克语—英语的机器翻译系统；
3. 进行基准评价。这个评价应该包含客观评价（统计模型困惑度）和主观评价（质量的人工判断），并试图使二者互相联系。还要产生一个学习曲线，用于显示系统性能如何随着双语语料的数量发生变化。
4. 通过使用形态和句法转录机改进系统性能；
5. 在研讨班最后，在一天之内构造一个新语对的翻译系统。

研讨班最后完全达到了上述目标。除此之外，研讨班还完成了以下实验：提高双语训练的速度，使用双语词典，使用同源词。并构造了一些工具来支持以上实验，包括一个复杂的图形界面用于浏览词对词对齐的结果，一些语料库的准备和分析工具，和一个人工判断的评价界面。

EGYPT 工具包包含以下几个模块：

1. GIZA：这个模块用于从双语语料库中抽取统计知识（参数训练）。
2. Decoder：解码器，用于执行具体的翻译过程（在信源信道模型中，"解码"就是"翻译"）。
3. Cairo：整个翻译系统的可视化界面，用于管理所有的参数、查看双语语料库对齐的过程和翻译模型的解码过程。
4. Whittle：语料库预处理工具。

Egypt 是个免费的工具包，其源代码可以在网上自由下载。这为相关的研究工作提供了一个很好的研究基础。

2.4.4 Yamada 和 Knight 的工作——基于句法的统计翻译模型

南加州大学信息科学研究所(ISI/USC)的 Kevin Knight 是统计机器翻译的主要倡导者之一，在统计机器翻译方面做了一系列的研究和推广工作，他也是 JHU 的统计机器翻译夏季讨论班的主要组织者之一[Knight 1999]。

ISI/USC 在 1994 年至 1999 年间承担美国国防部支持的 GAZELLE 项目，目的是开发日语、阿拉伯语和西班牙语到英语的机器翻译系统，面向不受限的新闻文本。主要着眼于提高翻译质量和降低新语言机器翻译系统的开发难度。

目前他们又在美国国防部主持的 TIDES 计划下面承担了一个 ReWrite 项目，主要目标仍然是研究新的基于语料库的机器翻译方法以提高翻译精度和快速开发新语言的机器翻译系统，同时也试图将类似的基于语料库的技术用于自然语言生成和摘要。

Yamada, Knight 等人在 IBM 的统计翻译模型的基础上，提出了一种基于句法结构的统计翻译模型[Yamada 2001,2002]。其主要的思想是：

1. IBM 的信源信道模型中，噪声信道的输入和输出都是句子，而在基于句法结构的统计翻译模型中，噪声信道的输入是一棵句法树，输出是一个句子；
2. 在翻译过程中，对源语言句法树进行以下变换：
   a) 对句法树进行扁平化处理（将相同中心词的多层结点压缩到一层）；
   b) 对于源语言句法树上面的每一个结点的子节点进行随机地重新排列（N 个子节点就有 N!种排列方式），每一种排列方式都有一个概率；
   c) 对于句法树任何一个位置随机地插入任何一个新的目标语言单词，每一个位置、每一个被插入的单词都有不同的概率；
   d) 对于句法树上每一个叶节点上的源语言单词翻译成目标语言单词，每一个不同的译文词选择都有不同的概率；
   e) 输出句子，其概率为上述概率的乘积。

从现有的文章中看，他们的实验采用了一个从英日词典中抽取的例句语料库，一共只有 2121 个句子，平均句长不到 10 个词。虽然其结果比 IBM Model 5 更好，不过由于他们的实验规模还比较小，严格来说并不具有说服力。

2.4.5 Och 等人的工作

德国 RWTH Aachen – University of Technology 等人在统计机器翻译领域也开展很多的工作。

在德国主持开发的著名的语音机器翻译系统 Verbmobil 中，Och 所在的研究组承担了其中统计机器翻译模块[Och 1999]。与 IBM 的模型相比，他们主要做了以下改进：

1. 为了解决数据稀疏问题，他们采用了基于类的模型，利用一种自动的双语词聚类技术，将两种语言的每一个词都对应到一个类中[Och 1998]，总共使用了 400 个类；
2. 在语言模型上，采用了基于类的五元语法模型，采用回退（Back-off）平滑算法；
3. 在翻译模型上，采用了一种称为对齐模板（Alignment Template）的方法，实现了两种层次的对齐：短语层次的对齐和词语层次的对齐。对齐模板也采用基于类的对齐矩阵的形式表示，如下图所示。对齐模板的获取是自动进行的，在对训练语料进行词语对齐以后，所有可能的对齐模板都被保存下来，并根据其在语料库中出现的频率赋予不同的概率。对于一个新句子进行短语匹配的过程类似于一个汉语词语切分的过程，采用一个动态规划算法，寻找概率最大的路径。
4. 为了搜索的方便起见，他们对于 IBM 提出的统计机器翻译基本方程式进行了修改，用一个反向的翻译模型取代了正常的翻译模型，如下所示：

$$S = \max_{S} P(S)P(S \mid T)$$

通过实验他们发现，这种改变并没有降低总体的翻译正确率。

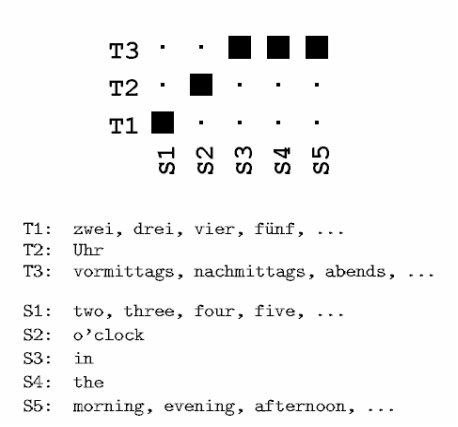


图 2.7　Och 的对齐模板

2.5 基于最大熵模型的统计机器翻译方法

正如上一节所述，Och 等人在进行统计机器翻译实验时发现，把 IBM 统计机器翻译基本方程式中的翻译模型换成反向的翻译模型，总体的翻译正确率并没有降低，这用信源信道理论是无法解释的。于是，他们借鉴了[Papineni, 1997][Papineni, 1998]**中统计自然语言理解的一种思路，提出了基于最大熵模型的统计机器翻译方法**[Och 2002]**。这是一种比基于信源信道的统计机器翻译方法更为一般化的方法，基于信源信道的方法可以看作是基于最大熵方法的一个特例。**

与基于信源信道的方法不同，基于最大熵的方法没有语言模型和翻译模型的划分（虽然也可以将它们作为特征），因而是一种直接翻译模型。

最大熵，又称最大熵原理、最大熵模型，或者最大熵方法，是一种通用的统计建模的方法。这里简单介绍一下最大熵方法的基本思想[Berger 1996]**。**

对于一个随机事件，假设已经有了一组样本，我们希望建立一个统计模型，来模拟这个随机事件的分布。为此需要选择一组特征函数，并用这组特征函数的线性组合的指数形式来模拟这个随机事件的概率分布。

假设 e**、**f **是机器翻译的目标语言和源语言句子，**$h_1(e,f), \ldots, h_M(e,f)$**分别是** e**、**f **上的** M **个特征，**$?_1, \ldots, ?_M$ **是与这些特征分别对应的** M **个参数（权值），那么翻译概率可以用以下公式模拟：**

$$\Pr(e \mid f) \approx p_{l_1 \ldots l_M}(e \mid f)$$

$$= \exp[\sum_{m=1}^{M} l_m h_m(e, f)] \Big/ \sum_{e'} \exp[\sum_{m=1}^{M} l_m h_m(e', f)]$$

这里的分母起到一个概率归一化的作用。

对于给定的 f**，其最佳译文** e **可以用以下公式表示：**

$$\hat{e} = \arg\max_e \{\Pr(e \mid f)\}$$

$$= \arg\max_e \{\sum_{m=1}^{M} l_m h_m(e, f)\}$$

可以看到，如果将两个特征分别取为 $\log p(e)$**和** $\log p(f|e)$**，并取** $?_1 = ?_2 = 1$**，那么这个模型就等价于信源信道模型。**

为了得到这个概率模型，需要对上述的参数（权值）进行估计。根据最大熵原理，参数

估计的原则是，在保证得到的统计模型在这一组特征上，与样本中的分布完全一致的前提下，同时又保证这个概率模型尽可能的“均匀”（也就是使模型的熵值达到最大），以确保除了这一组特征之外，这个模型没有其他的任何偏好。

在最大熵方法中最常用的做法是采用二值特征，可以用 GIS 算法或 IIS 算法进行参数训练。而在 Och 提出的基于最大熵的统计机器翻译中，采用的是一种基于实数值特征的最大熵模型，模型的参数不能使用通常 IIS 算法进行训练。为此[Och 2002]提出了采用了一种判别学习方法（Discriminative Training），其训练的优化准则为：

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \log p_{\lambda_1^M}(\mathbf{e}_s|\mathbf{f}_s) \right\}$$

这个判定准则是凸的，并且存在全局最优。

[Och 2002]介绍了他们在基于最大熵的统计机器翻译方法上的一系列实验：

1. 首先将信源信道模型中的翻译模型换成反向的翻译模型，简化了搜索算法，但翻译系统的性能并没有下降；
2. 调整参数 $?_1$ 和 $?_2$，系统性能有了较大提高；
3. 再依次引入其他一些特征，系统性能又有了更大的提高。

他们引入的其他特征包括：

1. 句子长度特征：对于产生的每一个目标语言单词进行惩罚；
2. 附加的语言模型特征：一个基于类的语言模型特征；
3. 词典特征：计算给定的输入输出句子中有多少词典中存在的共现词对。

可以看到，采用基于最大熵的统计机器翻译方法，确实比简单地采用信源信道模型可以较大地提高系统的性能。

基于最大熵的统计机器翻译方法为统计机器翻译的研究提供了一个更加广阔的视野，这篇论文获得了 ACL2002 的最佳论文奖。

2.6 多引擎机器翻译方法

由于各种不同的机器翻译方法各有特长，也各有缺点，没有哪一种单一的机器翻译方法能够达到理想的效果，因此采用多引擎的方法，希望各种方法能够互补，以达到总体效果的最优，就成为了一种自然的选择。目前多引擎的机器翻译目前已经被广泛采用，而实践证明这种方法也确实是有效的。

[Frederking 1994]提出了一种多引擎机器翻译的方法。系统结构如下图所示。

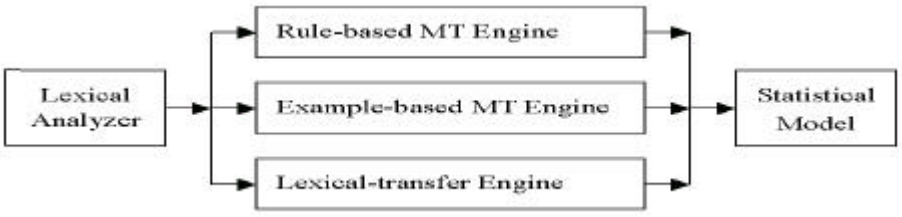


图 2.8 系统级的多引擎机器翻译结构

该方法基本思想描述如下：

1. 多个的翻译引擎同时对输入的句子进行翻译，不仅仅对整个句子进行翻译，而且对句子的任何一个片断也可以给出相应的译文，同时对这些译文片断给出一个评分。
2. 各个翻译引擎共享一个类似线图（Chart）的数据结构，根据其源文片断所处的位置，

将这些译文片断放在这个公共的线图结构之中。

3. 对各个引擎给出的片断的评分进行一致化处理，使之具有可比较性。
4. 采用一个动态规划算法（称为 Chart Walk 算法）选择一组刚好能覆盖整个源文输入句子，同时又具有最高总分的译文片断，作为最后输出的译文。

[Hogan 1998]通过一个简单的实验，证明这种方法确实可以得到比任何一种单一的方法都更高的准确率。

我们把上述这种采用多个完全独立的机器翻译系统进行集成的多引擎结构称为系统级多引擎结构。在很多多引擎的机器翻译系统中，并不都是采用完全独立的多个翻译引擎对源文进行翻译，而是在机器翻译系统的多个主要功能模块（部件）中分别采用多引擎技术，我们称之为部件级多引擎结构。

下面介绍两个典型的多引擎机器翻译系统。

2.6.1 Pangloss 系统

PANGLOSS 系统是美国卡内基梅隆大学等单位研制的一个多引擎的西班牙—英语的机器翻译系统[Brown 1995]。该系统采用系统级多引擎结构。

该系统总共包括三个翻译引擎：一个基于转换的翻译引擎（Transfer MT）、一个基于知识（中间语言）的翻译引擎（KBMT）和一个基于实例的翻译引擎（EBMT）。其系统结构如下图所示：

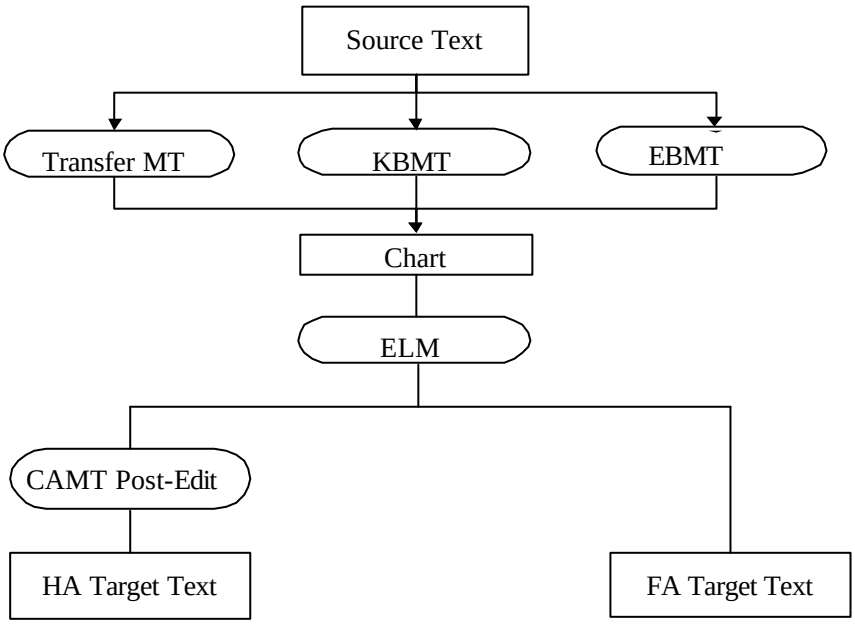


图 2.9　Pangloss 多引擎系统结构

其中 CAMT Post-Edit 是 Computed-Aided Post-Edit（计算机辅助译后编辑）。HA Target Text 指人工辅助产生的译文（Human-Aided Target Text），FA Target Text 指全自动产生的译文（Fully-Automated Target Text）。ELM 是英语语言模型（English Language Model）。

在 EBMT 翻译引擎中，不需要任何语言结构知识，需要的语言资源有一个双语例句库，一部双语电子词典和一个目标语言的同义词分类器。其中的双语例句库的规模达到了 72 万西班牙—英语语句对，主要来自于 LDC 提供的联合国多语言语料库（UN Multilingual Corpus, Linguistic Data Consortium）；而目标语言的同义词分类器是从 WordNet 中提取得到的，连同

双语词典一起来寻找源语言和目标语言语句中词语的关联。根据统计，该 EBMT 引擎对于不限制领域的输入能够达到 70.2%的覆盖率。

实现方法包括两个步骤：首先通过查找例句库的索引寻找同输入匹配的最长组块（chunk），然后进行语句片段的对齐，确定组块的译文。为了能有效的利用已有的例句，他们提出了对例句进行了泛化（Generalization）的思想。用一种较为一般化的模板来取代单纯的例句。

在 PANGLOSS 系统中除了 EBMT 翻译引擎外，还有一个基于转换的机器翻译引擎和一个基于知识的翻译引擎。对于一个输入语句，每一个翻译引擎都可能对其进行翻译，并产生一些片段的译文，然后把这些产生的片断译文放到一个统一的线图中，最后根据一个英语的统计语言模型（ELM）来决定线图（Chart）中的最佳路径作为译文输出，这样做可以尽量结合各个翻译引擎的优点，有利于产生最好质量的译文。与前述 Frederking 提出的 Chart Walking 算法的区别在于,这里用于选择最佳路径的函数不是根据各个翻译引擎给出的评分，而是根据英语的统计语言模型来决定，哪一个译文在英语中出现的可能性最大。

2.6.2 Verbmobil 系统

Verbmobil 是由德国教育与研究部（BMBF）资助的一个为期 8 年（1993-2000）语音机器翻译项目[Wahlster 2000]，涉及三种语言（德语、英语、日语）的双向翻译。世界三大洲的 31 个研究机构、369 名科学家和 919 名学生（硕士生、博士生和博士后）参与了这个项目的研究。

Verbmobil 系统与一般的文本翻译系统不同之处主要体现在：

1. 语音处理：要进行语音识别和语音合成。该系统的目标很高，实现了 GSM 语音条件下的自动翻译，除了一开始拨打 Verbmobil 语音服务电话以外，整个系统的服务可完全用 GSM 电话通过语音方式实现，无需任何按键操作；系统具有语音自适应能力，一开始使用与说话者无关的语音识别模块，通过一段时间对话后，自动适应说话者的口音，提供识别正确率；
2. 处理自然的语音：要考虑现实口语中的各种复杂现象，如停顿、重复、修正、丢词等等；
3. 要建立对话模型，理解句子的语义，并考虑上下文进行翻译，甚至要猜测说话者的意图；
4. 系统开发面向三个商务领域：约会安排、旅行计划和远程 PC 维护，复杂程度依次增加，如下表所示：

表 2.2　VerbMobil 系统的应用目标

场景一：约会安排	场景二：旅游计划	场景三：远程 PC 维护
何时？	何时？何地？如何？	何物？何时？何地？如何？
关注时间性表达	关注时间性和空间性表达	一个词汇受限的子语言
词汇量：2500/6000	词汇量：7000/10000	词汇量：15000/30000

可以看出，虽然领域受到限制，但这还是一个相当复杂的系统，不仅处理的问题面极其广泛，采用的技术也极其复杂多样。语音处理领域和自然语言处理领域中几乎所有的各种技术都在这个系统中有所反映。整个系统由 69 个互相交互的模块构成。其中用到的自然语言处理技术包括：组块分析、概率 LR 分析、HPSG 分析、对话行为（Dialog Act）分析、基于

统计的翻译、基于子串（substring）的翻译、基于模板的翻译、基于模板的转换、语义分析、上下文相关歧义的消解、基于规划的话语生成，等等。

整个系统采用一种复杂的部件级多引擎结构。如下图所示：

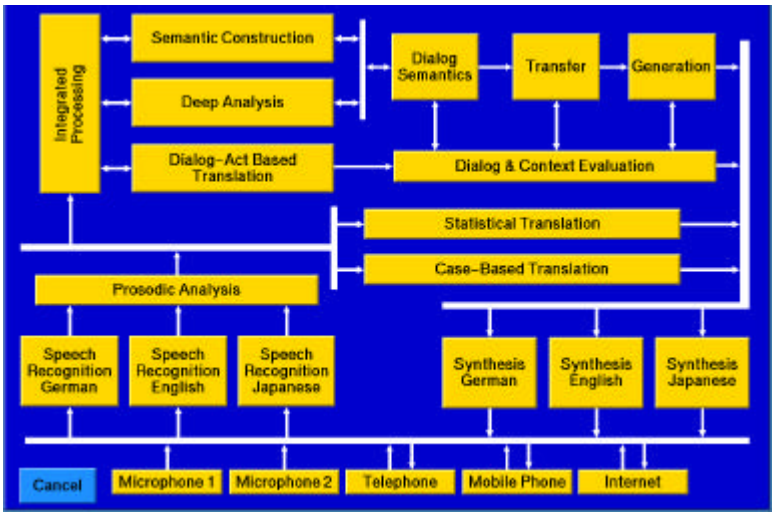


图 2.10　VerbMobil 的系统结构

系统采用一种多黑板结构用于模块之间的数据交互，模块之间不能直接通信。黑板结构还有利于各个模块之间的并行执行。总共采用了 198 个黑板结构用于 69 个不同模块之间的通讯。一种叫做 VIT（Verbmobil Interface Terms）的数据结构在中心黑板的深层处理中用作深层的语义表示形式。

句法分析阶段采用了多引擎技术，利用了三个分析引擎，分别是：组块分析器（Chunk Parser）、概率 LR 分析器和 HPSG 分析器。其中组块分析器鲁棒性最好，但结果质量最差，HPSG 分析器鲁棒性最差，但结果质量最好。

系统管理上采用了严格的软件工程方法，保证了项目的顺利实施。

项目组织上，Verbmobil 强调各个研究组之间的竞争，通过频繁的目标评价。项目中不同的团队对于每个特定的目标都提出了竞争的解决方案，通过正式的评价，从中选择最好的方案或者与次好的方案合并以改进总体的效果。

多种基准的测试以及大规模端对端评价实验令人信服地表明，Verbmobil 的最终版本系统中达到了所有的预定目标，有些目标甚至被超越。在大规模翻译实验中，正确翻译率达到大约 80%，在真实用户的端对端测试中，90%的对话任务获得成功。

2.7 机器翻译方法的分类

从上面的介绍可以看到，机器翻译方法是千差万别，丰 富多彩的。这里，本文试图从几个不同的角度对这些机器翻译方法进行分类。

2.7.1 按翻译转换的层面进行分类

对语言的分析理解存在很多不同的层面，而翻译转换可以在这些不同的层面上进行。根据翻译转换层面的不同，可以将机器翻译方法进行分类。

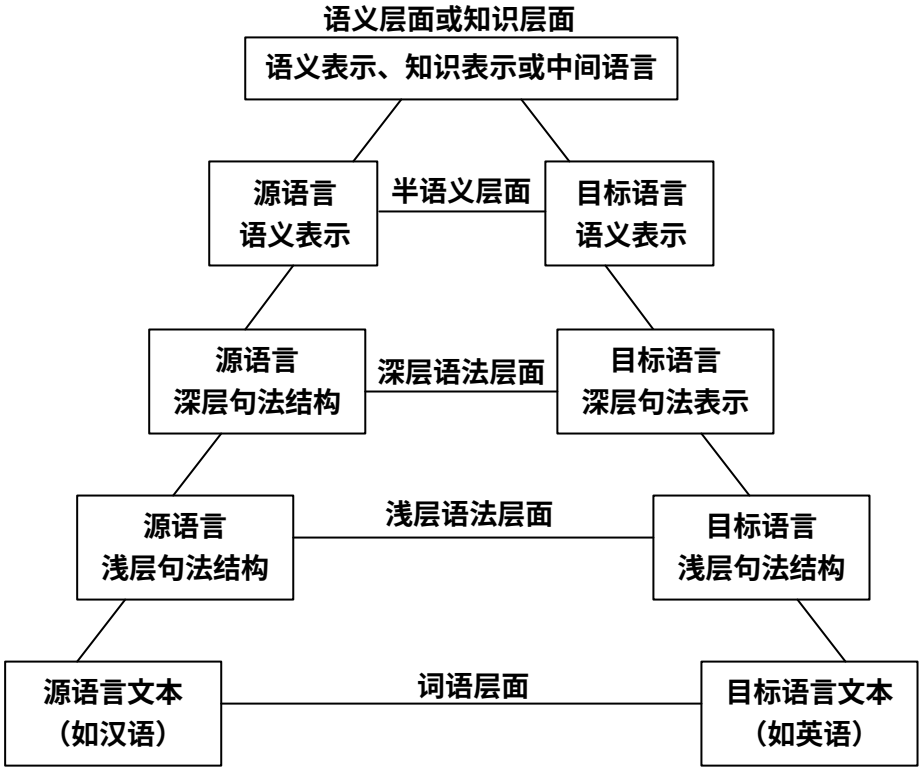


图 2.11 **机器翻译转换的层面**

一般而言，机器翻译的翻译转换操作可以在以下一些层面进行：

1. **词语层面：对输入的句子进行词法分析（词法分析可深可浅），不进行任何句法分析或语义分析，直接进行翻译。早期的直接翻译法就是在这个层面进行的，而** IBM **的基于信源信道模型的统计机器翻译方法，也可以认为是在这个层面进行的。对于结构相近，词序变化不大的语言之间的翻译，这种方法可以取得较好的效果。但对于词序变化较大的语言之间的翻译，这种方法效果较差；**
2. **浅层语法层面：对输入的句子进行浅层分析（组块分析），然后将翻译分为组块内翻译和组快间翻译两个步骤，在翻译的过程中保证组快的完整性（不可分割）。对于词序变化较大的语言之间的翻译，这种方法比单纯基于词语层面的转换效果更好；**
3. **深层语法层面：对输入的句子进行完全的句法分析，得到完全句法树（或者是没有标注词语间语义关系的依存树），将源语言的完全句法树（或依存树）转换成目标语言的完全句法树（或依存树），然后才生成目标语言句子。传统的基于转换的机器翻译系统大多数都是在这个层面上进行转换的；（注意，依存树实际上是一种介于句法表示和语义表示之间的形式，这里我们把没有标注词语间语义关系的依存树当作是句法表示形式，而把标注了词语间语义关系的依存树当作是语义表示形式）；**
4. **半语义层面：对输入的句子进行句法和语义分析，得到源语言的某种语义表示形式，如语义网络、逻辑形式或者标注了词语间语义关系的依存树，然后将这种语义形式转换成目标语言的相应的语义表示形式，然后再生成目标语言；（注意，这里的语义表示形式是一种和具体语言相关的语义表示形式，我们称之为半语义表示形式）；**
5. **知识层面：对输入句子进行彻底的句法语义分析和常识推理，得到一种与语言无关的语义表示形式、知识表示形式或中间语言表示形式，由于这种表示形式是具体语**

言无关的，因此无需经过转换阶段，可以直接生成目标语言句子。

一些常见的机器翻译范式，如基于转换的方法、基于实例的方法和基于统计的方法，并不一定能够简单地归入以上各个层面。

基于转换的方法一般是在深层句法层面或者半语义层面进行转换的；

基于实例的方法就更灵活了，有的方法是基于字面匹配的，可以认为是一种在词语层面转换的方法，而有的是基于组块分析或者完全句法分析甚至语义分析的，因此也可以认为是在浅层句法层面、深层句法层面或者半语义层面进行转换的方法；

基于统计的方法中，IBM 的信源信道方法（模型 1～5）虽然引入了简单的词法分析和词序调整，不过还是只能归入基于词语层面转换的方法；王野翊和 Och 对 IBM 统计翻译模型改进后的方法是在浅层句法层面进行转换的；Yamada 和 Knight 试图将 IBM 的方法改造成在深层句法曾经进行转换的方法，不过这种改造还不彻底，只是在源语言方面利用了深层句法结构，但并不转换成目标语言的句法结构，而是直接生成目标语言句子。

2.7.2 按语言知识的表示形式进行分类

从知识工程的角度看机器翻译，机器翻译可以认为是某种知识表示、知识获取和知识推理的过程，其中，知识表示形式是一个核心问题。这里我们可以从语言知识表示形式的角度对机器翻译进行分类，大致可以分为三类：实例表示形式、符号表示形式、数值表示形式。

实例表示形式直接使用人类翻译的实例作为知识表示形式，不进行任何抽象。实例表示形式的优点是正确率高，缺点是覆盖率低；

符号表示形式使用某种符号化的表示方法，最常见的符号化知识表示方法就是语言规则。规则表示方法的优点是形式直观、可以人工修改、颗粒度大小灵活可变、概括性强；缺点是不够全面和细致，导致系统鲁棒性差，人工编写知识库费时费力（当然某些规则形式的知识也可以自动获取）；

数值表示形式包括人工神经元网络形式和统计形式，目前以统计形式占主流地位。统计表示方法优点是学习方便、知识表示全面细致、导致系统鲁棒性好，知识库往往采用自动方法获取，语料库加工一旦完成，知识获取工作可以利用不同的统计方法重复进行，省时省力；缺点是数据稀疏问题严重，不方便表示深层语言知识，对语料库依赖性强，对不同领域知识的概括性差。

2.8 小结

目前，机器翻译三种主流的方法是基于规则的方法、基于实例的方法和基于统计的方法。

基于规则的方法虽然也还有人在继续研究[马红妹 2002]，不过总体上已经比较成熟，其主要的瓶颈在于通过人工编写的方式获得大规模语言知识的成本太高，在研究上很难取得更大的突破。

基于实例的方法也是目前的研究热点之一，不过，基于实例的方法研究总体上比较分散，各个系统的做法相差比较大，还没有形成一种或几种比较清晰的、大家都公认的范式。总的来说，基于实例的方法面临的主要问题是覆盖率问题，所以在很多情况下，基于实例的机器翻译引擎通常是作为其他翻译引擎的补充，而很少作为独立的机器翻译引擎存在。

IBM 提出的统计机器翻译方法，不仅仅是对机器翻译，而是对于整个的自然语言处理，都产生了长远而深刻的影响。由于各方面的原因，虽然 IBM 的统计机器翻译实验在初期取得了很大的成功（在 ARPA 测试中获得了超过 Systran 的结果），但 IBM 的统计机器翻译工

作并没有坚持下来，人们对统计机器翻译的怀疑也始终挥之不去。虽然统计方法在自然语言的很多其他领域都取得了成功，不过在机器翻译领域，统计方法的有效性并没有很快得到普遍的承认。

应该说，IBM 的统计机器翻译工作是有一定超前性的。IBM 在 1980 年代末到 1990 年代初就开始进行统计机器翻译工作，在 IBM 公布其实验结果之后的很多年，都没有人能够重复类似的结果。随着时间的推移，到了 1990 年代末，相关的研究工作开始得到重视，王野翊、Malamed、Knight、Och 等人都重复了 IBM 的工作并进行了一定的改进。统计机器翻译方法又出现了一个小的高潮。在 JHU 的夏季研讨班的总结报告[Al-Onaizan 1999]上有一段话耐人寻味：

> 当解码器的原形系统在研讨班上完成时，我们很高兴地惊异于其速度和性能。在 1990 年代早期在 IBM 公司举行的 DARPA 机器翻译评价时，我们曾经预计只有很短（10 个词左右）的句子可以用统计方法进行解码，即使那样，每个句子的解码时间也可能是几个小时。在早期 IBM 的工作过去将近 10 年后，摩尔定律、更好的编译器以及更加充足的内存和硬盘空间帮助我们构造了一个能够在几秒钟之内对 25 个单词的句子进行解码的系统。为了确保成功，我们在搜索中使用了相当严格的阈值和约束，如下所述。但是，解码器相当有效这个事实为这个方向未来的工作预示了很好的前景，并肯定了 IBM 的工作的初衷，即强调概率模型比效率更重要。

在 2002 年的 ACL 会议上，Och 等人关于统计机器翻译的论文[Och 2002]获得了大会最佳论文奖。在 2002 年 NIST 举办的机器翻译评测中，Och 所在的德国亚琛大学（RWTH Aachen – University of Technology）提交的系统获得了最好的成绩，统计机器翻译方法的优势得到了明显的体现。在 2003 年 NIST 的机器翻译评测中，Och 所在的美国南加州大学信息科学学院（USC/ISI）提交的系统又在各项评测中都取得了最好的成绩。

应该说，统计机器翻译方法到现在为止还没有走到尽头，统计机器翻译方法的潜力还远远没有发挥出来。现有的统计机器翻译方法还只是利用语言的浅层结构信息，利用深层结构信息的统计机器翻译方法还刚刚起步[Yamada 2001,2002]，模型还很不完善，实验规模也很小，说服力不强。作者相信，随着各种深层的句法语义信息逐渐引入到统计机器翻译的模型中来，机器翻译的译文质量还会有一个较大的提高。这也是本文工作的努力方向。

第3章 基于层叠隐马尔可夫模型（Cascaded HMM）的汉语词法分析

3.1 汉语分析技术概述

3.1.1 汉语词法分析的难点

对源语言的分析，是机器翻译的重要步骤之一。同样，汉语的分析对于汉英机器翻译来说也是至关重要的。

语言的分析一般可分为词法分析（形态分析）、句法分析、语义分析、语用分析等几个层次，本章主要讨论汉语的词法分析技术。

汉语是一种孤立语（又称分析语），与作为曲折语和黏着语的其他一些语言相比，汉语在语法上有一些明显的特点。

从语言的形式上看，这种特点主要体现在以下几个方面：

a) 汉语的基本构成单位是汉字而不是字母。常用汉字就有 3000 多个（GB2312 一级汉字），全部汉字达数万之多（UNICODE 编码收录汉字 20000 多）；
b) 汉语的词与词之间没有空格分开，也可以说，从形式上看，在汉语中“词”不是一个良定义（well-defined）的语言单位；
c) 汉语词没有形态上的变化（或者说形态变化非常弱），同一个词在句子中充当不同语法功能时，形式是完全相同的；
d) 汉语句子没有形式上唯一的谓语中心词。

从语言单位的层次来看，汉语中语言单位各层次之间的界限非常模糊。一般而言，语言单位从小到大可以划分成语素、词、短语、句子等几个层次。在英语中，这些层次的划分是非常清楚的，对于任意给定的语言单位，很容易归入到以上几个层面之中去，而很少出现不易判定的情况。而在汉语中，除了汉字是一个稳定的、有明确定义的语言单位之外，上述的几个层次的语言单位都没有明确的定义，互相之间的界限也非常模糊[刘群 1998]。从词这一个层次看，首先，词和语素的界限并不分明。汉语的语素基本上都是汉字，而大部分汉字在作为语素的同时，本身又可以独立成词，这也是汉语未定义词识别困难的主要原因。其次，词和短语的界限也很模糊。例如，人们通常认为“牛肉”是一个词，而“鸡肉”是否是一个词却通常很难判断；“洗澡”是一个词，因为“澡”是一个语素，并不能单独使用，但是却可以说“洗了一个澡”，这显然是个短语。

[朱德熙 1985]中指出了汉语语法上的两个很重要的特点，第一，在汉语中，词的构成原则和短语的构成原则基本一致，第二，由于汉语缺乏明显的形态变化，导致汉语的词类和语法功能之间缺少直接的一一对应关系。这就导致在汉语中很难像英语等语言一样采用简单的有限状态语法来处理词法分析问题，而必须引入更复杂的工具，特别是统计方法。就问题的复杂性而言，作者认为汉语的词法分析的问题性质和难度与英语的基本短语分析的问题性质和难度大体相同，而大大高于英语的词法分析的问题和难度。

有一种观点认为，既然汉语的词语这么不确定，那么在机器翻译中是否可以不进行词法分析，直接进行句法分析呢？实际并不是这样。因为如果这样做的话，会导致句法分析的搜

索空间急剧膨胀，以致无法承受。实际上，汉语文本中未定义词只占一小部分，绝大部分词都是可以在词典中找到的，如果这些词都要从头开始分析，势必给句法分析带来太多的负担。

3.1.2 汉语词法分析的任务和前人的工作

汉语词法分析包括以下几个任务：

1. 词典查询：词典查询算法对汉语词法分析的效率有一定影响。
2. 处理重叠词、离合词、前后缀：重叠词、离合词、前后缀都是常见的汉语构词方式。
3. 切分排歧：通常把汉语的词语切分歧义分成组合歧义（又称覆盖歧义）和交集歧义（又称交叉歧义）两种类型，更复杂的歧义现象通常都可以分解为这两种歧义类型的组合。
4. 未定义词识别，具体包括：
   a) 时间词、数词处理
   b) 中国人名识别
   c) 中国地名识别
   d) 译名识别
   e) 其他专名识别
   f) 缩略语识别
   g) 新涌现的通用词或专业术语等
5. 词性标注：给汉语词语标注词性（Part-of-Speech）。

上述任务中，词典查询和重叠词、离合词、前后缀的处理都比较简单。汉语词典查询算法方面可参考[孙茂松 2000]和[Ng & Lua 2002]；重叠词和前后缀的处理可以采用一般的规则方法。离合词的处理稍微复杂一些，可以采用根据离合词中的非自由语素字对上下文进行检测的方法。[白硕 1998]一文对离合词的处理提出了一条有益的思路。

汉语切分歧义的消解是一个研究得非常多的问题。常见的汉语切分歧义消解策略包括基于词典的机械切分方法、基于规则的方法和基于统计的方法。基于词典的机械切分方法包括最大匹配方法、最短路径法等等。其中最大匹配方法、最短路径方法都是基于词典的机械方法，在这类方法中，[王显芳 2001a]提出的交叉歧义检测法可以快速将句子中的交叉歧义和覆盖歧义进行分离，有利于在后续的算法中对这两种类型的歧义采用不同的方法分别处理。在排歧过程中，基于规则的方法是很常见的。这方面通过大规模语料库进行错误驱动的转换规则学习[Hockenmaier 1998]比纯粹通过人工的方法编写规则更加有效，还有的研究者甚至引入了深层的句法分析[Wu 1998]来解决词语切分问题。在统计方法中，基于记忆的交叉歧义排除法[孙茂松 1999]、n 元语法[孙茂松 1997][王显芳 2001b][高山 2001]、分类器的方法[Xue 2002]都是通过大规模语料库进行有指导的训练的方法，而最大压缩方法[Teahan 2000]、无词典的自监督学习方法[Peng 2001]则是无指导的训练方法。从理论上看，不使用词典的无指导方法好像可以一揽子解决切分排歧和未定义词识别问题。不过在应用中，由于这种方法放弃使用蕴涵丰富专家知识的词典信息，实际上往往很难取得理想的效果。

汉语未定义词识别的研究也很多，有些是专门针对某一类的未定义词识别，如人名识别[李建华 2000][孙茂松 1995][张俊盛 1992][郑家恒 1993][宋柔 1993]、地名识别[沈达阳 1995]、译名识别[孙茂松 1993]、机构名识别[张艳丽 2001]等等，有的是不针对某一特定类型的通用的未定义词识别方法[Wu 2000] [Chen 2002][Ye 2002][Sun2002]。从方法上来说，无外乎基于规则的方法、基于统计的方法两种。基于规则的方法一般需要使用一些特征字和特征词来触发和识别某类特定的未定义词，比如采用中国人的姓氏、称呼来触发中国人名的识别，利用中国人名和地名的用字规律来判断人名和地名，或者对初始切分后落单的汉字来触

发未定义词识别模块等等。规则一般来源于观察到的语言现象或者是大规模的专名库。采用规则方法的未定义词识别中也经常引入各种统计策略[宋柔 1993][Wu 2000]来改善识别的效果。基于大规模语料库进行统计的未定义词识别方法[Ye 2002][Sun 2002]往往也可以达到较好的效果。

词性标注也是研究得比较充分的一个课题。总体上说，汉语的词性标注和英语的词性标注在方法上没有明显的不同。在有大规模标注语料库的情况下，很多方法（特别是统计方法）都可用于解决词性标注问题，而且结果通常也都很好。比较典型的方法包括隐马尔可夫模型（HMM）[Marshall 1983] [Steven 1988]和基于转换的错误驱动的学习方法（TBL）[Brill 1995]。

不过，汉语词法分析的更大困难在于，以上这些问题并不是孤立存在的，而是互相交织在一起的。一些常见的歧义交织现象包括：

1．汉语未定义词和上下文构成交集歧义：如“克林顿对内塔尼亚胡说”，未定义词“内塔尼亚胡”和“胡说”构成交集歧义；“费孝通向人大常委会提交书面报告”，未定义词“费孝通”和“通向”构成交集歧义；
2．汉语未定义词本身成词：如“王国维”、“高峰”、“汪洋”、“张朝阳”这些人名中，“王国”、“高峰”、“汪洋”、“朝阳”本身都可成词；
3．汉语未定义词类型歧义：如“河北省刘庄”，这里“刘庄”是地名，但很容易被识别为人名。

虽然对于汉语词法分析的各个独立问题研究已经很多，但由于这些问题是互相交织在一起的，因此对于一个完整的汉语词法分析系统，如果把这些问题分阶段逐个处理，显然无法达到理想的效果。如何在一个集成的框架下解决所有这些问题，并达到一个整体的最优效果，是一个非常重要的课题。[Sproat 1994]提出了一种基于随机有限状态自动机的汉语词法分析方法，不过这种方法中对于未定义词的识别策略过于简单，而且由于当时条件的限制也没有充分利用大规模的语料库进行训练。[Sun 1997]提出了一种基于 Agent 的汉语词法分析框架。不过在这种框架下词法分析的各个模块之间只是一种松散的耦合，多种未定义词之间的竞争采用的一种经验性的评分策略，缺乏严格的统计模型支持。 [孙健 2002]提出了一种基于类的语言模型作为汉语词法分析的整体框架，取得了较好的效果，不过由于这种模型的参数空间太大，数据稀疏问题较为严重，搜索空间太大，系统性能较低。

本文提出了一种基于层叠隐马尔可夫模型（Cascaded HMM）的方法，旨在将汉语分词、切分排歧、未登录词识别、词性标注等词法分析任务融合到一个相对统一的理论模型中。首先，在预处理的阶段，采取 N-最短路径粗分方法，快速的得到能覆盖歧义的最佳 N 个粗切分结果；随后，在粗分结果集上，采用底层隐马尔可夫模型识别出普通无嵌套的人名、地名，并依次采取高层隐马尔可夫模型识别出嵌套了人名、地名的复杂地名和机构名；然后将识别出的未登录词以科学计算出来的概率加入到词汇化的细切分隐马尔可夫模型中，未登录词与歧义均不作为特例，与普通词一起参与各种候选结果的竞争。最后在全局最优的分词结果上进行词性的隐马尔可夫模型标注。该方法取得了很好的分词和标注效果。采用该方法的系统在 2002 年国家 973 专家组机器翻译第二阶段的评测和 2003 年 5 月 SIGHAN 举办的第一届汉语分词大赛中名列前茅，取得了很好的成绩。

3.2 汉语词法分析的层叠隐马尔可夫模型

3.2.1 隐马尔可夫模型简介

隐马尔可夫模型是一种在自然语言处理等领域中被广泛应用的统计模型。由于该模型是

层叠隐马尔可夫模型的基础，因此这里对其做一简单介绍。更多的细节和算法请参见[Rabiner 1986,1989]。

假设有一个观察值序列$W = w_1...w_n$（比如词语序列），同时有一个状态值序列$T = t_1...t_n$（比如词性标记或角色标记序列），观察值和状态值一一对应。

隐马尔可夫模型有如下三个假设：

假设 1：马尔可夫假设（状态构成一阶马尔可夫链）：$p(t_i \mid t_{i-1}...t_1) = p(t_i \mid t_{i-1})$

假设 2：不动性假设（状态与具体时间无关）：$p(t_i \mid t_{i-1}) = p(t_j \mid t_{j-1})$，对任意 i,j 成立

假设 3：输出独立性假设（输出仅与当前状态有关）：$p(w_1...w_n \mid t_1...t_n) = \prod_{i=1}^{n} p(w_i \mid t_i)$

隐马尔可夫模型的逻辑结构如下图所示：

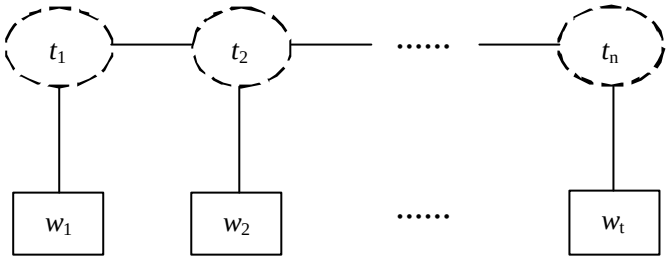


图 3.1 隐马尔可夫模型的逻辑结构

一个隐马尔可夫模型被定义为一个五元组：$l = (\Omega_T, \Omega_W, \mathrm{A}, \mathrm{B}, \Pi)$

其中：

$\Omega_T = \{q_1...q_S\}$：状态的有限集合

$\Omega_W = \{v_1...v_R\}$：观察值的有限集合

$\mathrm{A} = \{a_{jk}\}, a_{jk} = p(t_{i+1} = q_j \mid t_i = q_k)$：转移概率矩阵

$\mathrm{B} = \{b_{lk}\}, b_{lk} = p(w_i = v_l \mid t_i = q_k)$：输出概率矩阵

$\Pi = \{p_k\}, p_k = p(t_1 = q_k)$：初始状态分布

有时为了表示方便起见，引入一个唯一的起始状态 q_0，于是可以记：

$$p_k = p(t_1 = q_k \mid t_0 = q_0) = a_{k0}$$

这样就将初始状态分布Π合并到了状态转移矩阵 A 中。

隐马尔可夫模型需要求解三个基本问题：

评估问题： 对于给定模型，求某个观察值序列的概率 $p(W \mid l)$ ；（语言模型）

解码问题： 对于给定模型 l 和观察值序列W，求可能性最大的状态序列T ；

学习问题： 对于给定的一个观察值序列W，调整参数l，使概率$p(W \mid l)$最大。

对于评估问题，根据隐马尔可夫模型的三个假设，可以得到：

$$p(W)=\sum_{T} p(W \mid T)\, p(T)=\sum_{T}\prod_{i=1}^{n} p(w_i \mid t_i)\, p(t_i \mid t_{i-1})$$

评估问题可以采用一种特定的动态规划算法——向前算法解决，时间复杂度为：$O(S^2N)$，其中S为状态值集合的大小，N为观察值序列长度；

对于解码问题，我们有：

$$T^{*}=\arg\max_{T} p(T \mid W)=\arg\max_{T} p(W \mid T)\, p(T)=\arg\max_{T}\prod_{i=1}^{n} p(w_i \mid t_i)\, p(t_i \mid t_{i-1})$$

解码问题也可以采用一种动态规划算法——Viterbi 算法来解决，时间复杂度同样是$O(S^2N)$。

对于学习问题，在没有已标注的数据的情况下，可以采用一种特定的 EM 算法——向前向后算法（Baum-Welch 算法）来解决。如果已经有已标注的数据，可以直接用最大似然估计来求解模型参数。

3.2.2 层叠隐马尔可夫模型的结构

隐马尔可夫模型具有很强的表达和处理能力。n 元语法可以看成是隐马尔可夫模型的一种退化情形（状态值和输出值一一对应，输出概率都是 1）。汉语词法分析中的切分排歧、未定义词识别和词性标注都可以用隐马尔可夫模型或 n 元语法来解决。本文提出的层叠隐马尔可夫模型（Cascaded Hidden Markov Model，简称 Cascaded HMM）就是试图用统一的隐马尔可夫模型来解决汉语词法分析中的各个问题，并在这些隐马尔可夫模型中建立起一定的联系，以形成一个一体化的汉语词法分析系统框架。

基于层叠隐马尔可夫模型（Cascaded HMM）的汉语词法分析流程如下图所示。具体说明如下：

1．整个算法采用词图作为核心的数据结构；

2．整个词法分析过程分为多个阶段，每个阶段都采用隐马尔可夫模型作为基本统计模型；

(1) 每一层隐马尔可夫模型都采用改进的 Viterbi 算法（N-Best）,输出最好的若干个结果作为下一级隐马尔可夫模型的输入；

(2) 每一层隐马尔可夫模型的状态值和输出值各不相同，这些状态值、输出值和词图的边之间都存在一个映射过程；

3．这个词法分析过程分为五级隐马尔可夫模型，自底向上分别称为第一级到第五级隐马尔可夫模型，它们分别是：

(1) 第一级：粗切分采用基于一元语法的 N 最短路径方法；

(2) 第二级：简单未定义词识别采用基于角色的隐马尔可夫模型；

(3) 第三级：复合未定义词识别也采用基于角色的隐马尔可夫模型；

(4) 第四级：细切分采用词汇化的隐马尔可夫模型；

(5) 第五级：词性标注采用基于词性的隐马尔可夫模型。

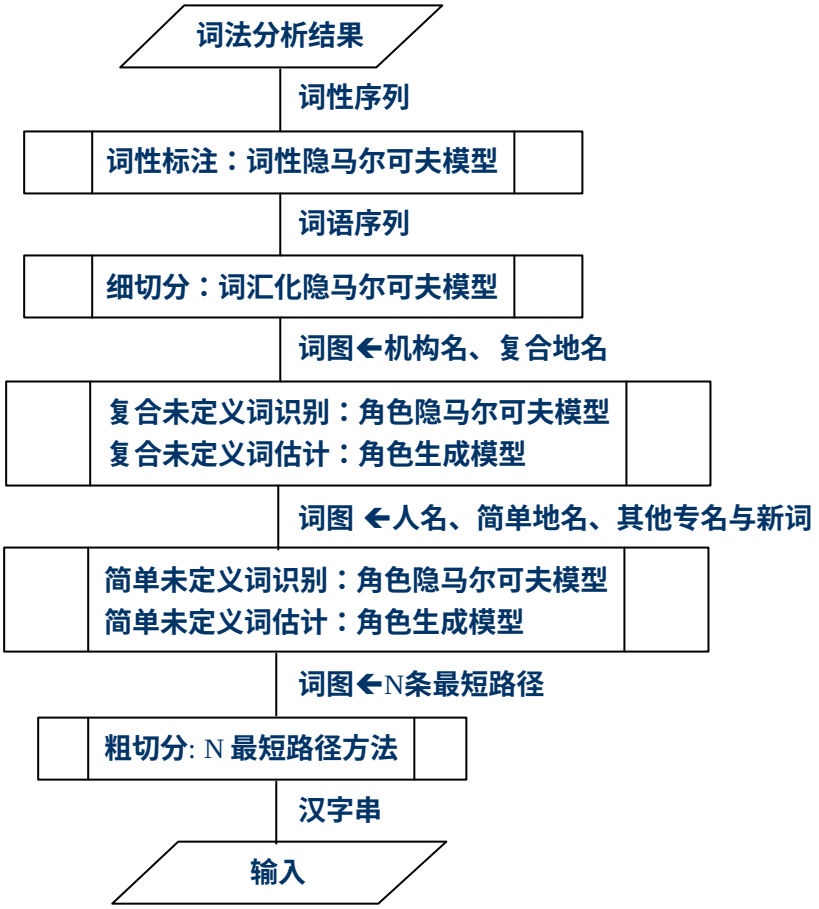


图3.2 基于层叠隐马尔可夫模型的汉语词法分析流程图

3.2.3 层叠隐马尔可夫模型的核心数据结构——词图

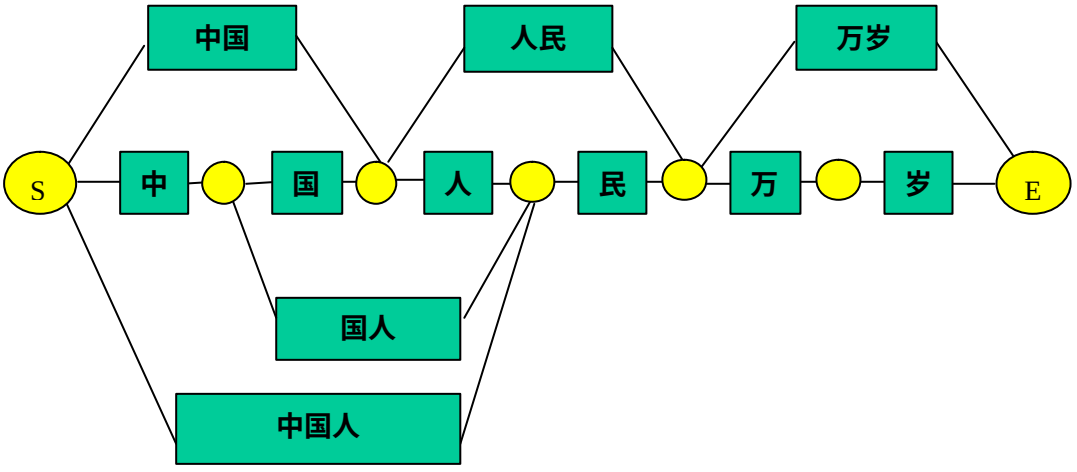


图 3.3 词图

基于层叠隐马尔可夫模型的汉语词法分析算法的核心数据结构是词图（Word Graph）。词图是一种在汉语词法分析中普遍采用的数据结构，类似于语音识别中的词格（Word Lattice）和句法分析中的线图（Chart）。在一个词图中，结点是汉字之间的位置，连接两个

结点的边是一个候选的词语，如上图所示。

3.2.4 层叠隐马尔可夫模型的参数训练

对于每一层的隐马尔可夫模型，我们都采用有指导的训练方法。为此，我们需要一个已标注各层角色标记的语料库。

我们所使用的语料库是北京大学计算语言学研究所研制的《人民日报》标注语料库，该语料库含《人民日报》1998 年上半年的全部语料并对其进行了词语切分和词性标注。为了将该语料库用于每个层次的隐马尔可夫模型，我们需要对该语料库进行改造，将语料库中的词性标注改造为相应模型的标注。大部分情况下，这种改造都是可以自动进行的，个别情况下需要进行很少量的人工校对。

3.3 粗切分：基于一元语法的 N 最短路径方法

在整个词法分析过程中，要进行两个阶段的切分排歧。第一个阶段的切分排歧位于未定义词识别之前，称为粗切分，采用基于一元语法的 N 最短路径方法。第二个阶段的切分排歧谓语未定义词识别之后，称为细切分，采用词汇化的隐马尔可夫模型。

粗切分的目的主要有两个：一是词典词（指词典中收录的词，是一个与未定义词相对的概念）的召回率要高，二是速度要快。为此我们选择了简单的一元语法作为算法框架。一元语法只需使用词频信息，存储空间小，查询速度快。切分路径的概率仅仅是路径上各个词语词频的乘积，计算非常简单，搜索效率高。

具体来说，包括以下工作：

1．词典查询并处理离合词、重叠词和前后缀；

2．根据一元语法选择概率最大的 N 条路径（称为 N 最短路径）；

3．将所得路径上的所有词语加入词图。

常用切分算法往往过于武断，过早地在初始阶段做出是否切分的判断，只保留一个自己认为最优的结果，而这一结果往往会因为存在歧义或未登录词而出错，这时候，后期补救措施往往费时费力，效果也不会很好。而我们的粗切分方法在保持高效率的前提下，通过保留少量大概率的粗分结果，可以最大限度地保留各种歧义字段和未登录词的边界，取得了很好的效果。

下表给出了测试语料为 200 万汉字时 N 最短路径（N=8）与常用算法在切分结果包容歧义方面的对比测试结果。其中“切分最大数”指的是被测试的句子可能的最大切分结果数。“切分平均数”指的是单个句子平均的切分结果数。“正确切分覆盖率”指的是正确切分中的所有词典词在粗切分结果全部被召回的句子比例。

表 3.1 N 最短路径(N=8)与常用算法对比

方法	切分最大数	切分平均数	正确切分覆盖率
最大匹配	1	1	85.46%
最少切分	1	1	91.80%
最大概率	1	1	93.50%
全切分	>3,424,507	>391.79	100.00%
8-最短路径	8	5.82	99.92%

由此可见，基于一元语法的 N 最短路径粗切分完全能够满足我们的要求，为下一步的工作打下了很好的基础。

3.4 未定义词识别：基于角色的隐马尔可夫模型

3.4.1 模型的定义

汉语未定义词识别同样分成两个阶段：第一个阶段是简单未定义词识别，第二个阶段是复合未定义词识别。简单未定义词包括人名、简单地名、其他专名和新词语。复合未定义词包括机构名和复合地名。复合未定义词中可以包含简单未定义词作为组成部分。

对于所有的未定义词识别，我们采用一种统一的策略：基于角色的隐马尔可夫模型（简称角色隐马尔可夫模型）。在这个模型中，观察值是经过前一阶段切分产生的词串，状态值是我们称之为“角色”的标记。对于不同类型的未定义词识别，我们要采用不同类型的角色标记。所以实际上，在未定义词识别的两个阶段上，每个阶段各有若干个基于角色的隐马尔可夫模型在同时工作。

在第一阶段，我们有以下几个基于角色的隐马尔可夫模型：

1. 数词识别；
2. 时间识别；
3. 中国人名识别；
4. 日本人名识别；
5. 简单中国地名识别；
6. 音译人名识别；
7. 音译地名识别；
8. 其他未定义词识别。

在第二阶段，我们有以下两个基于角色的隐马尔可夫模型：

1. 复合中国地名识别；
2. 机构名识别。

3.4.2 角色的选取

基于角色的隐马尔可夫模型中，角色标记集的选取是非常关键的。

角色标注的做法，类似于组块分析中给词语标注 B、E、I、O（分别表示组块的开始词语、结束词语、内部词语、外部词语）几个标记的做法。对于未定义词识别来说，这种标记过于简单，不易取得好的效果。为此，我们需要根据每种特定的未定义词类型定义专门的角色标记集。

角色标记集的选取，需要根据所要识别的未定义词类型，结合专家知识，科学地设定。然后再通过实验不断调整，某些角色需要合并，某些角色可能需要细分，反复实验调整，最后才能得到一个理想的角色标记集。

以中国人名识别为例，我们最后使用的角色标记集是：

表 3.2 中国人名识别的角色标记集合

角色	意义	例子
B	姓氏	张华平先生

C	双名的首字	张华平先生
D	双名的末字	张华平先生
E	单名	张浩说：“我是一个好人”
F	前缀	老刘、小李
G	后缀	王总、刘老、肖氏、吴老师、叶帅
K	人名的上文	又来到于洪洋的家。
L	人名的下文	新华社记者黄文摄
M	两个中国人名之间的成分	编剧邵钧林和稽道青说
U	人名的上文和姓成词	这里有关天培的壮烈
V	人名的末字和下文成词	龚学平等领导，邓颖超生前
X	姓与双名的首字成词	王国维、
Y	姓与单名成词	高峰、汪洋
Z	双名本身成词	张朝阳
A	以上之外其他的角色	

其中，V 和 X 两个标记非常特殊，主要用于处理中国人名和上下文构成交集歧义的情况。

再看看我们使用的机构名识别的角色标记集：

表 3.3 机构名识别的角色标记集合

角色	意义	例子
A	上文	参与亚太经合组织的活动
B	下文	中央电视台报道
X	连接词	北京电视台和天津电视台
C	一般性前缀	北京电影学院
F	译名性前缀	美国摩托罗拉公司
G	地名性前缀	交通银行北京分行
H	机构名前缀	中共中央顾问委员会
I	特殊性前缀	中央电视台
J	简称性前缀	巴政府
D	后缀（机构名特征词）	国务院侨务办公室
Z	其他非机构名成份	

在这个标记集中，一些简单地名和音译名可以被用作其组成部分。因此机构名识别作为符合未定义词识别，在简单未定义词识别完成之后进行。

3.4.3 角色的标注

假设输入的词语串是 $W=w_1\ w_2\ \ldots\ w_n$，相应的角色标记序列是 $T=T=t_1t_2\ldots\ t_n$，根据隐马尔可夫模型的定义，我们可以得到最佳角色标记序列 T^* 的计算公式：

$$T^* = \arg\max_T p(T|W) = \arg\max_T p(W|T)p(T)$$

$$= \arg\max_T \prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_{i-1})$$

角色标注的过程就是隐马尔可夫模型的解码问题的求解过程。使用常用的 Viterbi 算法，可以得到最佳的角色标注集。

还是以中国人名识别为例，我们利用前面的角色表对下面的句子进行角色标注：

馆 内 陈列 周 恩 来 和 邓 颖 超生 前 使用 过 的 物品 。

其结果为：

馆/A 内/A 陈列/K 周/B 恩/C 来/D 和/M 邓/B 颖/C 超生/V 前/A 使用/A 过/A 的/A 物品/A 。/A

在中国人名识别的这个例子中，我们要对两个标记做特殊处理：U 和 V。因为这两个标记分别表示人名和上下文成词的情况，因此我们需要把这两个标记对应的词进行分裂：U 分裂成 KB，V 分裂成 DL（左侧标记为 C）或 EL（左侧标记不为 C）。经过这样处理，上面的句子标记改为：

馆/A 内/A 陈列/K 周/B 恩/C 来/D 和/M 邓/B 颖/C 超/D 生/L 前/A 使用/A 过/A 的/A 物品/A 。/A

3.4.4 未定义词的提取

得到词语序列的角色标记后，还需要根据这些角色标记提取未定义词。对于某一类特定的未定义词，我们需要根据其角色标记集定义一个模式集（Pattern Set），其中每一个模式是一个角色标记的序列。还是以中国人名的标记为例，我们定义的模式集为：{BBCD, BBE, BBZ, BCD, BEE, BE, BG, BXD, BZ, CD, EE, FB, Y, XD}。下面给出一些常见模式的例子：

表 3.4 中国人名识别的模式集合

模式	例子	模式	例子
BBCD	陈方安生	BZ	张朝阳
BBE	陈鲁豫	CD	建华
BCD	董建华	FB	老李
BE	李鹏	Y	高峰
BG	李总	XD	王国维

在完成对输入词语序列的角色标注后，我们可以对其角色标记序列进行搜索，一旦发现一个子序列和某一个模式能够匹配，就可以将其识别为一个该类型的未定义词。如果发现多个模式能够同时被匹配，我们选择最长的模式。

于是在上面的例子中，我们识别出了两个人名：周恩来和邓颖超，模式都是 BCD。

3.4.5 参数训练

角色标注模型的参数训练包括两部分：一是隐马尔可夫模型的参数训练，二是模式集的

提取。

如前所述，我们利用《人民日报》标注语料库，首先进行标记的转换，然后进行训练。

例如：语料库中的有这样一个句子：

政务司/n　司长/n　陈/nr　方/nr　安生/nr　出任/v　委员会/n　主席/n

转换后得到：

政务司/A 司长/K 陈/B 方/B 安/C 生/D 出任/L 委员会/A 主席/A

在人名识别的例子中，我们还要处理的一个特殊情形是人名和上下文成词的问题。为此我们可能需要将人名的首字和尾字与上下文合并，产生一个新的标记（U 或 V）。

有了训练语料库后，隐马尔可夫模型的参数训练非常简单，采用最大似然估计（MLE），直接利用相应的频率进行计算即可：

$$p(t_{i+1} \mid t_i) = \frac{c(t_i, t_{i+1})}{c(t_i)}$$

$$p(w_i \mid t_i) = \frac{c(w_i, t_i)}{c(t_i)}$$

这里 $c(t_i)$是语料库中角色标记 t_i 出现的次数，$c(t_i,t_{i+1})$是角色标记 t_i和 t_{i+1} 相邻出现的词数，$c(w_i,t_i)$是词语 w_i 标记为角色 t_i 的次数。

模式集的提取也很容易。由于《人民日报》语料库中对大部分未定义词都给出了相应的标记，只要将这些词对应的角色标记序列提取出来即可。不过对于某些未定义词的类型还是要进行人工判别，比如人名中，中国人名、日本人名和音译人名在原始语料库中是没有区分的，对此我们进行了某种人机互助的判别，将这三种人名加以区分以训练三种不同的角色隐马尔可夫模型。

3.5 未定义词的概率估计：基于角色的词语生成模型

3.5.1 问题的由来

如果仔细研究上面的基于角色的隐马尔可夫模型就会发现一个问题。在复合未定义词识别过程中，如果用到了某个简单未定义词作为其组成部分，假设这个简单未定义词是 w_i，其相应的角色标记是 t_i，我们会发现输出概率 $p(w_i|t_i)$是无法从语料库中统计得到的，这是因为 w_i 本身是个未定义词，肯定没有在词典和语料库中出现过。为此我们需要引入一个新的模型来估计这个输出概率，我们称这个模型为基于角色的词语生成模型（简称角色生成模型）。

实际上，基于角色的词语生成模型和基于角色的隐马尔可夫模型是一一对应的。对于每一个角色隐马尔可夫模型而言，都需要一个相应的角色生成模型，用于计算其所识别出的未定义词的输出概率。而且，这两个模型之间可以共享某些参数。

3.5.2 模型的定义

假设我们有一个基于角色的隐马尔可夫模型，其对应的角色标记集为 $T=\{t_1,\ldots,t_n\}$，模式集为：$PatternSet(Type)=\{PT_1,\ldots,PT_m\}$。同时假设这个模型识别出了一个未定义词

$W=w_1w_2\ldots w_l$，而且 W 所对应的模式为 $PT_k=t_{k1}t_{k2}\ldots t_{kl}$，其中 t_{ki} 为模式 PT_k 中的第 i 个标记。于是我们有：

$$p(W \mid Type) = p(PT_k \mid Type)p(W \mid PT_k)$$

这里 $p(PT_k)$是在这种类型的未定义词识别中，模式 PT_k 被使用的概率。$p(PT_k)$满足归一性假设：

$$\sum_{k=1}^{m} p(PT_k \mid Type) = 1$$

根据对应的角色隐马尔可夫模型的输出独立性假设，我们得到：

$$p(W \mid Type) = p(PT_k \mid Type)\prod_{i=1}^{l} p(w_i \mid t_{ki})$$

我们可以看到，这个模型中总共有两个参数：$p(PT_k|Type)$和 $p(w_i|t_{ki})$。其中 $p(w_i|t_{ki})$就是对应的角色隐马尔可夫模型中的输出概率，而 $p(PT_k|Type)$可以从语料库中统计得到。

根据上述模型，人名识别中“陈鲁豫”的生成概率计算如下：

$$p(\text{陈鲁豫}|\text{中国人名}) = p(\text{BBE}|\text{中国人名})\, p(\text{陈}|\text{B})\, p(\text{鲁}|\text{B})\, p(\text{豫}|\text{E})$$

基于角色的词语生成模型除了在复合未定义词的识别过程中要被使用以外，在词法分析后面的各个阶段中还要被多次用到，下面我们介绍时会分别提到。

3.6 细切分：词汇化的隐马尔可夫模型

3.6.1 模型的定义

细切分在未定义词识别完成之后尽心，是第二遍也是最后一遍切分排歧。这时候，所有的候选的词典词和未定义词都已加入到词图中，细切分的目的就是选取一条从词图起点到词图终点的最优路径。

为了对各种切分路径进行比较，我们需要计算每一条切分路径的概率值。这个概率值我们还是用一个隐马尔可夫模型来计算，这个隐马尔可夫模型我们称之为词汇化的隐马尔可夫模型（Lexicalized HMM）。在这个模型中，观察值就是每一个词语，状态值是一种类别标记，具体分为两类：对于词典词而言，类别标记就是该词本身；而对于未定义词而言，类别标记是该未定义词的类型。

具体来说，类别标记定义如下：

表 3.5 词汇化隐马尔可夫模型的标记集

类别标记	解释
w_i（词语本身）	词典词
PER	未定义词：人名
LOC	未定义词：地名
ORG	未定义词：机构名
NUM	未定义词：数词

TIME	未定义词：时间
OTHER	未定义词：其他类别
START	句子开始
END	句子结束

给定一个分词原子序列 S，S 的某个可能的分词结果记为 $W=(w_1,\cdots w_n)$, W 对应的类别标记序列记为 $C=(c_1,\cdots c_n)$。我们取概率最大的分词结果$(W,C)^*$作为最终的分词结果：

$$(W,C)^* = \underset{W,C}{\operatorname{argmax}}\, P(W,C)$$

利用贝叶斯公式进行展开，得到：

$$(W,C)^* = \underset{W,C}{\operatorname{argmax}}\, P(W|C)\,P(C)$$

将词类看作状态，词语作为观测值，利用一阶 HMM 展开得到：

$$(W,C)^* = \underset{W,C}{\operatorname{argmax}} \prod_{i=1}^{n} p(w_i \mid c_i)\,p(c_i \mid c_{i-1})$$

3.6.2 最短路径的求解

在下图中，我们给出了“毛泽东 1893 年诞生”的切分路径选择图。最终所求的分词结果就是从初始节点 S 到结束节点 E 的最短路径，这是个典型的最短路经问题，可以采取动态规划算法快速求解。

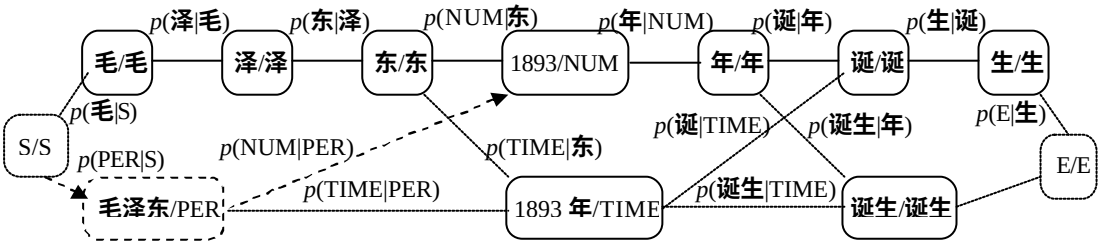


图 3.4 词汇化隐马尔可夫模型切分路径选择图
（原始字串为“毛泽东 1893 年诞生”）

说明：1. 节点中表示的是“词语/类”(即 w_i / c_i),节点的权值为类到词语的概率 $p(w_i \mid c_i)$;

2. 有向边的权值为相邻类的转移概率 $p(c_i \mid c_{i-1})$; S 为初始节点；E 为结束节点

3. “毛泽东/PER”相关的虚线部分是人名识别 HMM 作用过之后产生的。

3.6.3 参数估计

根据隐马尔可夫模型的定义，我们需要对状态转移概率矩阵 $p(c_i|c_{i-1})$和输出概率矩阵 $p(w_i|c_i)$进行参数估计。

状态转移概率 $p(c_i|c_{i-1})$可以通过语料库的训练得到。通过将《人民日报》语料库的词性标记转换为上面定义的类别标记，我们很容易计算出各个状态之间的转移概率。

输出概率 $p(w_i|c_i)$的估计分为两种情况。根据类别 c_i 的定义，如果 w_i 在核心词典收录，可以得到 $c_i=w_i$，因此 $p(w_i|c_i)=1$。如果 w_i 是一个未定义词，那么我们就不可能通过语料库

的统计来获得这个输出概率，因为这个词在训练语料库中很可能从来没有出现过。这时我们可以使用前面介绍的基于角色的词语生成模型来进行估计。

3.7 词性标注：基于词性的隐马尔可夫模型

3.7.1 基于隐马尔可夫模型的词性标注

词性标注我们采用基于词性的隐马尔可夫模型，其中隐含的状态就是词性标记。

假设输入的词语串是 $W=w_1 w_2 \ldots w_n$，相应的词性序列是 $T=T=t_1 t_2 \ldots t_n$，根据隐马尔可夫模型的定义，我们可以得到最佳词性序列 T^* 的计算公式：

$$T^* = \arg\max_T p(T|W) = \arg\max_T p(W|T)p(T)$$

$$= \arg\max_T \prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_{i-1})$$

词性标注的搜索算法还是采用 Viterbi 算法，这里不再详细介绍。

在参数训练中，唯一的问题还是对未定义词的输出概率的估计问题。我们仍然采用前面介绍的基于角色的词语生成模型来估计未定义词的输出概率，这里也不再重复。

3.7.2 词性标记集的选择与转换

3.7.2.1 对现有词性标记集的分析

汉语词性问题，一直是汉语语言学界争论的焦点之一，到目前虽已形成一些初步的共识，但也没有完全达成一致。这反映在自然语言处理上，就是词性标记集也没有形成统一的标准。目前已经有了一些影响比较大的词性标记集，如北京大学《人民日报》语料库所采用的词性标记集和教育部语言文字应用研究所制订的信息处理用现代汉语词类及词性标记集规范（目前已经被批准为国家推荐标准）。

我们在机器翻译的研究中发现，现有的这些词性标记集都或多或少存在各种问题，难以满足机器翻译研究的需要。为此，我们决定制订一个自己的词性标记集。为了制订一个比较科学的、适合汉外机器翻译研究的汉语词性标记集，我们对已有的一些比较著名的汉语词性标记集进行了详细的分析和比较。这些词性标记集包括：

1．北京大学《人民日报》标注语料库词性标记集（以下简称 PKUPOS）[俞士汶 1999，2000]；

2．清华大学《汉语树库》词性标记集（以下简称 THUPOS）[清华大学 1998]；

3．教育部语言文字应用研究所词性标记集（以下简称 YYSPOS）[语用所 2002]；

4．宾州大学汉语树库词性标记集（以下检查 PennPOS）[Xia 2000]。

通过对这些词性标记集的研究，我们发现其中存在的不适合于机器翻译的主要问题有：

第一，没有全面贯彻“按照词的语法功能进行分类”的原则。

[朱德熙 1985]明确提出，“划分词类的根据只能是词的语法功能”，这一论断无疑是十分正确的，特别是对机器翻译来说。PKUPOS 中的简称略语 j、习用语 l、成语 i 三个标记、YYSPOS 中的缩略语 j、习用语 l 两个标记，都是与功能完全无关的标记，这必定会给机器

翻译后续的句法分析等工作带来极大的困扰，即使对于汉语词性标记任务本身，这种划分也是非常不利的，会导致词性标记正确率的降低，因为这种标记在与上下文组合的时候区分性很差。在这一点上，THUPOS 和 PennPOS 做得比较好，这可能是因为这两套标记都是为句法树库的构建服务的，更多地考虑了句法分析的需求吧；

第二，“动词名物化”及类似问题的处理。

[朱德熙 1985]明确反对所谓动词名物化的说法并给出了科学的论证。其实这一观点在语言学界已被普遍接受。不过在自然语言处理中人们还是自觉或不自觉地采用‘动词名物化”及类似的做法。[周锡令 2003]明确提出对朱德熙先生观点的质疑，PennPOS 完全采用了“动词名物化”及相关的做法（在 PennPOS 中，甚至要表示区分修饰性定语的助词“的”和小句做定语的助词“的”，这在词法分析阶段是很难做到的）。PKUPOS 和 THUPOS 总体上采用了朱德熙先生的观点，但为了兼顾另外一方的观点，引入了名动词、副动词、名形词、副形词等二级标记，但对这几个标记的使用做了严格的限制。国内其他一些单位在词性标记集定义中虽然没有对这个问题做出要求，但在实际操作中往往还是采用了“动词名物化”一类的做法。对于这一问题，我们认为，从机器翻译实践的角度看，朱德熙先生的观点更为合理。实际上，单纯从操作角度看，采用这两种观点似乎各有利弊。如果采用“动词名物化”的做法，无疑对句法分析带来很大的好处，不过给词法分析造成了更大的困难。而采用朱德熙先生的做法，会为词法分析带来较大的方便，但会给句法分析增加更大的负担。不过从本质上说，一个动词是否处于句子主宾语的位置上，是应该由句法分析模块而不是由词法分析模块来判断的，因为在词法分析中，我们进行词性标记时一般仅仅考虑与该词相邻的几个词作为参考的依据，而不会考虑更多的上下文。而仅仅根据一个动词相邻的几个词来判断其是否处于句子主宾语的位置上显然是不合适的，这个问题只有放到句法分析中，把整个句子考虑进来才有可能做出比较准确的判断。实际上，我们的词法分析实验也说明了这一点。我们采用 PKUPOS 做词性标记实验时，各个标记的正确率分别为（测试语料库中有些标记没有出现，所以没有给出数据）：

表 3.6 各类词性标记的标注正确率

标记	词性	正确率	标记	词性	正确率
a	形容词	83.33%	h	前接成分	86.21%
b	区别词	94.14%	k	后接成分	97.25%
z	状态词	88.45%	i	成语	98.50%
s	处所词	95.21%	j	简称略语	89.77%
t	时间词	98.01%	l	习用语	85.10%
f	方位词	95.84%	v	动词	86.49%
r	代词	97.94%	vd	副动词	89.44%
d	副词	91.62%	vn	名动词	81.55%
m	数词	96.39%	p	介词	93.53%
q	量词	93.92%	u	助词	99.12%
n	名词	95.14%	c	连词	91.15%
nr	人名	93.04%	o	拟声词	68.25%
ns	地名	98.56%	e	叹词	71.43%
nt	机构团体	91.40%	y	语气词	95.21%
nz	其他专名	74.57%	x	非语素字	94.55%
Ng	名语素	0.47%	w	标点符号	99.77%

从这个表中可以看出，动词和形容词的标记正确率都比名词低，这是因为这两类词和其他类词的兼类现象比较严重。副动词的标记正确率是比较高的，这是由于汉语中副词的语法功能是非常单一的，只能作为动词和形容词的修饰词，所以比较容易判断。名动词的标记正确率比一般动词的标记正确率还要低将近 5 个百分点。这说明，要在词法分析阶段判断一个动词是否在句子中处于主宾语的位置确实是比较困难的。也就是说，采用“动词名物化”的做法虽然表面上看可以给句法分析带来方便，不过由于在词法分析阶段，“动词名物化”的标记正确率比正常的词性标记正确率低得多，因此实际上句法分析并不能真正享受到“动词名物化”标记所带来的好处，反而使得词法分析的标记正确率也有所下降。因此，我们认为，在词性标记时采用“动词名物化”之类的做法是不合适的。

第三，词类划分的颗粒度过粗。

语言学上给出的汉语词性划分，有些词性实际上是有一些语法功能相当不同的词组成的。最典型的代词和助词。代词的功能五花八门，统一归入到代词类中，对机器翻译是很不利的。助词也是一样，实际上，每个助词都有很强的个性，把他们简单地用一个标记来表示是不合适的。另外，标点符号的语法功能也相差很大，如果都使用相同的标记，也不太合理。我们做过一个简单的实验，在一个概率上下文无关语法（PCFG）中，如果为每一个标点符号定义一个单独的标记，比把所有的标点符号用同一个标记来表示，句法分析的标记正确率会提高大约 2～3 个百分点。同样，如果把介词中的“把”和“被”用单独的标记来表示，也可以提高句法分析的正确率。这就说明标记的选择对句法分析有很大的影响，如果能够在词性标记上将一些句法功能相差很大的词区分开来，对后面的句法分析是非常有利的。

3.7.2.2 词性标记集的制订

根据上面的分析，我们在词法分析系统中使用了一个自定义的词性标记集 ICTPOS（见附录）。这个词性标记集的定义主要考虑了以下几个原则：

1. 真正贯彻“按功能分类”的思想，这样做有助于提高汉语词法分析和句法分析的正确率；
2. 易于从北大《人民日报》语料库词性标记集进行转换；
3. 对于语法功能不同的词，在不造成词法分析和句法分析歧义区分困难的情况下，尽可能细分子类。

在具体做法上，我们是在 PKUPOS 的基础上进行改造和扩充，得到了一个含 22 个一级标记，77 个二级标记，总共 99 个标记的词性标记集。

我们对 PKUPOS 的改造主要包括：

1．删去了简称略语 j、习用语 l、成语 i、语素字 g 这几个大类，把这几个大类中的词按照其功能划分到其他词类中，在那些词类中增加一个二级标记来表示短语和语素；

2．对某些语法功能差异较大的词类，尽可能详细地划分子类，给每个子类一个二级标记。这些词类包括代词（10 个二级标记）、助词（15 个二级标记）和标点符号（16 个二级标记）。特别是对于助词类，我们几乎给每一个常见的一些助词都分派了一个二级标记。

3．动词类中，对一些语法功能比较特殊的动词或子类也给出了专门的二级标记，如动词“是”和“有”分别给出了一个二级标记，“趋向动词”、“补助动词”、“形式动词”、“内动词（不及物动词）”也都给出了单独的标记；

4．名词类中，对专有名词做了更详细的划分，区分了中国人名、姓氏、名字、日本人名、音译人名、地名、音译地名等；

5．其他一些细微的修改，比如将连词区分为并列连词和从属连词等。

在“动词名物化”的处理方面，我们照搬了 PKUPOS 中的做法，没有做大的改动，因为 PKUPOS 的做法总体上是符合朱德熙先生的观点的，这也是我们所认同的做法。

3.7.2.3 词性标记的转换

确定了词性标记集 ICTPOS 以后，我们需要进行两方面的词性标记转换：

一方面是训练语料库的词性标记转换，要将我们的训练语料库（《人民日报》语料库）中的词性标记从 PKUPOS 转换到 ICTPOS。

另一方面是系统运行时的词性标记转换，为了适应不同的需要（如评测的需要），在汉语词法分析系统输出的时候，要将 ICTPOS 的标记转换成输出所要求的标记。

运行时的转换比较简单，因为 ICTPOS 的标记比现有的各种标记集都更加详细，要转化成其他标记集大部分情况下是一种多对一的映射，通过一个映射表很容易实现。不过，如果应用所要求的词性标记体系和我们的体系相差较大（比如说采用‘动词名物化’之类的做法），那么这种转换就比较困难，效果不会太好。

训练语料库的词性标记转换比较复杂，因为这是一种一对多的映射。在这个转换中，针对不同的标记，我们使用了三种方法：

第一，采用‘PKUPOS＋具体词➔ICTPOS’的做法。这种做法之所以可行，是因为 ICTPOS 中的一些标记所对应的词语是一个封闭集合，如标点符号二级标记、助词二级标记、代词二级标记、动词“是”、动词“有”、趋向动词等；

第二，采用“PKUPOS＋汉语词语语法信息➔ICTPOS”的做法。这种做法利用了北京大学计算语言所开发的“汉语语法信息词典”[俞士汶 1998，2000]，例如“内动词”，就可以利用语法信息词典中“内外＝内”这个属性进行判断；对于语法信息词典中已有的成语、习用语、缩略语，利用语法信息词典中的信息，也都可以转换成相应的功能标记；

第三，采用人工校对的办法，主要是对一些“语法信息词典中”没有收录的词，需要采用这种方法进行转换。这种情况比较少，在全部的《人民日报》半年语料库中，只有几千个词（每次出现算一个词）需要进行这种转换，一个人花两天左右的时间就全部完成了。而且转换完成后将人工转换的这些词放入词典中，以后遇到类似的情形就可以自动处理了。

新的词性标记集 ICTPOS 比原有的 PKUPOS 多了一倍多的标记，可以认为，新标记所含有的信息量比原有标记要丰富得多，这种信息对于后续的语法分析等处理是非常有益的。而且我们的实验表明，采用含更多标记的新标记集后，汉语词法分析系统的词性标记正确率不仅没有下降，反而略有提高。这也达到了我们对词性标记集进行改造的初衷，就是提高系统的性能，更重要的是为后续的处理（特别是句法分析）提供更丰富的信息。

3.8 实验结果

我们开发了一个基于层次隐马尔可夫模型的汉语词法分析系统 ICTCLAS，下面介绍我们用 ICTCLAS 做的实验以及 ICTCLAS 参加国内外公开测试的结果。

3.8.1 各层隐马尔可夫模型的对比实验

我们使用北京大学计算语言学研究所加工的《人民日报》语料库进行了训练和测试。在人民日报 1998 年一月份共计 1,108,049 词的新闻语料库上，我们进行了如下四种条件下的性能测试：

1） BASE: 基准测试，即仅仅做二元语法分词和词性标注，不引入未定义词识别；

2） +PER: 在 BASE 的基础上引入人名识别 HMM。

3） +LOC: 在+PER 的基础上引入地名识别 HMM。

4） +ORG: 在+LOC 的基础上引入机构名识别 HMM。

下图给出了四种条件下，词法分析的分词正确率 SEG、上位词性标对率 TAG1、下位词性标对率 TAG2，人名识别的 F-1 值 FP、地名识别的 F-1 值 FL 以及机构名识别的 F-1 值 FO。另外要说明的是，由于词典中已经收录了很多人名、地名和机构名，所以在基准测试 BASE 中，FP、FL、FO 的值并不是零。

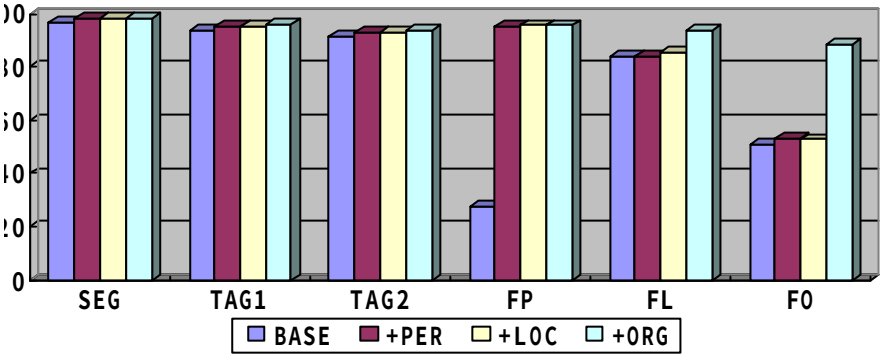


图 3.5 层叠隐马尔可夫模型的汉语词法分析结果

从图中我们可以发现：

1) 随着各层隐马尔可夫模型的逐层加入，词法分析的效果逐步提升。其中，人名识别引入后，切分正确率 SEG 在 96.55%的基础上，增加到 97.96%，增幅最大。人名、地名、机构名等识别过程均加入后，切分正确率 SEG、上位词性标对率 TAG1、下位词性标对率 TAG2 分别达到了 98.38%,95.76%,93.52%。这表明：各层隐马尔可夫模型对最终词法分析均发挥了积极作用。

2）同时，随着各层隐马尔可夫模型的加入，不仅极大的提高了本层模型的最终性能，还改进了低层模型的处理精度。人名识别加入后，人名识别的 F-1 值，立即从 27.86%提升到 95.40%，低层的分词模型的正确率也提高了 1.41%；机构名识别引入后，机构名识别的 F-1 值提高了 35.59%，同时低层的地名识别也提高了 8.49%，人名识别的 F-1 值也达到了最高点 95.58%。其原因在于：高层隐马尔可夫模型的成功应用在解决当前问题的同时，也消除了低层模型的部分歧义，排除了低层模型所产生的部分错误结果。例如：在人名识别中很容易将“刘庄的水很甜” 中的“刘庄”错误识别为人名，然而，高层的地名识别模型会正确地将“刘庄”作为地名召回，因此达到了排歧的作用。

3.8.2 在 973 评测中的测试结果

2002 年 7 月 6 日，我们开发的基于层叠隐马尔可夫模型的汉语词法分析系统 ICTCLAS 参加了国家 973 英汉机器翻译第二阶段的开放评测，测试结果如下：

领域	词数	SEG	TAG1	RTAG
体育	33,348	97.01%	86.77%	89.31%
国际	59,683	97.51%	88.55%	90.78%
文艺	20,524	96.40%	87.47%	90.59%
法制	14,668	98.44%	85.26%	86.59%
理论	55,225	98.12%	87.29%	88.91%
经济	24,765	97.80%	86.25%	88.16%
总计	208,213	97.58%	87.32%	89.42%

表 3.7　973 评测中的测试结果

要说明的几个问题：

1) 数据来源：国家 973 英汉机器翻译第二阶段评测（2002 年）的评测总结报告；
2) 词性标注相对正确率是指在排除分词错误的情况下的词性标注正确率：RTAG=TAG1/SEG*100%
3) 由于我们采用的词性标注体系和 973 评测专家组的标注体系有较大出入，所以词性标注的正确率比我们自己的测试结果有较大差距。

3.8.3 第一届国际分词大赛的评测结果

为了比较和评价不同方法和系统的性能，国际计算语言学联合会(Association for Computational Linguistics, ACL)下设的中文处理特别兴趣研究组(the ACL Special Interest Group on Chinese Language Processing, SIGHAN)于 2003 年 4 月 22 日至 25 日举办了第一届国际汉语分词评测大赛(First International Chinese Word Segmentation Bakeoff) [Sproat 2003]。报名参赛的分别是来自于大陆、台湾、美国等 6 个国家和地区，共计 19 家研究机构，最终提交结果的是 12 家参赛队伍。

大赛采取大规模语料库测试，进行综合打分的方法，语料库和标准分别来自北京大学（简体版）、宾州树库（简体版）、香港城市大学（繁体版），台湾中研院（繁体版）。每家标准分两个测试项目(Track)：封闭训练测试项目(Close Track)和开放训练测试项目(Open Track)。

ICTCLAS 分别参加了简体的所有四个测试项目，和繁体的两个封闭训练测试项目。其中在宾州树库封闭训练测试项目中综合得分 0.881，名列第一；在北京大学封闭训练测试项目中综合得分 0.951，名列第一；北京大学开放训练测试项目中综合得分 0.953，名列第二。由于我们对于繁体汉字的处理没有经验，在繁体测试项目中我们的名词不算太好，不过综合得分(0.938)比前几名相差并不大。

其中，ICTCLAS 参加北京大学封闭训练测试项目的结果为：

Site	word count	R	c_r	P	c_p	F	OOV	R_{oov}	R_{iv}
S01	17,194	0.962	±0.0029	0.940	±0.0036	0.951	0.069	0.724	0.979
S10	17,194	0.955	±0.0032	0.938	±0.0037	0.947	0.069	0.680	0.976
S09	17,194	0.955	±0.0032	0.938	±0.0037	0.946	0.069	0.647	0.977
S07	17,194	0.936	±0.0037	0.945	±0.0035	0.940	0.069	0.763	0.949
S04	17,194	0.936	±0.0037	0.942	±0.0036	0.939	0.069	0.675	0.955
S08	17,194	0.939	±0.0037	0.934	±0.0038	0.936	0.069	0.642	0.961
S06	17,194	0.933	±0.0038	0.916	±0.0042	0.924	0.069	0.357	0.975
S05	17,194	0.923	±0.0041	0.867	±0.0052	0.894	0.069	0.159	0.980

ICTCLAS 参加北京大学开放训练测试项目的结果为：

Site	word count	R	c_r	P	c_r	F	OOV	R_{oov}	R_{iv}
S10	17,194	0.963	±0.0029	0.956	±0.0031	0.959	0.069	0.799	0.975
S01	17,194	0.963	±0.0029	0.943	±0.0035	0.953	0.069	0.743	0.980
S08	17,194	0.939	±0.0037	0.938	±0.0037	0.938	0.069	0.675	0.959
S04	17,194	0.933	±0.0038	0.942	±0.0036	0.937	0.069	0.712	0.949
S03	17,194	0.940	±0.0036	0.911	±0.0043	0.925	0.069	0.647	0.962
S11	17,194	0.905	±0.0045	0.869	±0.0051	0.886	0.069	0.503	0.934

上面两个表中，S01 是 ICTCLAS，R 是分词结果的召回率，P 是分词结果的正确率，OOV（out-of-vocabulary）是未定义词（在测试语料中出现但未在训练语料中出现的词）占测试语料词数的比例，R_{oov} 是未定义词的召回率，R_{iv} 是已定义词（in-vocabulary）的召回率。C_r 和 C_p 分别表示要区分两个系统，在可信度达到 95%的情况下，它们的召回率和正确率所应该达到的差距。

跟 973 测试的结果相比，可以看到 SigHan 测试得到的分词正确率要低一些，这主要的原因是由于 973 测试和 SigHan 测试采用了不同的标准。在 973 测试中，测试的正确率最后是由人来判断的，所以切分过程中只要不是硬伤，都算正确。而在 SigHan 测试中，完全根据测试语料提供单位给出的标准答案进行评分，因此正确率要低一些。

注：上面两个表引自[Sproat 2003]，其中第二个表中的第二个 C_r 应为 C_p，原文有误。

3.9 汉语词法分析小结

我们提出的基于层叠隐马尔可夫模型的汉语分词方法，是一个基于统计方法的、集成的一体化汉语词法分析解决方案。这种方法的主要特色在于：

1. 采用集成的算法框架，即层叠隐马尔可夫模型。整个词法分析过程分为五个阶段，每个阶段都是以隐马尔可夫模型作为基本的算法模型；整个算法的时间复杂度和隐马尔可夫模型的时间复杂度相同，分析时间随着输入串长度的增长而线性增长，速度非常快；
2. 各层隐马尔可夫模型之间以下两种方式互相关联，形成一种紧密的耦合关系：每一层隐马尔可夫模型都采用 N-Best 策略，将产生的最好的若干个结果送到词图中供高层模型使用，在效率和质量之间取得了较好的平衡；低层的隐马尔可夫模型为高层隐马尔可夫模型的参数估计提供支持，通过基于角色的词语生成模型较好地解决了高层隐马尔可夫模型中未定义词的输出概率估计问题；
3. 所有各层隐马尔可夫模型都采用《人民日报》标注语料库作为训练语料库，通过对该语料库进行不同形式的改造以适应各层隐马尔可夫模型的使用，而这种改造绝大部分都是自动进行的，只需要介入很少量的人工校对；
4. 采用了我们自己定义的词性标记集，这个词性标记集克服了现有各种词性标记集的一些问题，比现有的词性标记集更详细，也更适合于机器翻译的应用。而且这种词性标记集可以方便地转换到现有的各种主要词性标记集。

我们自己的实验和国内外汉语词法分析评测的结果都表明，层叠隐马尔可夫模型是解决汉语词法分析问题的一种有效方法。

当然，和自然语言中的其他问题一样，汉语词法分析问题是不可能有完美的解决办法的。基于层叠隐马尔可夫模型的汉语词法分析方法也还有很多改进的余地。我们最近希望做的两项改进是：

1. 提高系统的领域适应性：基于统计的系统总是对训练语料有一定的偏向。我们希望

能够开发一种领域自适应技术，使得系统在对新领域的语料进行处理时能够通过自学习的方法提高正确率；

2．**研制更快速的汉语分词方法：对于某些需求（如搜索引擎）来说现有的系统性能还是不能满足要求。我们希望通过对现有系统的适当裁剪，能够大大提高系统的效率，而不至于对系统的性能产生太大影响。**

第4章　基于《知网》的词汇语义相似度计算

4.1 引言

基于实例的机器翻译中的一个关键问题是相似例句的查找，这其中要用到的一种基本的技术就是句子的相似度计算。而句子的相似度计算最终都要归结为词语的相似度计算。本章给出一种基于《知网》的词汇语义相似度计算方法。

看一个简单的例子。要翻译“张三写的小说”这个短语，通过语料库检索得到译例：

1）李四写的小说／the novel written by Li Si

2）去年写的小说／the novel written last year

通过相似度计算我们发现，“张三”和“李四”都是具体的人，语义上非常相似，而“去年”的语义是时间，和“张三”相似度较低，因此我们选用“李四写的小说”这个实例进行类比翻译，就可以得到正确的译文：

the novel written by Zhang San

如果选用后者作为实例，那么得到的错误译文将是：

* the novel written Zhang San

通过这个例子可以看出相似度计算在基于实例的机器翻译中所起的作用。

另外，在基于实例的翻译喝基于统计的机器翻译中，还有一项重要的工作是双语对齐。在双语对齐过程中也要用到两种语言的词语相似度计算，这里不再做详细的介绍。

除了机器翻译以外，词语相似度计算在自然语言处理的其他很多领域中也都有广泛的应用，例如信息检索、信息抽取、文本分类、词义排歧等等[Gauch 1995] [Li 1995] [王斌 1999] [李娟子 1999]。

4.2 词语相似度及其计算的方法

4.2.1 什么是词语相似度

什么是词语相似度？

我们认为，词语相似度是一个主观性相当强的概念。脱离具体的应用去谈论词语相似度，很难得到一个统一的定义。因为词语之间的关系非常复杂，其相似或差异之处很难用一个简单的数值来进行度量。从某一角度看非常相似的词语，从另一个角度看，很可能差异非常大。

不过，在具体的应用中，词语相似度的含义可能就比较明确了。例如，在基于实例的机器翻译中，词语相似度主要用于衡量文本中词语的可替换程度；而在信息检索中，相似度更多的要反映文本或者用户查询在意义上的符合程度。

本文的研究主要以基于实例的机器翻译为背景，因此在本文中我们所理解的词语相似度就是两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。两个词语，如果在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性越大，二者的相似度就越高，否则相似度就越低。

相似度是一个数值，一般取值范围在[0,1]之间。一个词语与其本身的语义相似度为 1。

如果两个词语在任何上下文中都不可替换，那么其相似度为 0。

相似度这个概念，涉及到词语的词法、句法、语义甚至语用等方方面面的特点。其中，对词语相似度影响最大的应该是词的语义。

4.2.2 词语相似度与词语距离

度量两个词语关系的另一个重要数量指标是词语的距离。

一般而言，词语距离是一个[0,∞)之间的实数。

一个词语与其本身的距离为 0。

词语距离与词语相似度之间有着密切的关系。

两个词语的距离越大，其相似度越低；反之，两个词语的距离越小，其相似度越大。二者之间可以建立一种简单的对应关系。这种对应关系需要满足以下几个条件：

1) 两个词语距离为 0 时，其相似度为 1；
2) 两个词语距离为无穷大时，其相似度为 0；
3) 两个词语的距离越大，其相似度越小（单调下降）。

对于两个词语 W_1 和 W_2，我们记其相似度为 $Sim(W_1,W_2)$，其词语距离为 $Dis(W_1,W_2)$，那么我们可以定义一个满足以上条件的简单的转换关系：

$$Sim(W_1,W_2)=\frac{a}{Dis(W_1,W)+a} \qquad \cdots\cdots \quad (4.1)$$

其中 α 是一个可调节的参数。 α 的含义是：当相似度为 0.5 时的词语距离值。

这种转换关系并不是唯一的，我们这里只是给出了其中的一种可能。

在很多情况下，直接计算词语的相似度比较困难，通常可以先计算词语的距离，然后再转换成词语的相似度。

4.2.3 词语相似度与词语相关性

度量两个词语关系的另一个重要的数量指标是词语的相关性。

词语相关性反映的是两个词语互相关联的程度。可以用这两个词语在同一个语境中共现的可能性来衡量。

词语相关性也是一个[0,1]之间的实数。

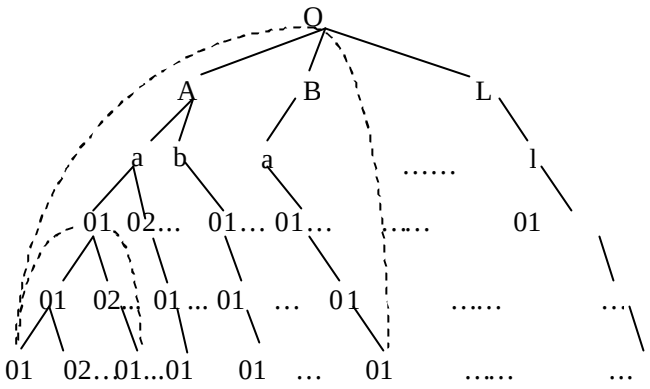
词语相关性和词语相似性是两个不同的概念。例如“医生”和“疾病”两个词语，其相似性非常低，而相关性却很高。可以这么认为，词语相似性反映的是词语之间的聚合特点，而词语相关性反映的是词语之间的组合特点。

同时，词语相关性和词语相似性又有着密切的联系。如果两个词语非常相似，那么这两个词语与其他词语的相关性也会非常接近。反之，如果两个词语与其他词语的相关性特点很接近，那么这两个词一般相似程度也很高。

4.2.4 词语相似度的计算方法

词语距离有两类常见的计算方法，一种是根据某种世界知识（Ontology）或分类体系（Taxonomy）来计算，一种利用大规模的语料库进行统计。

根据世界知识（Ontology）或分类体系（Taxonomy）计算词语语义距离的方法，一般是利用一部同义词词典（Thesaurus）。一般同义词词典都是将所有的词组织在一棵或几棵树状的层次结构中。我们知道，在一棵树形图中，任何两个结点之间有且只有一条路径。于是，这条路径的长度就可以作为这两个概念的语义距离的一种度量。



虚线用于标识某上层节点到下层节点的路径

图 4.1 《同义词词林》语义分类树形图

[王斌 1999]采用这种方法利用《同义词词林》来计算汉语词语之间的相似度（如图 4.1 所示）。有些研究者考虑的情况更复杂。[Agirre 1995]在利用 Wordnet 计算词语的语义相似度时，除了结点间的路径长度外，还考虑到了其他一些因素。例如：

1) 概念层次树的深度：路径长度相同的两个结点，如果位于概念层次的越底层，其语义距离较大；比如说：“动物”和“植物”、“哺乳动物”和“爬行动物”，这两对概念间的路径长度都是 2，但前一对词处于语义树的较高层，因此认为其语义距离较大，后一对词处于语义树的较低层，其语义距离更小；
2) 概念层次树的区域密度：路径长度相同的两对结点，如果一对位于概念层次树中低密度区域，另一对位于高密度区域，那么前者的语义距离应大于后者。引入区域密度的原因在于，有些概念层次树中概念描述的粗细程度不均，例如在 Wordnet 中，动植物分类的描述及其详尽，而有些区域的概念描述又比较粗疏，这会导致语义距离计算的不合理。

另一种词语相似度的计算方法是大规模的语料来统计。例如，利用词语的相关性来计算词语的相似度。事先选择一组特征词，然后计算这一组特征词与每一个词的相关性（一般用这组词在实际的大规模语料中在该词的上下文中出现的频率来度量），于是，对于每一个词都可以得到一个相关性的特征词向量，然后利用这些向量之间的相似度（一般用向量的夹角余弦来计算）作为这两个词的相似度。这种做法的假设是，凡是语义相近的词，他们的上下文也应该相似。[李涓子 1999]利用这种思想来实现语义的自动排歧；[鲁松 2001]研究了如何如何利用词语的相关性来计算词语的相似度。[Dagan 1995,1999]使用了更为复杂的概率模型来计算词语的距离。

这两种方法各有特点。基于世界知识的方法简单有效，也比较直观、易于理解，但这种方法得到的结果受人的主观意识影响较大，有时并不能准确反映客观事实。另外，这种方法比较准确地反映了词语之间语义方面的相似性和差异，而对于词语之间的句法和语用特点考虑得比较少。基于语料库的方法比较客观，综合反映了词语在句法、语义、语用等方面的相似性和差异。但是，这种方法比较依赖于训练所用的语料库，计算量大，计算方法复杂，另外，受数据稀疏和数据噪声的干扰较大，有时会出现明显的错误。

本文主要研究基于《知网（Hownet）》的词语相似度计算方法，这是一种基于世界知识

的方法。

4.3 《知网（Hownet)》简介

按照《知网》的创造者——董振东先生自己的说法[杜飞龙 1999]：

《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

《知网》中含有丰富的词汇语义知识和世界知识，为自然语言处理和机器翻译等方面的研究提供了宝贵的资源。不过，在我们真正试图利用《知网》来进行计算机处理时，发现还是会遇到不少困难。我们的感觉是，《知网》确实是一座宝库，但另一方面，《知网》的内容又非常庞杂。尽管《知网》的提供了详细的文档，但由于这些文档不是以一种形式化的方式说明的，很多地方多少显得有些混乱。当我们阅读这些文档时，很容易一下子陷入大量的细节之中，而很难对《知网》有一个总体的把握。这使得我们在进行计算的时候觉得很不方便。因此，我们在试图利用《知网》进行计算的过程中，也在逐渐加深我们对于《知网》的认识，并试图整理出一个关于《知网》的比较清晰的图象。

本节中，我们将主要通过对《知网》的知识描述语言的分析，利用集合、特征结构等抽象数据形式，将《知网》的知识描述语言表示成一种更为直观、更为结构化的形式，以便于后面的相似度计算。

4.3.1 《知网》的结构

董振东先生反复强调，《知网》并不是一个在线的词汇数据库，《知网》不是一部语义词典。

在介绍《知网》的结构之前，我们首先要理解《知网》中两个主要的概念：“概念”与“义原”。

“概念”是对词汇语义的一种描述。每一个词可以表达为几个概念。

“概念”是用一种“知识表示语言”来描述的，这种“知识表示语言”所用的“词汇”叫做“义原”。

“义原”是用于描述一个“概念”的最小意义单位。

与一般的语义词典（如《同义词词林》，或 Wordnet）不同，《知网》并不是简单的将所有的“概念”归结到一个树状的概念层次体系中，而是试图用一系列的“义原”来对每一个“概念”进行描述。

《知网》一共采用了 1500 义原，这些义原分为以下几个大类：

1) Event|事件
2) entity|实体
3) attribute|属性值
4) aValue|属性值
5) quantity|数量
6) qValue|数量值
7) SecondaryFeature|次要特征
8) syntax|语法
9) EventRole|动态角色
10) EventFeatures|动态属性

对于这些义原，我们把它们归为三组：第一组，包括第 1 到 7 类的义原，我们称之为“ 基本义原 ”，用来描述单个概念的语义特征；第二组，只包括第 8 类义原，我们称之为“ 语法义原 ”，用于描述词语的语法特征，主要是词性（Part of Speech）；第三组，包括第 9 和第 10 类的义原，我们称之为“ 关系义原 ”，用于描述概念和概念之间的关系（类似于深层格语法中的格关系）。

除了义原以外，《知网》中还用了一些符号来对概念的语义进行描述，如下表所示：

表 4.1 《知网》知识描述语言中的符号及其含义

,	多个属性之间，表示“ 和 ”的关系
#	表示“ 与其相关 ”
%	表示“ 是其部分 ”
$	表示“ 可以被该‘ V ’处置，或是该“ V ”的受事，对象，领有物，或者内容
*	表示“ 会‘ V ’或主要用于‘ V ’，即施事或工具
+	对 V 类，它表示它所标记的角色是一种隐性的，几乎在实际语言中不会出现
&	表示指向
~	表示多半是，多半有，很可能的
@	表示可以做“ V ”的空间或时间
?	表示可以是“ N ”的材料，如对于布匹，我们标以“ ?衣服 ”表示布匹可以是“ 衣服 ”的材料
{}	(1) 对于 V 类，置于 [ ] 中的是该类 V 所有的“ 必备角色 ”。如对于“ 购买 ”类，一旦它发生了，必然会在实际上有如下角色参与：施事，占有物，来源，工具。尽管在多数情况下，一个句子并不把全部的角色都交代出来 (2) 表示动态角色，如介词的定义
()	置于其中的应该是一个词表记，例如，(China|中国)
^	表示不存在，或没有，或不能
!	表示某一属性为一种敏感的属性，例如：“ 味道 ”对于“ 食物 ”，“ 高度 ”对于“ 山脉 ”，“ 温度 ” 对于“ 天象 ”等
[]	标识概念的共性属性

我们把这些符号又分为几类，一类是用来表示语义描述式之间的逻辑关系，我们称之为“ 逻辑符号 ”，包括以下几个符号：, ~ ^ ，另一类用来表示概念之间的关系，我们称之为“ 关系符号 ”，包括以下几个符号：# % $ * + & @ ? ! ，第三类包括几个无法归入以上两类的“ 特殊符号 ”：{} () [] 。

我们看到，概念之间的关系有两种表示方式：一种是用“ 关系义原 ”来表示，一种是用表示概念关系的“ 关系符号 ”来表示。按照我们的理解，前者类似于一种深层格关系，后者大部分是一种深层格关系的“ 反关系 ”，例如“ $ ”我们就可以理解为“ 施事、对象、领有、内容 ”的反关系，也就是说，该词可以充当另一个词的“ 施事、对象、领有、内容 ”。

义原一方面作为描述概念的最基本单位，另一方面，义原之间又存在复杂的关系。在《知网》中，一共描述了义原之间的 8 种关系：上下位关系、同义关系、反义关系、对义关系、属性-宿主关系、部件-整体关系、材料-成品关系、事件-角色关系。可以看出，义原之间组成的是一个复杂的网状结构，而不是一个单纯的树状结构。不过，义原关系中最重要的还是的上下位关系。根据义原的上下位关系，所有的“基本义原”组成了一个义原层次体系（如图 4.2）。这个义原层次体系是一个树状结构，这也是我们进行语义相似度计算的基础。

从表面上看，其他的语义词典，如《同义词词林》和 Wordnet，也有一个树状的概念层

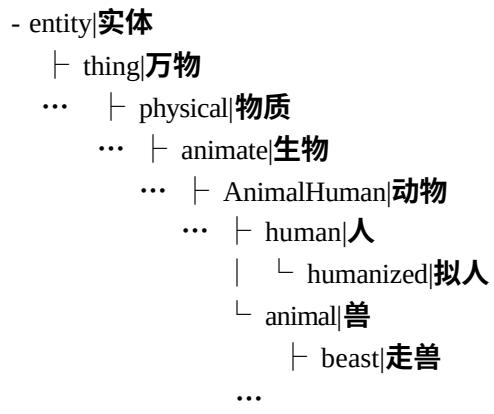


图 4.2 树状的义原层次结构

次体系，好像《知网》和它们很相似，但实际上有着本质的不同。在《同义词词林》和 Wordnet 种，概念就是描写词义的最小单位，所以，每一个概念都是这个概念层次体系中的一个结点。而在《知网》中，每一个概念是通过一组义原来表示的，概念本身并不是义原层次体系中的一个结点，义原才是这个层次体系中的一个结点。而且，一个概念并不是简单的描述为一个义原的集合，而是要描述为使用某种专门的“知识描述语言”来表达的一个语义表达式。也就是说，在描述一个概念的多个义原中，每个义原所起到的作用是不同的，这就给我们的相似度计算带来了很大的困难。下面我们就对这个描述概念的知识描述语言进行一些考察。

4.3.2 《知网》的知识描述语言

《知网》对概念的描述是比较复杂的。在《知网》中，每一个概念用一个记录来表示，如下所示：

NO.=017144
W_C=打
G_C=V
E_C=~网球，~牌，~秋千，~太极，球~得很棒
W_E=play
G_E=V
E_E=
DEF=exercise|锻练,sport|体育

其中 NO.为概念编号，W_C，G_C，E_C 分别是汉语的词语、词性和例子，W_E、G_E、E_E 分别是英语的词语、词性和例子，DEF 是知网对于该概念的定义，我们称之为一个语义表达式。其中 DEF 是知网的核心。我们这里所说的知识描述语言也就是 DEF 的描述语言。

在《知网》的文档中，对知识描述语言做了详尽的介绍。不过，由于该文档过于偏重细节，不易从总体上把握。本节中我们试图对于这种知识描述语言给出一个简单的概括。

我们看几个例子：

表 4.2：《知网》知识描述语言实例

打	017144	exercise|锻练,sport|体育
男人	059349	human|人,family|家,male|男
高兴	029542	aValue|属性值,circumstances|境况,happy|福,desired|良
生日	072280	time|时间,day|日,@ComeToWorld|问世,$congratulate|祝贺
写信	089834	write|写,ContentProduct=letter|信件
北京	003815	place|地方,capital|国都,ProperName|专,(China|中国)
爱好者	000363	human|人,*FondOf|喜欢,#WhileAway|消闲
必须	004932	{modality|语气}
串	015204	NounUnit|名量,&(grape|葡萄),&(key|钥匙)
从良	016251	cease|停做,content=(prostitution|卖淫)
打对折	017317	subtract|削减,patient=price|价格,commercial|商,(range|幅度=50%)
儿童基金会	024083	part|部件,%institution|机构,politics|政,#young|幼,#fund|资金,(institution|机构=UN|联合国)

从这些例子我们可以看到，《知网》的知识描述语言是比较复杂的。我们将这种知识描述语言归纳为以下几条：

1) 《知网》收入的词语主要归为两类，一类是实词，一类是虚词；
2) 虚词的描述比较简单，用“{句法义原}”或“{关系义原}”进行描述；
3) 实词的描述比较复杂，由一系列用逗号隔开的“语义描述式”组成，这些“语义描述式”又有以下三种形式：
   a) 基本义原描述式：用“基本义原”，或者“(具体词)”进行描述；
   b) 关系义原描述式：用“关系义原=基本义原”或者“关系义原=(具体词)”或者“(关系义原=具体词)”来描述；
   c) 关系符号描述式：用“关系符号 基本义原”或者“关系符号(具体词)”加以描述，我们还注意到，可以有多个关系符号描述式采用同一个关系符号；
4) 在实词的描述中，第一个描述式总是一个基本义原描述式，这也是对该实词最重要的一个描述式，这个基本义原描述了该实词的最基本的语义特征。

根据以上分析，我们将《知网》对一个实词的描述重新表示如下：

$$
\text{单词}:\begin{bmatrix}
\text{第一基本义原描述}=\text{基本义原}_a \\
\text{其他基本义原描述}=\{\text{基本义原}_b,\text{基本义原}_c,\ldots\} \\
\text{关系义原描述}=\begin{bmatrix}\text{关系义原}_1=\text{基本义原}_x\mid\text{具体词}_x \\ \text{关系义原}_2=\text{基本义原}_y\mid\text{具体词}_y \\ \ldots\end{bmatrix} \\
\text{关系符号描述}=\begin{bmatrix}\text{关系符号}_1=\{\text{义原}_u\mid\text{具体词}_u,\text{义原}_v\mid\text{具体词}_v,\ldots\} \\ \text{关系符号}_2=\{\text{义原}_s\mid\text{具体词}_s,\text{义原}_t\mid\text{具体词}_t,\ldots\} \\ \ldots\end{bmatrix}
\end{bmatrix}
$$

在上面的表达式中，“[……]”表示特征结构，“{……}”表示集合，“|”表示“或”。特征结构和集合是这个表达式中使用的两种抽象数据结构，也是下面我们进行相似度计算时面对的主要问题。

4.4 基于《知网》的语义相似度计算方法

从上面的介绍我们看到，与传统的语义词典不同，在《知网》中，并不是将每一个概念对应于一个树状概念层次体系中的一个结点，而是通过用一系列的义原，利用某种知识描述语言来描述一个概念。而这些义原通过上下位关系组织成一个树状义原层次体系。我们的目标是要找到一种方法，对用这种知识描述语言表示的两个语义表达式进行相似度计算。

利用《知网》计算语义相似度一个最简单的方法就是直接使用词语语义表达式中的第一基本义原描述式，把词语相似度等价于第一基本义原的相似度。这种方法好处是计算简单，但没有利用知网语义表达式中其他部分丰富的语义信息。

[Li 2002]中提出了一种词语语义相似度的计算方法，计算过程综合利用了《知网》和《同义词词林》。在义原相似度的计算过程中，不仅考虑了义原之间的上下文关系，还考虑了义原之间的其他关系。在计算词语相似度时，加权合并了《同义词词林》的词义相似度、《知网》语义表达式的义原相似度和义原关联度。这种算法中，《同义词词林》和《知网》采用完全不同的语义体系和表达方式，词表也相差较大，把它们合并计算的合理性值得怀疑。另外，把语义关联度加权合并计入义原相似度中，也未必合理。

4.4.1 词语相似度计算

对于两个汉语词语 W_1 和 W_2，如果 W_1 有 n 个义项（概念）：S_{11}，S_{12}，……，S_{1n}，W_2 有 m 个义项（概念）：S_{21}，S_{22}，……，S_{2m}，我们规定，W_1 和 W_2 的相似度各个概念的相似度之最大值，也就是说：

$$Sim(W_1,W_2)=\max_{i=1..n,j=1...m} Sim(S_{1i},S_{2j}) \qquad \cdots\cdots (4.2)$$

这样，我们就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题。当然，我们这里考虑的是孤立的两个词语的相似度。如果是在一定上下文之中的两个词语，最好是先进行词义排岐，将词语标注为概念，然后再对概念计算相似度。

4.4.2 义原相似度计算

由于所有的概念都最终归结于用义原（个别地方用具体词）来表示，所以义原的相似度计算是概念相似度计算的基础。

由于所有的义原根据上下位关系构成了一个树状的义原层次体系，我们这里采用简单的通过语义距离计算相似度的办法。假设两个义原在这个层次体系中的路径距离为 d，根据公式(4.1)，我们可以得到这两个义原之间的语义距离：

$$Sim(p_1,p_2)=\frac{a}{d+a} \qquad \cdots\cdots (4.3)$$

其中 p_1 和 p_2 表示两个义原（primitive），d 是 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数。 α 是一个可调节的参数。

用这种方法计算义原相似度的时候，我们只利用了义原的上下位关系。实际上，在《知网》中，义原之间除了上下位关系外，还有很多种其他的关系，如果在计算时考虑进来，可能会得到更精细的义原相似度度量，例如，我们可以认为，具有反义或者对义关系的两个义原比较相似，因为它们在实际的语料中互相可以互相替换的可能性很大。对于这个问题这里

我们不展开讨论。

另外，在知网的知识描述语言中，在一些义原出现的位置都可能出现一个具体词（概念），并用圆括号( )括起来。所以我们在计算相似度时还要考虑到具体词和具体词、具体词和义原之间的相似度计算。理想的做法应该是先把具体词还原成《知网》的语义表达式，然后再计算相似度。这样做将导入函数的递归调用，这会使算法会变得很复杂。由于具体词在《知网》的语义表达式中只占很小的比例，因此，在我们的实验中，为了简化起见，我们做如下规定：

- 具体词与义原的相似度一律处理为一个比较小的常数（γ）；
- 具体词和具体词的相似度，如果两个词相同，则为 1，否则为 0。

4.4.3 虚词概念的相似度的计算

我们认为，在实际的文本中，虚词和实词总是不能互相替换的，因此，虚词概念和实词概念的相似度总是为零。

由于虚词概念总是用“{句法义原}”或“{关系义原}”这两种方式进行描述，所以，虚词概念的相似度计算非常简单，只需要计算其对应的句法义原或关系义原之间的相似度即可。

4.4.4 实词概念的相似度的计算

由于实词概念是用一个语义表达式来描述的，因此其相似度计算变得非常复杂。

如何计算两个语义表达式的相似度呢？

从前面的分析可知，《知网》的知识描述语言可以通过义原和集合、特征结构这两种抽象数据结构来表达。义原之间的相似度计算问题已经解决，剩下的问题就是集合和特征结构的相似度问题了。

我们的基本设想是：整体相似要建立在部分相似的基础上。把一个复杂的整体分解成部分，通过计算部分之间的相似度得到整体的相似度。

假设两个整体 A 和 B 都可以分解成以下部分：A 分解成 A_1，A_2，……，A_n，B 分解成 B_1，B_2，……，B_m，那么这些部分之间的对应关系就有 $m \times n$ 种。问题是：这些部分之间的相似度是否都对整体的相似度发生影响？如果不是全部都发生影响，那么我们应该如何选择那些发生影响的那些部分之间的相似度？选择出来以后，我们又如何得到整体的相似度？

我们认为：一个整体的各个不同部分在整体中的作用是不同的，只有在整体中起相同作用的部分互相比较才有效。例如比较两个人长相是否相似，我们总是比较它们的脸型、轮廓、眼睛、鼻子等相同部分是否相似，而不会拿眼睛去和鼻子做比较。

因此，在比较两个整体的相似性时，我们首先要做的工作是对这两个整体的各个部分之间建立起一一对应的关系，然后在这些对应的部分之间进行比较。

还有一个问题：如果某一部分的对应物为空，如何计算其相似度？我们的处理方法是：

- 将任一非空值与空值的相似度定义为一个比较小的常数（δ）；

下面我们分别考虑集合和特征结构的相似度计算问题。

4.4.4.1 特征结构的相似度计算

特征结构可以理解为一个“属性：值”对（Attribute-Value Pair）的集合，我们将一个

“属性：值”对称为一个“特征”(Feature)。在一个特征结构中，每个“特征”的“属性”是唯一的。

计算两个特征结构的相似度，首先要在两个特征结构的特征之间建立起一一对应的关系。由于每个特征结构的各个特征都具有不同的属性，因此这种一一对应关系通过特征的属性很容易建立起来：属性相同的特征之间一一对应，如果没有属性相同的特征，那么该特征的对应物为空。

这样，特征结构的相似度就转化为各个特征的相似度的加权平均。其中的权值反映出该属性在特征结构中的重要程度。比如，人们常说“眼睛是心灵之窗”，在人脸的相似度计算中，“眼睛”的相似度的权值应该比“鼻子”、“耳朵”的相似度要高。

剩下的问题就是计算两个特征的相似度。特征由“属性”和“值”组成。由于“属性”相同，于是，两个特征的相似度可以等价于其“值”的相似度。

4.4.4.2 集合的相似度计算

集合的相似度计算比特征结构更为复杂，因为集合的元素是无序的，要在两个集合的元素之间建立一一对应关系并不容易。

两个集合的相似度计算模型，必须满足我们对于集合相似度计算的一些直观要求。这里我们列出以下两条：

1. 一个集合和它本身的相似度为 1；
2. 假设两个集合都有 n 个元素，其中 m（m<n）个元素相同，又假设两个元素的相似度只能是 0（不同）或 1（相同），那么这两个集合的相似度应该是 m/n。

要计算两个集合的相似度，最容易想到的方法是首先计算两个集合的所有元素两两之间的相似度，然后再进行加权平均。但是这样会带来一个问题，就是一个集合和它本身的相似度可能不为 1，除非它的任意两个元素之间的相似度都为 1。这个结果当然是不合理的。这也从另一个角度说明我们先前定义的原则（首先在两个集合的元素之间建立一一对应关系）的合理性。

在本文中，我们采用以下算法来为两个集合的元素之间建立一一对应关系：

1. 首先计算两个集合的所有元素两两之间的相似度；
2. 从所有的相似度值中选择最大的一个，将这个相似度值对应的两个元素对应起来；
3. 从所有的相似度值中删去那些已经建立对应关系的元素的相似度值；
4. 重复上述第 2 步和第 3 步，直到所有的相似度值都被删除；
5. 没有建立起对应关系的元素与空元素对应。

根据上述算法建立起两个集合元素的一一对应关系后，我们就很容易计算两个集合的相似度了：集合的相似度等于其元素对的相似度的加权平均。又因为集合的元素之间都是平等的，所以我们可以将所有的权值取成相同的，于是：集合的相似度等于其元素对的相似度的算术平均。

容易看出，按照这种方法得到的相似度是满足上面提到的两个直观要求的。

4.4.4.3 实词概念相似度的计算

由前面的分析我们知道，在《知网》中对一个实词的描述可以表示为一个特征结构，该特征结构含有以下四个特征：

1) 第一基本义原描述：其值为一个基本义原，我们将两个概念的这一部分的相似度记

为 $Sim_1(S_1,S_2)$；

2) 其他基本义原描述：对应于语义表达式中除第一基本义原描述式以外的所有基本义原描述式，其值为一个基本义原的集合，我们将两个概念的这一部分的相似度记为 $Sim_2(S_1,S_2)$；
3) 关系义原描述：对应于语义表达式中所有的用关系义原描述式，其值是一个特征结构，对于该特征结构的每一个特征，其属性是一个关系义原，其值是一个基本义原，或一个具体词。我们将两个概念的这一部分的相似度记为 $Sim_3(S_1,S_2)$；
4) 关系符号描述：对应于语义表达式中所有的用关系符号描述式，其值也是一个特征结构，对于该特征结构的每一个特征，其属性是一个关系义原，其值是一个集合，该集合的元素是一个基本义原，或一个具体词。我们将两个概念的这一部分的相似度记为 $Sim_4(S_1,S_2)$。

于是，两个概念语义表达式的整体相似度记为：

$$Sim(S_1, S_2) = \sum_{i=1}^{4} b_i Sim_i(S_1, S_2) \qquad \cdots\cdots \quad (4.4)$$

其中，β_i（$1 \le i \le 4$）是可调节的参数，且有：$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1 \ge \beta_2 \ge \beta_3 \ge \beta_4$。后者反映了 Sim_1 到 Sim_4 对于总体相似度所起到的作用依次递减。由于第一基本义原描述式反映了一个概念最主要的特征，所以我们应该将其权值定义得比较大，一般应在 0.5 以上。

在实验中我们发现，如果 Sim_1 非常小，但 Sim_3 或者 Sim_4 比较大，将导致整体的相似度仍然比较大的不合理现象。因此我们对公式(4.4)进行了修改，得到公式如下：

$$Sim(S_1, S_2) = \sum_{i=1}^{4} b_i \prod_{j=1}^{i} Sim_j(S_1, S_2) \qquad \cdots\cdots \quad (4.5)$$

其意义在于，主要部分的相似度值对于次要部分的相似度值起到制约作用，也就是说，如果主要部分相似度比较低，那么次要部分的相似度对于整体相似度所起到的作用也要降低。但可以保证一个词和它本身的相似度仍为 1。

下面我们再分别讨论每一部分的相似度。

1) 第一基本义原描述：就是两个义原的相似度，按照公式(4.3)计算即可；
2) 其他基本义原描述：其值为一个集合，转换为两个基本义原集合的相似度计算问题；
3) 关系义原描述：其值为一个特征结构，转换为两个特征结构的相似度计算问题。而这个特征结构中特征的值就是基本义原或具体词，因此这两个特征结构的相似度计算也可以最终还原到基本义原或具体词的相似度计算问题。这里，由于无法区分关系义原之间的重要程度，我们将对各个特征的相似度取算术平均；
4) 关系符号描述：其值为一个特征结构，转换为两个特征结构的相似度计算问题。而这个特征结构中特征的值又是一个集合，集合的元素才是基本义原或具体词，因此这两个特征结构的相似度计算也可以最终还原到基本义原或具体词的相似度计算问题。同样，由于无法区分关系符号之间的重要程度，我们将对各个特征的相似度取算术平均；

到此为止，我们已经讨论了基于《知网》的词语相似度计算的所有细节，具体的算法我们不再详细说明。

4.5 实验及结果

根据以上方法，我们实现了一个基于《知网》的语义相似度计算程序模块。

词语相似度计算的结果评价，最好是放到实际的系统中（如基于实例的机器翻译系统），观察不同的相似度计算方法对实际系统的性能的影响。这需要一个完整的应用系统。在条件不具备的情况下，我们采用了人工判别的方法。

我们使用了三种方法来计算词语相似度，并把它们的计算结果进行比较：

方法 1：仅使用《知网》语义表达式中第一基本义原来计算词语相似度；

方法 2：[Li 2002] 中使用的词语语义相似度计算方法；

方法 3：本文中介绍的语义相似度计算方法；

在实验中，根据在多次尝试中取得的经验，我们将几个参数值设置如下：

$\alpha = 1.6$

$\beta_1 = 0.5$, $\beta_2 = 0.2$，$\beta_3 = 0.17$，$\beta_4 = 0.13$

$\gamma = 0.2$

$\delta = 0.2$

实验结果如下表所示：

表 4.3 实验结果（一）

词语 1	词语 2	词语 2 的语义	方法 1	方法 2	方法 3
男人	女人	人,家,女	1.000	0.668	0.833
男人	父亲	人,家,男	1.000	1.000	1.000
男人	母亲	人,家,女	1.000	0.668	0.833
男人	和尚	人,宗教,男	1.000	0.668	0.833
男人	经理	人,#职位,官,商	1.000	0.351	0.657
男人	高兴	属性值,境况,福,良	0.016	0.024	0.013
男人	收音机	机器,*传播	0.186	0.008	0.164
男人	鲤鱼	鱼	0.347	0.009	0.208
男人	苹果	水果	0.285	0.004	0.166
男人	工作	事务,$担任	0.186	0.035	0.164
男人	责任	责任	0.016	0.005	0.010

考察方法 3 的结果，我们可以看到，“男人”（取义项“人，家，男”）和其它各个词的相似度与人的直觉是比较相符合的。

将方法 3 和方法 1、方法 2 的结果相比较，可以看到：方法 1 的结果比较粗糙，只要是人，相似度都为 1，显然不够合理；方法 2 的结果比方法 1 更细腻一些，能够区分不同人之间的相似度，但有些相似度的结果也不太合理，比如“男人”和“工作”的相似度比“男人”和“鲤鱼”的相似度更高。从可替换性来说，这显然不合理，至少“男人”和“鲤鱼”都是有生命物体，而“工作”只可能是一个行为或者一个抽象事物。方法 2 出现这种不合理现象的原因在于其计算方法把部分语义关联度数值加权计入了相似度中。另外，方法 2 的结果中，“男人”和“和尚”的相似度比“男人”和“经理”的相似度高出近一倍，而方法 3 的结果中，这两个相似度的差距更合理一些。

下表中给出另外一些测试结果，供读者参考：

表 4.4 实验结果（二）

词语 1	词语 2	相似度	词语 1	词语 2	相似度
工人	教师	0.722	粉红	红	1
工人	科学家	0.576	粉红	红色	1
工人	农民	0.722	粉红	绿	0.861
工人	运动员	0.722	粉红	颜色	0.059

教师	科学家	0.576	绿	颜色	0.059
教师	农民	0.722	十分	非常	1
教师	运动员	0.722	十分	特别	0.624
科学家	农民	0.576	思考	考虑	1
科学家	运动员	0.6	思考	思想	0.074
农民	运动员	0.722	考虑	思想	0.074
中国	美国	0.936	跑	跳	0.444
中国	联合国	0.136	跑	跳舞	0.127
中国	安理会	0.114	跑	运行	0.444
中国	欧洲	0.733	运行	跳舞	0.151

可以看到，绝大部分结果还是比较合理的，但也有部分结果不够合理，例如“中国”和“联合国”、“中国”和“安理会”的相似度都过低，这是因为，“中国”、“联合国”、“安理会”在《知网》中的第一基本义原分别是“地方”、“机构”、“部件”。“跑”和“跳”的相似度也较低，这是因为这两个词被简单定义为两个基本义原，而缺少其它信息。这也从一个侧面反映了知网的某些定义不合理或不一致之处。

需要声明的是，上述试验中，每个词都只取了一个最常见的义项，而不是考虑所有义项。

4.6 小结

与传统的语义词典不同，《知网》采用了 1500 多个义原，通过一种知识描述语言来对每个概念进行描述。

为了计算用知识描述语言表达的两个概念的语义表达式之间的相似度，我们采用了“整体的相似度等于部分相似度加权平均”的做法。首先将一个整体分解成部分，再将两个整体的各个部分进行组合配对，通过计算每个组合对的相似度的加权平均得到整体的相似度。我们具体讨论了特征结构和集合这两种抽象数据结构中各个组成部分的组合配对方式。通过对概念的语义表达式反复使用这一方法，可以将两个语义表达式的整体相似度分解成一些义原对的相似度的组合。对于两个义原的相似度，我们采用根据上下位关系得到语义距离并进行转换的方法。

实验证明，我们的做法充分利用了《知网》中对每个概念进行描述时的丰富的语义信息，得到的结果与人的直觉比较符合，词语相似度值刻划也比较细致。

第5章 一种双语短语结构对齐的搜索算法[1]

对齐（Alignment）无论是对基于实例的机器翻译方法还是对基于统计的机器翻译方法都是一项基础性的研究工作。对齐需要在文本的各个层次上进行，一般包括段落、句子、词语、短语等几个层次。目前，句子对齐技术已相当成熟，达到了很高的正确率。而词语对齐和短语对齐的正确率还不是很高。相对而言，词语对齐的研究比较多，而短语对齐的研究则比较少。

由于本文的目标是建立是基于短语结构层面转换的统计机器翻译模型，因此，短语结构的对齐是本项研究的一项重要内容。

5.1 双语对齐技术概述

5.1.1 各种层次的语言单位上的对齐技术

段落对齐和句子对齐中，两种最主要的方法是基于长度的对齐[Gale 1993]和基于词汇的对齐。单纯基于长度的句子对齐算法已经可以达到较高的正确率，而且不依赖于语言，效率非常高。单纯基于词汇的句子对齐算法由于词典信息的稀疏性，正确率比基于长度的对齐算法略低，性能也比较差。在基于长度的算法基础上辅以词语互译信息，可以达到更高的正确率[王斌 1999]。

词语对齐技术目前的研究也比较充分。一般来说，词语对齐要利用两方面的信息：一方面是词语相似度信息，另一方面是词语在句子中的位置变化（称为扭曲）信息。IBM 的统计机器翻译模型（IBM Model 1～5）[Brown 1990,1993]本质上都是一种词语对齐模型。[Och 1999]和[Wang 1998a]都是在 Brown 工作的基础上的改进，基本的思路都是引入组块（Chunk）模型，在组块内和组块间分别进行对齐。[Vogel 1996]引入了一种基于隐马尔可夫模型（HMM）的对齐方法，主要是用隐马尔可夫模型来反映词语之间的位置扭曲关系。[Ker 1997]提出了一种基于类（Class-based）的词语对齐模型，通过两种语言的义类词典来弥补双语词典信息的不足。

结构对齐试图在词语对齐的基础上引入双语短语之间的对应关系。不过，由于语言结构表示方法不同，结构对齐实际上也分成很多种。常见的结构对齐包括：短语结构对齐[Kaji 1992][Imamura 2001]；依存结构对齐[Matsumot 1993][Grishman 1994][Meyers 1996][Watanabe 2000]和组块结构对齐[Och 1999][Wang 1998a][王伟 2002]。

我们希望能够在双语短语结构转换的层次上建立起机器翻译的统计模型，因此在这里主要关心短语结构上的对齐算法。组块对齐是对词语对齐的一种扩充，相对短语结构的对齐和依存结构的对齐来说比较简单一些。依存结构的对齐和短语结构的对齐有很大的不同（一般不考虑交叉约束），我们这里也不做详细的介绍。下面我们主要讨论一下短语结构的对齐技术。

[1]本文中所用的汉语词“对齐”有时是指“对齐”这个动作，有时是指“对齐”这个动作所产生的结果，即两种语言单位之间的对应关系。如这里的“对齐”就属于后一种解释。请读者在阅读时自己加以判别。

5.1.2 短语结构对齐的定义

所谓短语结构，就是按照类似于乔姆斯基（Chomsky）的短语结构语法（PSG）进行句法分析所形成的句法树结构。句法树的叶结点全部由句子中的单词和词性标记组成，非叶结点都是非终结符，也就是短语类标记。短语结构和依存结构是句法树最常见的两种形式。不过为方便起见，在不引起混淆的情况下，我们有时也把短语结构树简称为句法树。

所谓短语结构的对齐，也就是在源语言和目标语言的短语结构树的结点之间寻找对应关系（即对齐）。一般来说，这种对齐应该满足：

1. **对齐模式：包括** 1:1，1:0，0:1。**也就是说，允许一个结点没有对齐的结点或者有唯一的一个对齐结点，但不允许一个结点有两个或两个以上的对齐结点；**
2. **全局约束（global constraint）原则：两棵短语结构树的根结点对齐；**
3. **交叉约束（crossing constraint）原则：如果结点** A **和结点** B **对齐，结点** A’**和结点** B’**对齐，**A **是** A’**的祖先结点，那么** B **必须是** B’**的祖先结点。**

这里我们通过一个例子来加以说明。假设语料库中有一对对齐的汉英句子：

汉语：我们可以比照其它工厂的做法拟定计划。

英语：We can draw up our plan in the light of the experience of other factories.

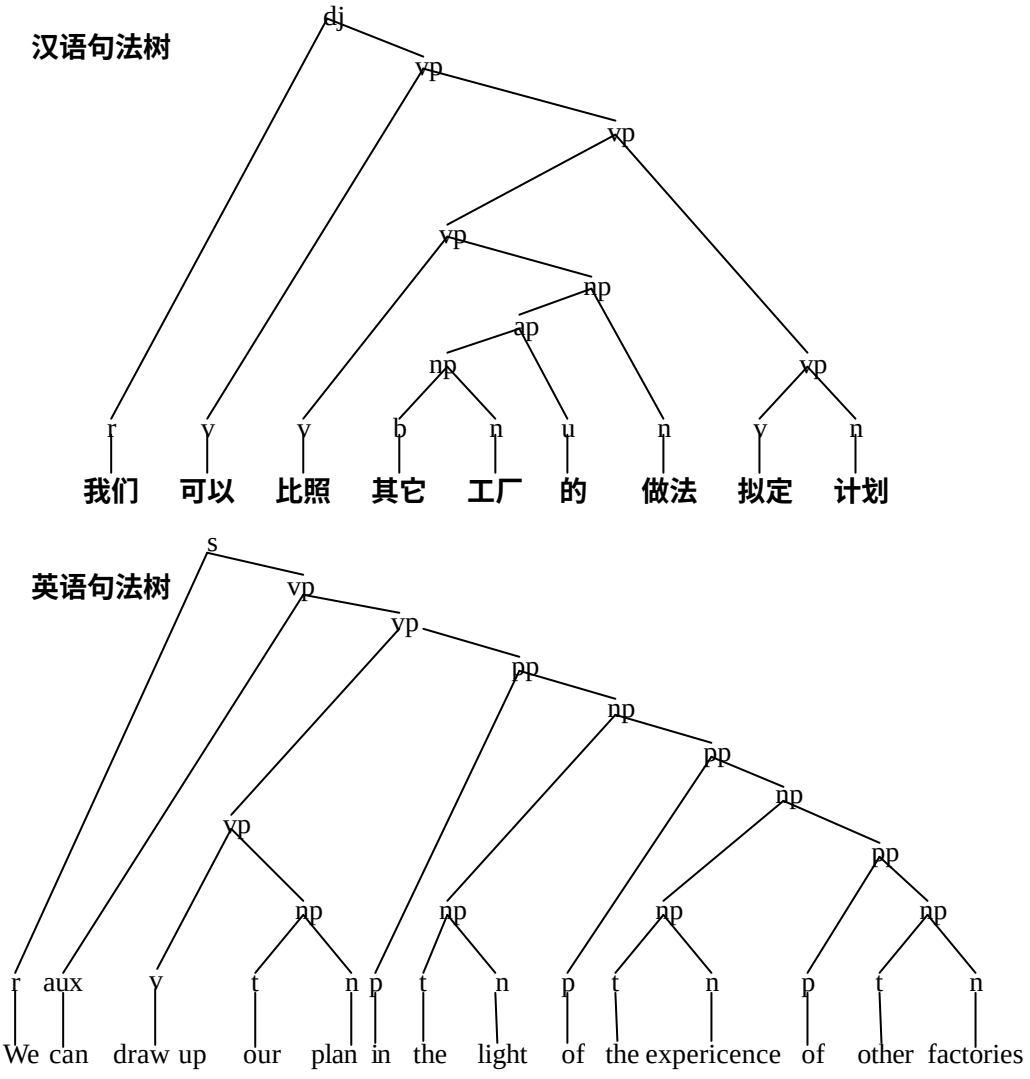


图 5.1 **两棵对应的句法树**

需要说明的是，在词典中可以一般查到“ in the light of ”是一个固定搭配。有些句法分析器把它当作一个短语来处理，这样做在工程上固然是很方便的，不过在理论上并不合理。比如说，我们可以在这个固定搭配中插入一个短语：“ in the XX and the light of ”。这有点像汉语中的离合词。对于这一点，这里不展开讨论。这里只是简单假设事先并不知道‘ in the light of ”是一个固定搭配，并试图通过短语对齐来发现这个固定搭配和“比照”之间的互译关系。

在这两棵句法树上，理想的短语结构对齐应该是：

{ r(**我们**) / r(we) }　{ v(**可以**) / v(can) }　{ a(**其它**) / a(other) }　{ n(**工厂**) / n(factories) }

{ v(**拟定**) / v(draw up) }　{ n(**计划**) / n(plan) }　{ u(**的**) / p(of) }　{ n(**做法**) / n(experience) }

{ vp(**拟定计划**) / vp(draw up our plan) }　{ np(**其它工厂**) / np(other factories) }

{ ap(**其它工厂的**) / pp(of other factories) }

{ np(**其它工厂的做法**) / np(the experiences of other factories) }

{ vp(**比照其它工厂的做法**) / pp(in the light of the experiences of other factories) }

{ vp(**比照其它工厂的做法拟定计划**) / vp(draw up our plan in the light of the experience of other factories) }

{ vp(**可以比照其它工厂的做法拟定计划**) / vp(can draw up our plan in the light of the experience of other factories.) }

{ dj(**我们可以比照其它工厂的做法拟定计划**) / s(We can draw up our plan in the light of the experience of other factories.) }

通常在词典中我们可以找到以下对应关系：

{ v(**比照**) / p(in the light of) }

不过由于“ in the light of ”并不是短语结构树中的一个结点，所以根据对齐模式的要求，这种对应关系无法反映到短语对齐当中。

5.1.3 短语结构对齐的过程

我们用上面的例子来简单说明一下短语对齐的过程。在短语结构对齐之前，我们一般都要完成以下工作：

1．源语言的句法分析（短语结构分析）；

2．目标语言的句法分析（短语结构分析）；

3．源语言和目标语言的词语对齐。

如果以上三项工作都完全没有出错，也就是说正确率是 100%，那么词语对齐工作是比较简单的。还是看上面的例子。

通过词语对齐，我们得到以下对齐的词语结点：

{ r(我们) / r(we) }　{ v(可以) / v(can) }　{ a(其它) / a(other) }　{ n(工厂) / n(factories) }

{ v(拟定) / v(draw up) }　{ n(计划) / n(plan) }　{ u(的) / p(of) }

有一些词的对应关系是很难通过词语对齐得到的，例如：

{ n(做法) / n(experience) }

根据全局约束原则和交叉约束原则，可以排除掉大部分不可能的对齐，并明确得到一些短语的对齐关系：

{ dj(我们可以比照其它工厂的做法拟定计划。) / s(We can draw up our plan in the light of the experience of other factories.) }

{ vp(拟定计划) / vp(draw up our plan) }　{ np(其它工厂) / np(other factories) }

{ vp(可以比照其它工厂的做法拟定计划) / vp(can draw up our plan in the light of the experience of other factories) }

{ vp(比照其它工厂的做法拟定计划) / vp(draw up our plan in the light of the experience of other factories) }

不过，仅仅依靠这两条原则，还是有很多短语的对齐无法确定。

没有确定对齐关系的汉语短语包括：

ap(其它工厂的)

np(其它工厂的做法)

pp(比照其它工厂的做法)

没有确定对齐关系的英语短语包括：

np(the light)

np(the experiences)

pp(of other factories)

np(the experiences of other factories)

pp(of the experiences of other factories)

np(the light of the experiences of other factories)

pp(in the light of the experiences of other factories)

所以这时真正需要我们通过短语对齐算法找到的对齐只有以下几个：

{ ap(其它工厂的)
pp(of other factories) }

{ np(其它工厂的做法)
np(the experiences of other factories) }

{ vp(比照其它工厂的做法)
pp(in the light of the experiences of other factories) }

其他未对齐的短语都对齐到空。

5.1.4 短语结构对齐的问题和难点

我们看到，交叉约束实际上是一个很强的约束条件。在种理想的情况下，假设句法分析和词语对齐全部正确，那么仅根据交叉约束就可以得到很多（也许是大部分）正确的短语对齐，但还是有一些短语对齐的歧义需要消解。

实际上，句法分析和词语对齐都是很容易出错的，而且错误率还相当高。如果我们简单地在句法分析和词语对齐的基础上进行短语对齐，那么短语对齐的正确率和召回率都会非常低。这是短语对齐所面临的主要困难所在。以下我们分别讨论如何克服这两种错误给结构对齐带来的困扰。

1．句法分析错误给短语对齐带来的困扰

为了解决克服句法分析的错误给短语对齐带来的困扰，现有的办法一般都是在句法分析阶段采取 N-Best 策略，也就是说，允许句法分析器输出多个可能的结果，并保留在一个类似线图的数据结构中，然后再由短语对齐程序在整个线图中搜索正确的短语对齐，最后输出最优的短语对齐结果。这样做不仅可以大大提高短语对齐的正确率和召回率，反过来还可以利用双语短语对齐的信息，对句法分析和词语对齐阶段给出的多个结果进行排歧，提高句法分析和词语对齐的正确率。

举例来说，假设汉语和英语的两个句子：

汉语：男孩用望远镜看见了那个女孩。

英语：The boy saw the girl with a telescope.

这里英语句子是一个典型的 PP-Attachment 歧义句子。“with a telescope”可以修饰动词短语“saw the girl”，也可以修饰名词短语“the girl”。不过通过双语短语对齐我们发现，“with a telescope”可以对齐到“用望远镜”，“saw the girl with telescope”可以对齐到“用望远镜看见了那个女孩”，但是“the girl with a telescope”却没有对齐的短语，这样我们就可以排除英语中的歧义结构（“with a telescope”修饰“the girl”）。

2．词语对齐错误给短语对齐带来的困扰

对短语对齐造成困扰的第二个因素是词语对齐。目前的词语对齐技术都还不够成熟，很难达到较高的正确率和召回率。词语对齐主要存在的问题有：

(1) 错误对齐：错误对齐通常是由于译文中出现了源文词的多个相同或者不同的对译词所造成的。比如说，一个源文句子中可能经常会出现两个或两个以上意思相近或者完全相同的词语，对应地，译文也就有多个意思相近或者完全相同的对译词，这些源文词和译文词之间的对译关系很容易造成混淆，从而导致错误的对齐。请看下面两个句子：

汉语：如果省略了路径名，那么就假设是当前路径。

英语：If the path name is ommited, the current path is assumed.

这里汉语中有两个“路径”，英语中有两个“path”，相应有两种可能的对齐关系：

A． 第一个“路径”⇔ 第一个“path”，第二个“路径”⇔ 第二个“path”

B． 第一个“路径”⇔ 第二个“path”，第二个“路径”⇔ 第一个“path”

还有的时候，虽然源语言中的词各不相同，但译文句子中也可能恰好出现了两个词都有可能是某个源文词的翻译，这时候同样也容易造成错误对齐。

(2) 部分对齐：由于源语言和目标语言的词语单位的大小不同，很可能造成一种语言中的词语对应到另一种语言中的一个短语，而仅仅进行词语对齐时可能会造成部分对齐，比如说如果汉语中将“南朝鲜”作为一个词，很可能会和英语中“South Korea”中的词语“Korea”对应起来。

(3) 覆盖率问题：人类语言的翻译是非常灵活的，双语词典只能覆盖实际翻译中的一小部分。引入统计方法或者义类词典，可以适当提高词语对齐的覆盖率，但同时也带来了正确率的下降。不管是双语词典还是通过共现得到的统计信息，都很难保证覆盖一个词所有翻译方法。在双语对齐中，覆盖率不高始终是一个很严重的问题。

词语对齐错误会给短语对齐带来非常严重的困扰。特别是错误的词语对齐结果，根据交叉约束原则，会一直往上传递到各个层次的短语对齐中，导致大量的短语对齐的错误。

从理论上说，短语对齐对于词语对齐也有排歧的作用。在上面给出的错误对齐的例子中，如果我们能够在短语对齐中判断出短语对齐“路径名”和“path name”对齐，“当前路径”和“current path”对齐，那么我们就可以排除 B 了。

不过，根据交叉约束原则，短语对齐要依赖于词语对齐的结果。如果没有正确的词语对齐结果，又如何能得到正确的短语对齐结果呢？这里存在一个“鸡生蛋”还是“蛋生鸡”的问题。

5.1.5 现有的短语结构对齐技术

目前在短语结构对齐方面的研究并不多。这方面的研究工作主要有[Kaji 1992]和[Imamura 2001]。

5.1.5.1 Kaji 的工作

[Kaji 1992]的工作目的是为基于实例的机器翻译（EBMT）抽取模板。整个对齐工作分为以下几个阶段：

1． 日语句子的句法分析

2． 英语句子的句法分析：

这里的两种语言的句法分析都采用 CYK 算法，得到的结果是一个三角形的矩阵，其中可以保留多种句法分析的结果，歧义保留到短语对齐阶段解决；

3． 日语句子和英语句子的词语对齐：

词语对齐采用词典查询的方法，当存在两个以上的对译词时，歧义保留到短语配对阶段解决；

4． 日语句子和英语句子的短语对齐：

短语对齐按照以下方式进行：自底向上搜索日语和英语的句子分析表，以发现对应的短语。对于每一个日语分析表中的短语 X，搜索英语句子分析表，以找到一个短语 Y，包含 X 的所有对应单词，而不含有所有 X 以外单词的对应单词。如果 Y 找到了，那么 X 和 Y 相互配对。如下图所示：

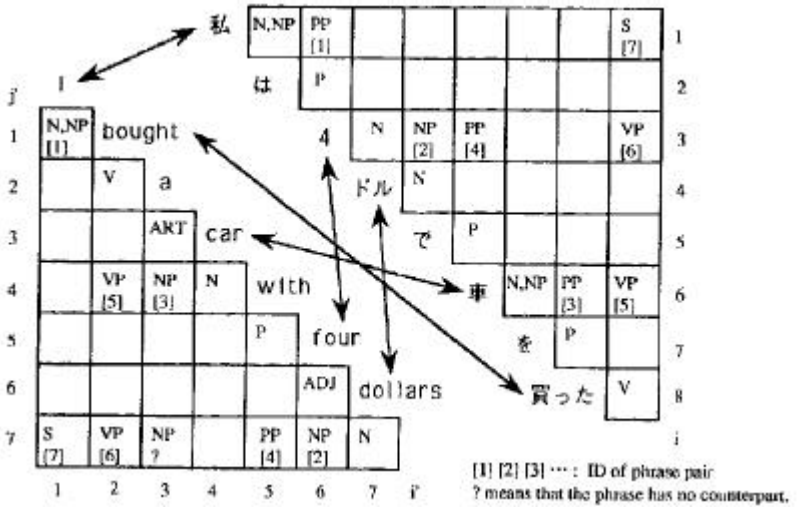


图 5.2 Kaji 所采用的基于 CYK 算法的句法分析和短语结构一体化算法

对于短语对齐过程中遇到的一些歧义现象，文中提出了一些原则性的解决办法：

对于词语对齐的歧义，文中假设，可以通过更大的短语的对齐信息来消除词语对齐的歧义。假设日语句子中的词语 J 在英语句子中有多个对应词。当一个包含 J 的短语 X 配对到英语句子中的短语 Y 时，就假设 J 的正确配对在 Y 中。

对于句法分析的歧义，文中假设，一个短语如果在对应的译文中找不到对齐的短语，那么该短语倾向于被拒绝。

对于短语对齐本身的歧义,就是说,如果一个短语存在着两个以上的可能被对齐的短语，文中提出的原则是：低层短语与低层短语、高层短语与高层短语对应。

Kaji 的工作是最早的双语短语结构对齐的工作，具有一定的开创性意义。它提出了短语结构对齐的基本思想，和处理对齐中遇到的某些问题的原则性解决办法。不过这篇文章总体上只是提出了一种解决问题的思路，想法还不够具体，有些问题没有考虑到，也缺乏实验数据的支持。

Kaji 的方法中对于短语结构对齐中的一个主要的难点问题——词语对齐错误（包括错误对齐、部分对齐和覆盖率问题）造成的困扰，并没有给出有效的解决办法。虽然提出了利用短语对齐的结果来排除词语对齐的歧义的设想，但这个设想还很初步，没有告诉我们应该如何解决词语对齐和短语对齐的相互依赖（即前面所说“鸡生蛋、蛋生鸡”）问题。另外，如果词语对齐中出现是一个没有歧义的错误对齐，那么由于交叉约束的影响，还是会对短语对齐造成很严重的干扰。

5.1.5.2 Imamura 的工作

[Imamura 2001]给出的双语对齐系统的流程如下：

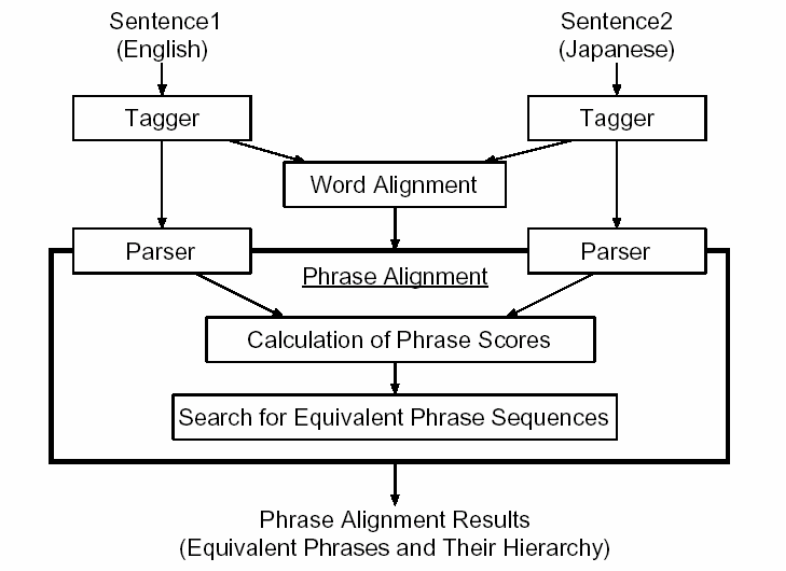


图 5.3 Imamura **的双语对齐系统结构**

文中提出的对齐过程是这样的：

1．**句法分析：标记并分析英语句子和日语句子，采用线图（**Chart**）分析法，多种分析结果可以并存。**

2．**词语对齐：通过词语对齐抽取得到对齐的词对（称为“词语链接”）。我们假设** W **个词语链接被抽取出来。**

3．**抽取对齐的短语：**

(1) **从所有的** W **个词语链接中任意选择** i **个词语链接（**$1 < i = W$**），抓取所有包含这些链接的句法结点（非终结符），并从英语树和日语树的叶结点中排除所有的其他单词链接。**

(2) **比较在步骤** 3(1)**中抓取的所有英语和日语结点的句法类型。当发现相同的结点类型时，就把它们作为对齐的短语。如果一个句子或者助动词短语的候选被获得，那么覆盖最大区域的候选就被选择出来。在其他有歧义的情况下，覆盖最小区域的候选被选择出来。**

(3) **对于所有的单词链接组合尝试步骤** 3 **和** 4**。**

4．**根据短语对齐的结果对句法分析得到的短语结构进行评分。**

5．**选择评分最高的源语言和目标语言的短语结构；如果没有形成完整的分析树，那么搜索评分最高的部分句法树序列。**

其中，两个短语对齐的基本条件是：

1. **相同的语义信息，即对应序列中的单词既没有不足（**deficiency**）也没有超出；**
2. **相同的短语类型，即短语属于相同的句法类型。**

这里，第一个条件相当于前面介绍的交叉约束原则；第二个原则是这篇文章提出的一个经验性原则，认为大部分情况下总是相同类型的短语互相对齐。

这篇文章给出了一个实验型的结果，实验规模用了不到 300 **个旅游用语句子，日语的句法分析正确率为** 44%**，英语的句法分析正确率为** 52%**。实验结果为短语对齐的正确率大约为** 87%**，每个句子平均抽取的短语数为** 5.6**。**

Imamura **的方法的方法比** Kaji **的工作有了较大的进步，提出了具体的实现算法，并给出了实验结果。不过这个算法在一些关键的步骤上还是没有交待清楚。对于短语结构对齐中的一个主要的难点问题——词语对齐错误（包括错误对齐、部分对齐和覆盖率问题）造成的**

困扰，Imamura 给出算法虽然可以在一定程度上解决，但还是存在很多问题。下面我们进行具体的分析。

首先分析一下 Imamura 算法中的步骤 3(1)。按照这个步骤，要“从所有的 W 个词语链接中任意选择 i 个词语链接（$1 < i = W$），抓取所有包含这些链接的句法结点（非终结符）”。这句话实际上有两种理解方式：

第一种理解，这里的“任意选择”是完全随机的任意选择，没有任何条件。按照这种理解，是有可能排除掉错误的词语对齐的。不过，这样的话，算法的时间复杂就是 $O(2^W)$，这个时间复杂度太高，特别是句子长度比较长的时候。

第二种理解，这里的“任意选择”是指选择任意几个连续的词语链接。这样的话，算法的复杂度将是 $O(W^2)$，这样，算法复杂度虽然不高，但是无法排除掉错误的词语对齐。

以下图为例，假设正确的词语对齐应该是：C⇔c，D⇔d，E⇔e，F⇔f，而正确的短语对齐应该是：A⇔a，B⇔b。但由于词语对齐错误，得到的词语对齐是：C⇔c，D⇔f，E⇔e，F⇔d。如果按照上面的第一种理解，当我们选择词语链接 C⇔c 和 E⇔e 的时候，我们就可以得到正确的短语对齐 B⇔b。不过，如果我们按照上面的第二种理解，我们就不可能得到这个词语对齐（B⇔b）。

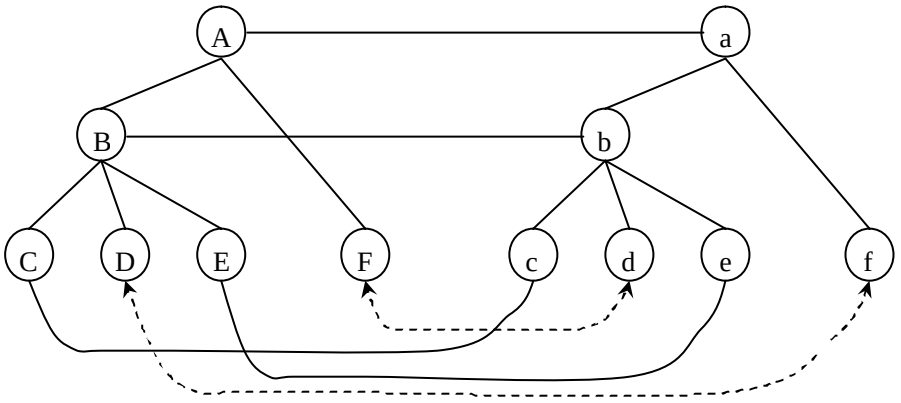


图 5.4 Imamura 的算法中词语对齐的组合对短语对齐的影响

Imamura 的文章中，并没有明确指出采用的是上面哪一种理解。从文章给出实验结果看，如果词语对齐的召回率提高，正确率下降，会导致短语对齐的总体正确率有所提高（但同时会造成所抽取的短语总数减少）。从这一点看，该算法按照上述第一种理解的可能性较大。由于 Imamura 的实验采用的是旅游对话语料，一般句子长度都不会很长（最终评价每个句子抽取的对齐短语数为 5.6），由于句子长度造成的效率问题可能也并不明显。不过，如果用这种方法来处理真实的文本语料，由于真实文本中的句子长度大大超过对话语料（几十个甚至上百个词的句子是很常见的），按照上述第一种理解，算法的时间复杂度 $O(2^w)$就会造成比较严重的问题。

我们再看一下 Imamura 算法的第 5 步。由于第三步是任意选择词语链的组合来获得短语对齐的候选，这就会使得短语对齐候选的数量会非常多，而且这些短语对齐候选之间会有大量的冲突（不满足交叉约束）现象存在。因此，如何选择一组评分值最高、同时又互相没有冲突（满足交叉约束）的短语对齐结果，并不是一个简单的问题，而是需要一个搜索算法。文章中并没有给出这个搜索算法的细节，也没有对这个算法的效果和复杂度进行分析。

另外，Imamura 文章中给出的实验只有 300 个句子，实验规模偏小，这些句子又都是旅游对话语料，句子较短，句子结构比较简单。如果面对大规模的文本语料，这个算法是否还能取得同样的效果，是无法确定的。

5.2 一种双语短语结构对齐的搜索算法

5.2.1 算法简介

本文中我们提出一种双语短语结构对齐的搜索策略。这种策略也是建立在双语句法（短语结构）分析和词语对齐的基础上，我们定义了一种局部对齐结构，采用柱形搜索（Beam Search）策略。和现有的方法相比，我们的算法优势在于：

1. 对词语对齐错误的容错能力：我们给出的方法可以排除词语对齐错误对短语对齐错误的干扰。因此，为了得到尽可能多的对齐短语，我们可以通过提高词语对齐召回率的办法，而不必太担心词语对齐正确率对短语对齐的影响；
2. 采用局部对齐结构，可以将完整的句法结构树的对齐问题很好地分解成每个结点及其子结点的对齐问题；
3. 由于采用了一种柱形搜索（Beam Search）策略，效率非常高，算法复杂度对于句子长度是线性的。

5.2.2 局部对齐

我们的算法执行过程是：在源语言短语结构树上，自底向上，计算每一个源语言结点的最佳的 N 个局部对齐。因为在每一个结点上都保留最佳的 N 个局部对齐结果，因此这是一种柱形搜索（Beam Search）。我们现在需要做的工作，就是定义一个局部对齐数据结构，能够将一棵完整的句法树的对齐分解成每个结点上的局部对齐。

一个局部对齐（Local Alignment）是一个数据结构，其中包含以下信息：

SrcNode（源语言结点）：每一个局部对齐对应唯一的源语言结点；

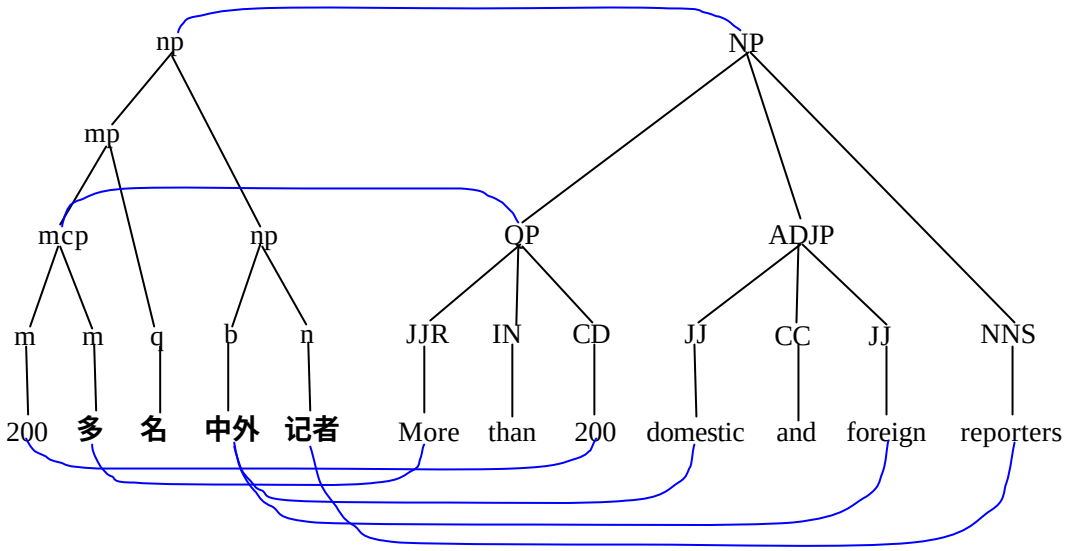
TgtNode（对应的目标语言结点）：与该源语言结点对应的目标语言结点，如果没有对应的目标语言结点，则此项为空；

TgtRange（对应的目标语言范围）：与该源语言结点对应的目标语言句子片断范围，包括第一个单词序号和最后一个单词序号；

Score（评分）：对当前子树对齐结构的评分；

Children（子结点局部对齐组合）：由于每个子结点都可能对应多个局部对齐，因此必须规定与当前局部对齐相关的各子结点的局部对齐。

以下图为例，我们给出下图中几个源语言结点的局部对齐结构：



(1)
SrcNode：m(200)
TgtNode：CD(200)
TgtRange：200
Score：1.0
Children：Empty

(2)
SrcNode：m(**多**)
TgtNode：JJR(more)
TgtRange：more
Score：1.0
Children：Empty

(3)
SrcNode：q(**名**)
TgtNode：Null
TgtRange：Null
Score：1.0
Children：Empty

(4)
SrcNode：b(**中外**)
TgtNode：JJ(domestic)
TgtRange：domestic
Score：0.0005
Children：Empty

(5)
SrcNode：b(**中外**)
TgtNode：JJ(foreign)
TgtRange：foreign
Score：0.03
Children：Empty

(6)
SrcNode：n(**记者**)
TgtNode：NNS(reporter)
TgtRange：reporter
Score：0.8
Children：Empty

(7)
SrcNode：mcp(200 **多**)
TgtNode：QP(more than 200)
TgtRange：more than 200
Score：…
Children：(1) (2)

(8)
SrcNode：mp(200 **多名**)
TgtNode：Null
TgtRange：more than 200
Score：…
Children：(7)

(9)
SrcNode：np(**中外记者**)
TgtNode：Null
TgtRange：domestic and … reporters
Score：…
Children：(4) (6)

(10)
SrcNode：np(**中外记者**)
TgtNode：Null
TgtRange：foreign reporters
Score：…
Children：(5) (6)

(11)	(12)
SrcNode：np(200 多名中外记者)	**SrcNode**：np(200 多名中外记者)
TgtNode：NP(more than … reporters)	**TgtNode**：NP(more than … reporters)
TgtRange：more than … reporters	**TgtRange**：more than … reporters
Score：…	**Score**：…
Children：(8) (9)	**Children**：(8) (10)

图 5.5 局部结构对齐示意图

从这个例子可以看到，每个源语言结点可以有多个候选的局部对齐。一旦确定了一个源语言结点的局部对齐，也就确定了其所有子孙结点的局部对齐。所以，搜索完整的句法分析树的对齐结果的过程就是求解其根结点的局部对齐的过程。

在局部对齐中，源语言结点（SrcNode）可以对应到一个目标语言结点（TgtNode），如(7)(11)(12)，也可以对应到空结点（TgtNode＝Null），如(8)(9)(10)。如果源语言结点对应到一个目标语言结点，那么其目标语言范围（TgtRange）就是目标语言结点所覆盖的范围；如果对应的目标语言结点为空，那么其目标语言范围（TgtRange）就是其所有子局部对齐（Children）所覆盖目标语言范围的总和。

对于叶结点来说（词语结点(1)～(6)），局部对齐的评分（Score）就是词语对齐的评分。对于非叶结点（短语结点(6)～(12)）来说，局部对齐的评分要根据其子局部对齐的评分和当前结点的对齐情况综合计算。评分的计算方法在后面讨论。

词语对齐的歧义对应到不同的词语局部对齐（(4)和(5)），这这种歧义会传播到短语对齐之中（(9)和(10)，(11)和(12)）。在这个例子中，这样处理实际上是不合适的，因为(4)和(5)实际上是部分对齐问题，而不是歧义的对齐。最好能够将“b(中外)”对齐到“ADJP(domestic and foreign)”。这是我们下一步要做的改进之一，这里不做详细的讨论。

5.2.3 短语结构对齐的柱形搜索(Beam Search)算法

整个短语结构对齐采用自底向上的柱形搜索（Beam Search）算法。

1. 对于每一个词语对齐，构造一个局部对齐，放入到源语言单词的局部对齐列表中；这个局部对齐的评分就是词语对齐的评分；
2. 对于每一个源语言单词，对其局部对齐列表按照评分从大到小进行排序；
3. 如果源语言单词的局部对齐不超过 N 个，那么给该源语言单词增加一个局部对齐，该局部对齐的目标语言结点为空（Null），评分（Score）为 0；
4. 自底向上对每一个源语言结点执行局部对齐归并算法，最终得到源语言根结点的对齐列表（见下一节的局部对齐归并算法）；
5. 取根结点局部对齐列表中评分最高的局部对齐，如果目标语言结点不是目标语言句法树的根结点，那么设置该目标语言结点为目标语言句法树的根结点；（全局约束原则）
6. 遍历根结点局部对齐列表中评分最高的局部对齐及其所有子孙结点局部对齐，如果目标语言结点不为空，那么输出该局部对齐的源语言结点和目标语言结点作为对齐的短语结点。

我们看到，初始化的过程中，我们会给每个源语言词语加入一个空词语对齐（除非与该源语言词语对齐的目标语言词语超过 N 个）。这样的话，就可以保证错误的词语对齐都有可

能在短语对齐的过程中被淘汰，而不会对短语对齐造成干扰。

5.2.4 局部对齐的归并

在搜索过程中，对于每一个源语言结点，保存一个局部对齐列表，根据柱形搜索的原则，表中最多保留 N 个局部对齐。

根据当前结点所有子结点的局部对齐列表，可以构造出当前结点的局部对齐列表，这个过程称为局部对齐的归并。

局部对齐的归并算法如下：

1．选择子结点局部对齐组合（Children）：对于当前结点的每一个子结点，选择唯一的一个局部对齐，构成当前局部对齐的子结点局部对齐组合；如果所有的子结点局部对齐组合都已选用，那么跳到 8，否则转到 2；
2．计算这些子的局部对齐的目标语言范围是否有重叠，如果重叠，则归并失败，回到 1，重新选择子局部对齐组合；
3．计算目标语言范围（TgtRange）：目标语言范围的左边界是其所有子局部对齐的左边界中的最小值，右边界是其所有子局部对齐的右边界中的最大值；
4．计算覆盖目标语言范围（TgtRange）的最小目标语言结点，作为目标语言结点（TgtNode）的首选（TgtNode1）；如果只有一个子结点局部对齐的目标语言结点不为空，那么就取它的父结点作为首选（避免当前结点和其子结点对应到同一个目标语言结点）；
5．取首选目标语言结点（TgtNode1）的父结点（TgtNode2）、祖父结点（TgtNode3）；
6．分别构造四个当前结点的局部对齐，其目标语言结点（TgtNode）分别取 Null、TgtNode1、TgtNode2 和 TgtNode3；
7．如果目标语言结点不为空，那么根据目标语言结点重新计算目标语言范围；
8．分别计算上述四个当前局部对齐的评分（见下一节局部对齐评分算法），并将它们加入到当前结点的局部对齐列表中，然后返回 1；
9．对当前结点的局部对齐列表根据评分从大到小进行排序，只保留其中评分最高的 N 个局部对齐，删除剩下的所有局部对齐;
10．返回。

5.2.5 局部对齐的评分

局部对齐的评分算法如下：

1．计算所有子结点局部对齐的评分之和作为基本评分（Score0）；
2．如果对应的目标语言结点为空（TgtNode=Null），那么将基本评分作为当前局部对齐的评分（Score=Score0），并返回；
3．如果目标语言单词（TgtNode）不为空，假设源语言结点（SrcNode）所含的单词数为 w1，目标语言范围（TgtRange）所含的单词数为 w2，如果 w2/ w1 > ? 或者 w1/w2>?（阈值 ?>1），那么将局部对齐的评分（Score）置为 0，并返回。
4．如果对应的目标语言结点不为空（TgtNode!=Null），那么将基本评分乘以一个系数 a 后作为当前局部对其的评分（Score=Score0×a ，1<a <2），并返回。

上面的第 3 步是为了排除明显的错误对齐，避免目标语言结点中所包含的单词数和源语言结点中所包含的单数比例过于悬殊。参数 ? 跟具体的语言和句子有关系。可以用当前目标

语言句子和源语言句子的长度比例适当放大得到。

第 4 步是对短语对齐的奖励，以保证尽可能多地得到对齐的短语。参数 a 越大，越倾向于大胆的进行短语对齐而忽略词语对齐，参数 a 越小，则表示尽可能信任词语对齐，而短语对齐则比较谨慎。

5.2.6 搜索算法的时间复杂度分析

搜索算法的时间复杂度分析如下：算法的时间消耗主要在第 4 步，就是短语自底向上的搜索阶段。假设句子长度（词数）为 L，那么短语结构树上的结点综述也是 $O(L)$。对于每一个结点，要执行局部对齐归并算法，假设有 m 个子结点，每个子结点都保留 N 个局部对齐，那么总共就有 N^m 种组合形式，不过我们可以将 m 个子结点同时归并的过程改成两两归并，两两归并只取其中 N 个最佳归并结果，再和其他子结点归并，这样归并的复杂度可以降低到 $O(N^2)$。于是，整个算法的复杂度是 $O(N^2L)$。可以看到，这个算法的复杂度跟句子的长度是线性关系。

5.3 实验及结果分析

5.3.1 实验方案

在进行短语对齐之前，我们要先进行双语句法分析和词语对齐工作。在我们的实验中，采用了以下技术完成这些任务：

1. 汉语词法分析：我们采用的汉语词法分析器就是前面介绍的基于层叠隐马尔可夫模型开发的 ICTCLAS。从前面的介绍可知，分词的正确率和召回率都在 95%左右，如果考虑词性标记的因素，正确率还更低一些。（下载网址：http://www.nlp.org.cn）；
2. 汉语句法分析：我们采用了清华大学周强博士开发的汉语句法分析器[周强 1999]，在分词和标注完全正确的情况下，该分析器的句法标记正确率和召回率都在 77%左右；
3. 英语词法分析：英语的句子切分和词语切分，我们利用了 IBM 公司公开的 Unicode 处理工具 ICU4C（下载网址：http://oss.software.ibm.com/icu/index.html）开发而成，英语的词性标注我们采用了 Eric Brill 的著名的英语词性标注软件[Brill 1995]，正确率非常高。（下载网址：http://www.cs.jhu.edu/~brill/）；
4. 英语句法分析：英语句法分析器我们采用 Charniak 公开的句法分析器[Charniak 2000]，短语标记的正确率和召回率约为 89%。
   （下载网址：http://www.cs.brown.edu/people/ec/）；
5. 汉语和英语的词语对齐：采用了我们自己开发的词语对齐程序。

由于我们是利用已有的句法分析程序，因此我们的短语对齐是建立在确定的句法分析的结果上，而不是像 Kaji 和 Imamura 那样建立在多个候选的句法分析结果基础上。不过我们给出的算法也完全适合有多个候选的句法分析结果的情形。

词语对齐程序也是我们课题组开发的，我们这里做一个简单的介绍。在词语对齐过程中，我们尝试了很多种办法，特别是我们重复了[Ker 1997]中提出的基于类的词语对齐实验，其中英语采用了 Wordnet 作为英语的义类词典（取顶部若干个层次），汉语采用了“同义词词林”作为义类词典，结果远不像论文中所声称的那样好。后来我们引入了一部采用人机互助

的方法从人读词典中提取出来的大规模英汉对照词典，使得对齐的正确率和召回率都有了很大的提高。从这一点也可以看出，专家知识（这里体现为双语词典）在自然语言处理中是可以起到非常重要的作用的。

目前，词语对齐系统主要使用了以下特征：

1、 词典信息；

2、 Dice 系数；

3、 相对扭曲度模型；

4、 两种语言的词性转移概率。

词语对齐的正确率大约为：80.2%，召回率大约是：66.4%，F1 值为 72.7%。

不过，现在这个词语对齐程序还是很不完善的，前面所介绍的词语对齐中的三个问题，即错误对齐、部分对齐和覆盖率问题，都没有很好解决。由于我们的词典比较大，相对来说，覆盖率问题解决得比较好一些。

由于我们的短语对齐程序对词语对齐具有容错的能力，因此，在词语对齐过程中，我们要求尽可能输出各种可能的词语对齐结果，而不必考虑这些对齐结果之间是否存在冲突。另外，词语对齐程序在给出一个词对的时候，同时会给出对该词对的评分（介于 0 到 1 之间）。

以下是一个词语对齐的例子：

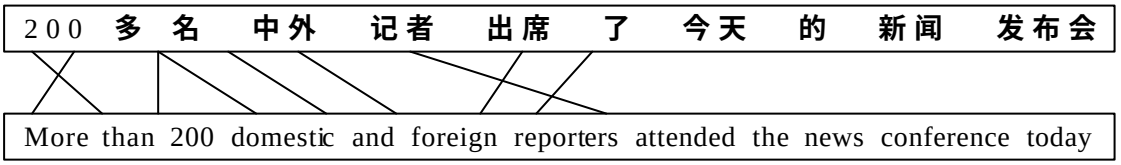


图 5.6 词语对齐结果举例

这个图上我们省略了词语对齐的评分。可以看到，汉语词“中外”被对齐到“domestic”和“foreign”两个词。实际上这里是个部分对齐的问题，并不是歧义或错误。不过形式上是一样的。

5.3.2 实验语料来源及规模

我们在一个含有 8009 个句子对的汉英双语语料库上进行了翻译模板抽取工作。这些语料库的来源包括：新闻报道：来自新华社、法新社、联合早报等，其中包含了宾州大学汉语树库（Chinese PennTreeBank）中所包含的 4000 多个句子及其译文，不过由于宾州树库所采用的词语和句法标记体系和规范与我们的句法分析器差异很大，为了实验效果的一致起见，我们没有利用该树库中的汉语分析树，而是采用了自动分析结果。

5.3.3 短语结构对齐的实例分析

在上面给出的 8009 个句子对的双语语料库上，我们首先进行了自动的汉语和英语词法分析、句法分析和双语词语对齐，然后在此基础上利用本章所介绍的算法对所有的句子对进行了短语对齐。

下面我们首先通过两个具体的例子来看一下对齐的结果。

第一个句对是一篇文章的标题，比较简单：

汉语：玻利维亚举行总统与国会选举

英语：Bolivia Holds Presidential and Parliament Elections

表 5.1 短语对齐的双语语料库样例（一）

<bead id="1">
<srctext no="1">**玻利维亚举行总统与国会选举** </srctext>
<srcword no="1">**玻利维亚**/nsf **举行**/v **总统**/n **与**/cc **国会**/n **选举**/vn
</srcword>
<srctree>dj_ZW:10(nsf:1|**玻利维亚** vp_PO:9(v:2|**举行** np_DZ:8(np_LH:7(n:3|**总统** cc:4|**与** n:5|**国会**) vn:6|**选举**)))</srctree>
<tgttext no="1">Bolivia Holds Presidential and Parliament Elections </tgttext>
<tgtword no="1">Bolivia/NNP Holds/NNP Presidential/JJ and/CC Parliament/NNP Elections/NNS</tgtword>
<tgttree>S:10(NNP:1|Bolivia VP:9(NNP:2|Holds NP:8(UCP:7(JJ:3|Presidential CC:4|and NNP:5|Parliament) NNS:6|Elections)))</tgttree>
<wordalignment>2:2/0.005786 3:3/0.03744 4:4/1 5:5/1 6:6/0.04684</wordalignment>
<phrasealignment>2:2/0.005786 3:3/0.03744 4:4/1 5:5/1 7:7/1.556 6:6/0.04684 8:8/2.404 9:9/3.615 10:10/3.977</phrasealignment>
</bead>

这里，结果文件是一个 xml **格式的文件，其中，**<bead>**表示一个句珠，**<srctext>**和**<tgttext>**分别表示汉语句子和英语句子，其中的属性** no **表示句子在句珠内部的编号，用于处理一个句珠内部出现两个或两个以上汉语或英语句子的情况。**<srcword>**和**<tgtword>**分别表示汉语词语切分标注和英语词语切分标注的结果，**<srctree>**和**<tgttree>**分别表示表示汉语和英语句法分析的结果。如果有多个汉语或英语句子，那么其句法树在都合并成了一棵树，所以**<srctree>**和**<tgttree>**没有** no **属性。**<wordalignment>**和**<phrasealignment>**分别是词语对齐和短语对齐的结果。其中每个对齐以“** a:b/score **”的形式给出，** a **和** b **分别是汉语词语（短语）和英语词语（短语）结点的编号（与句法树中的编号一致），** score **是这一对词语（短语）的评分。**

我们看到，上面这两个句子的词法分析、句法分析和词语对齐的结果都完全正确，词语对齐中有一个词语对齐没有找到：玻利维亚⇔Bolivia**。**

而短语对齐的结果也非常好：

总统与国会⇔ Presidential and Parliament

总统与国会选举⇔ Presidential and Parliament Elections

举行总统与国会选举⇔ Holds Presidential and Parliament Elections

玻利维亚举行总统与国会选举⇔ Bolivia Holds Presidential and Parliament Elections

所有对齐的短语都全部正确地找到了。

第二个句对的情况就复杂得多：

汉语：美国国家海洋暨大气总署发现，旨在供长距离侦测潜艇使用的低频声纳，如果在若干限制下运用，只会对海洋哺乳类造成微不足道影响。

英语：US State General Administration of Sea and Atmosphere found that the low-frequency sonar that was used for the long-distance monitoring of submarines would have little influence on sea mammals if used with some restrictions.

表 5.2 短语对齐的双语语料库样例（二）

<bead id="4">
<srctext no="1">**美国国家海洋暨大气总署发现，旨在供长距离侦测潜艇使用的低频声纳，如果在若干限制下运用，只会对海洋哺乳类造成微不足道影响。**</srctext>
<srcword no="1">[**美国**/ns **国家**/n **海洋**/n **暨**/cc **大气**/n **总署**/n]nt **发现**/v **，**/wd **旨在**/v **供**/vi **长距离**/d **侦**/vg **测**/v **潜艇**/n **使用**/v **的**/ude1 **低频**/b **声纳**/n **，**/wd **如果**/c **在**/p **若干**/m **限制**/vn **下**/f **运用**/v **，**/wd **只**/d **会**/v **对**/p **海洋**/n **哺乳类**/n **造成**/v **微不足道**/vl **影响**/vn **。**/wj </srcword>
<srctree>zj_XX:61(fj_BL:60(fj_BL:59(dj_ZW:42(nO:36(ns:1|**美国** n:2|**国家** n:3|**海洋** cc:4|**暨** n:5|**大气** n:6|**总署**) v:7|**发现**) wd:8|"，" vp_PO:58(v:9|**旨在** np_DZ:56(vp_PO:53(vi:10|**供** vp_LW:50(vp_ZZ:47(d:11|**长距离** vp_PO:43(vg:12|**侦** vp_PO:37(v:13|**测** n:14|**潜艇**))) v:15|**使用**)) ude1:16|**的** np_DZ:38(b:17|**低频** n:18|**声纳**)))) wd:19|"，" fj_BL:57(vp_XX:54(c:20|**如果** vp_ZZ:51(pp_JB:48(p:21|**在** sp_FW:44(np_DZ:39(m:22|**若干** vn:23|**限制**) f:24|**下**)) v:25|**运用**)) wd:26|"，" vp_ZZ:55(d:27|**只** vp_ZZ:52(v:28|**会** vp_ZZ:49(pp_JB:45(p:29|**对** np_DZ:40(n:30|**海洋** n:31|**哺乳类**)) vp_PO:46(v:32|**造成** vp_PO:41(vl:33|**微不足道** vn:34|**影响**))))))) wj:35|"。")</srctree>
<tgttext no="1">US State General Administration of Sea and Atmosphere found that the low-frequency sonar that was used for the long-distance monitoring of submarines would have little influence on sea mammals if used with some restrictions. </tgttext>
<tgtword no="1">US/NNP State/NNP General/NNP Administration/NNP of/IN Sea/NNP and/CC Atmosphere/NNP found/VBD that/IN the/DT low-frequency/JJ sonar/NN that/WDT was/VBD used/VBN for/IN the/DT long-distance/JJ monitoring/NN of/IN submarines/NNS would/MD have/VB little/JJ influence/NN on/IN sea/NN mammals/NNS if/IN used/VBN with/IN some/DT restrictions/NNS ./.</tgtword>
<tgttree>S:62(NP:47(NP:36(NNP:1|US NNP:2|State NNP:3|General NNP:4|Administration) PP:43(IN:5|of NP:37(NNP:6|Sea CC:7|and NNP:8|Atmosphere))) VP:61(VBD:9|found SBAR:60(IN:10|that S:59(NP:58(NP:38(DT:11|the JJ:12|"low-frequency" NN:13|sonar) SBAR:57(WDT:14|that VP:54(VBD:15|was VP:52(VBN:16|used PP:51(IN:17|for NP:48(NP:39(DT:18|the JJ:19|"long-distance" NN:20|monitoring) PP:44(IN:21|of NNS:22|submarines))))))) VP:56(MD:23|would VP:55(VB:24|have NP:49(NP:40(JJ:25|little NN:26|influence) PP:45(IN:27|on NP:41(NN:28|sea NNS:29|mammals))) SBAR:53(IN:30|if VP:50(VBN:31|used PP:46(IN:32|with NP:42(DT:33|some NNS:34|restrictions)))))))) ".":35|".")</tgttree>
<wordalignment>1:1/0.3864 2:2/0.06551 3:6/0.06551 5:8/0.04875 6:4/0.02294 7:9/1 11:19/1 12:20/0.03631 13:20/0.005562 14:22/1 15:16/0.07233

```
        17:12/2.624e-007 18:13/0.009986 20:30/1 21:15/0.01119 23:34/1
        24:3/0.0005462 25:32/0.0004867 27:23/5.422e-008 28:23/7.938e-005
        29:27/0.0867 30:28/0.2147 31:29/1 32:31/0.02271 33:24/0.00396
        33:25/2.779e-005 34:26/1</wordalignment>
<phrasealignment>1:1/0.3864 2:2/0.06551 3:6/0.06551 5:8/0.04875 6:4/0.02294
        36:47/0.8837 7:9/1 11:19/1 12:20/0.03631 14:22/1 37:44/1.1 47:51/3.204
        15:16/0.07233 50:52/4.915 53:54/5.407 17:12/2.624e-007 18:13/0.009986
        38:38/0.01498 56:58/8.133 20:30/1 23:34/1 39:42/1.1 44:46/1.21 51:50/1.331
        54:53/3.497 28:23/7.938e-005 29:27/0.0867 30:28/0.2147 31:29/1 40:41/1.822
        45:45/2.863 33:25/2.779e-005 34:26/1 41:40/1.5 49:49/6.545 57:56/15.06
        61:62/37.62</phrasealignment>
</bead>
```

在这个句子对中，英语的词法分析和句法分析基本上没有错误，而汉语的的词法分析也基本上没有错误。汉语的句法分析总体上也比较正确，不过还是出现了几处错误。具体来说就是：

表 5.3 **例句中的句法分析错误**

短语	标记结果	正确结果
测潜艇	vp	×
侦测潜艇	vp	np
长距离侦测潜艇	vp	np
长距离侦测潜艇使用	vp	×
供长距离侦测潜艇使用的低频声纳	vp	×
旨在供长距离侦测潜艇使用的低频声纳	vp	np
美国国家海洋暨大气总署发现……低频声纳	fj	×

我们再看一下词语对齐的结果：这个句子对产生了以下词语对齐的结果：

表 5.4 **例句的词语对齐结果**

美国⇔US	**国家**⇔State	**海洋**⇔Sea	**大气**⇔Atomsphere
总署⇔Administration	**发现**⇔found	**侦**⇔monitoring	**测**⇔monitoring
低频⇔low-frequency	**潜艇**⇔submarings	**使用**⇔used	**长距离**⇔long-distance
声纳⇔sonar	**如果**⇔if	**在**⇔was×	**限制**⇔restrictions
下⇔General×	**运用**⇔with×	**只**⇔would×	**会**⇔would
对⇔on	**海洋**⇔sea	**哺乳类**⇔mammals	**造成**⇔used×
微不足道⇔have×	**微不足道**⇔little	**影响**⇔influence	

可以看到，这里词语对齐的结果还是相当不错的，几个一对多和多对一的情况都给出了正确的结果。不过，也还是出现了几处错误（上面打×的对齐）。

我们再看一下本章所介绍的短语对齐程序所产生的短语对齐结果。这里一共得到了 17 **个短语对齐结果：**

表 5.5 例句的短语对齐结果

美国国家海洋暨大气总署	⇔	US State General Administration of Sea and Atmosphere	√
旨在供长距离侦测潜艇使用的低频声纳	⇔	the low-frequency sonar that was used for the long-distance monitoring of submarines	√
供长距离侦测潜艇使用	⇔	was used for the long-distance monitoring of submarines	√
长距离侦测潜艇使用	⇔	used for the long-distance monitoring of submarines	×
长距离侦测潜艇	⇔	for the long-distance monitoring of submarines	×
测潜艇	⇔	of submarines	×
低频声纳	⇔	the low-frequency sonar	√
如果在若干限制下运用，只会对海洋哺乳类造成微不足道影响	⇔	would have little influence on sea mammals if used with some restrictions	√
如果在若干限制下运用	⇔	if used with some restrictions	√
在若干限制下运用	⇔	used with some restrictions	√
若干限制下	⇔	with some restrictions	√
若干限制	⇔	some restrictions	√
对海洋哺乳类造成微不足道影响	⇔	little influence on sea mammals if used with some restrictions	×
对海洋哺乳类	⇔	on sea mammals	√
海洋哺乳类	⇔	sea mammals	√
微不足道影响	⇔	little influence	√
美国海洋……影响。	⇔	US State …restrictions.	√

我们看到，短语对齐的正确率还是相当高的。这里出现了三个短语对齐错误，主要是以下两个原因造成的：

1、 **句法分析错误，如“测潜艇⇔**of submarines**”、“长距离侦测潜艇使用⇔**used for the long-distance monitoring of submarines**”，其中，后面一对短语直接看上去并没有错误，不过在这个句子中，“旨在供长距离侦测潜艇使用”才是一个短语，“长距离侦测潜艇使用”在这个句子中并不是一个短语。**

2、 **空词语对齐对齐造成的错误，比如说：“长距离侦测潜艇⇔** for the long-distance monitoring of submarines**”。实际上，这里“长距离侦测潜艇”只应对齐到“**the long-distance monitoring of submarines**”。但由于英语词“**for**”没有对齐的汉语词，因此单纯从形式上看，这两种对齐可能性都存在，而这里选择了一个错误的对齐。就这个例子而言，我们可以通过引入汉语短语类到英语短语类的转移概率来解决这个问题，显然汉语名词短语对齐到英语名词短语的概率比对齐到英语介词短语的概率更高得多。**

词语对齐的错误基本上没有对短语对齐造成影响。相反，通过短语对齐，以下词语对齐被排除掉了：

表 5.6 被排除掉的短语对齐

测⇔monitoring	在⇔was×	下⇔General×	运用⇔with×
只⇔would×	造成⇔used×	微不足道⇔have×	

这里也有一个错误，原来下面的词语对齐是完全正确的：

侦⇔monitoring，**测**⇔monitoring

但由于汉语词法句法分析的错误，使得“侦”、“测”这两个词处在了汉语句法树的不同层次上，从而导致短语对齐的时候不可能对齐到同一个英语词。

5.3.4 实验结果及分析

为了对短语对齐的结果进行一个总体评价，需要构造一个较大规模的短语对齐语料库。而短语对齐语料库的构造涉及到词法分析、句法分析、词语对齐等多方面因素，非常复杂，因此需要耗费大量的人力物力。由于这种语料库缺少公共的资源，而在本文的工作中也没有足够的投入去单独开发一个这种语料库，因此我们从整个 8009 个句子对的语料库中随机抽取了 120 个句子对，通过人工进行词语对齐和短语对齐的方法，来估计整个算法的效果。

词语对齐和短语对齐的评价指标是正确率、召回率和 F1 值，这几个指标的定义如下：

$$\text{对齐正确率}=\frac{\text{找到的正确对齐数}}{\text{找到的所有对齐数}}$$

$$\text{对齐召回率}=\frac{\text{找到的正确对齐数}}{\text{实际的正确对齐数}}$$

$$\text{对齐F1值}=\frac{2\times\text{对齐正确率}\times\text{对齐召回率}}{\text{对齐正确率}+\text{对齐召回率}}$$

在整个测试集上计算上述指标时，我们首先对单个句子计算上述指标，然后再按照句子进行平均，得出整个测试集上的指标[1]。

我们对这 120 个句子的的测试集进行测试得到的结果为：

表 5.7 短语对齐的实验结果

短语对齐以前	词语对齐的正确率	73.4%
	词语对齐的召回率	74.7%
	词语对齐的 F1 值	73.8%
短语对齐以后	词语对齐的正确率	83.3%
	词语对齐的召回率	70.9%
	词语对齐的 F1 值	76.8%
	短语对齐的正确率	80.2%
	短语对齐的召回率	84.5%
	短语对齐的 F1 值	82.3%
	总对齐的正确率	81.5%
	总对齐的召回率	76.3%
	总对齐的 F1 值	78.7%

需要说明的是，我们这里统计短语对齐的召回率的时候，忽略了句法分析错误所造成的影响。也就是说，如果由于句法分析导致某些短语没有被正确分析出来，那么这些短语并不纳入正确的短语对齐的计数之中。不过，句法分析错误对短语对齐正确率的影响并没有被忽略。如果一个被句法分析模块错误地识别出来的短语被对齐到另一个短语，那么我们会把它纳入错误的短语对齐计数中。这样处理一方面是为了操作上的方便，因为我们没有足够的精

[1]这种指标称为宏平均指标，与之对应的是微平均指标，计算时首先统计整个语料库上正确的对齐数、找到的对齐数、和找到的正确对齐数的平均值，再用这几个平均值来计算正确率、召回率和 F1 值。从数值上看，宏平均指标一般比微平均的指标略低。

力去对错误的句法分析进行校正。另一方面，这样做也确实有其合理性，因为我们希望考察的是短语对齐算法的性能，而不是考虑短语对齐模块和句法分析模块的总体性能。

从以上测试结果的统计中我们可以看到，在短语对齐以后，词语对齐的指标有了较大变化，词语对齐的正确率提高了很多，而召回率略有下降。总体 F1 值有所提高。这说明短语对齐确实对词语对齐起到了排歧的作用。

单纯短语对齐的正确率和召回率都是比较高的，比词语对齐的正确率和召回率都要高很多。把短语对齐和词语对齐合并计算的总对齐正确率、召回率和 F1 值也比单纯的词语对齐对应指标要高很多，这说明这里给出的短语对齐搜索算法确实能够在较大程度上避免短语对齐错误对词语对齐的干扰。

我们简单对比一下这个实验结果和[Immamura 2001]给出的实验结果。[Immamura 2001]的实验结果中，短语对齐的正确率大约为 85%左右，看上去比我们的结果要好（Immamura 的论文中没有给出召回率指标）。不过，他们的测试数据比我们的数据要简单得多。他们只采用了 300 个日英句子对，而且这些句子全部是基本的旅行对话用语，不仅句子长度比较短，而且领域也非常集中。而我们的语料库中的句子全部是从真实的新闻语料中选取的句子，平均的句子长度在 20 个词以上。而语料库的规模也大得多，高出一个数量级以上。我们的语料库都是新闻语料，涉及的领域也比单纯的旅游领域广泛得多。所以实际上我们的实验规模和实验数据的复杂程度远远高于[Immamura 2001]的实验。在这样数据和规模上取得目前的实验结果，应该说是令人满意的。

另外，Immamura 的方法中采用句法分析和短语对齐的一体化技术，这也有利于克服句法分析错误对短语对齐的影响。而我们的实验由于条件所限没有采取这种技术，这对我们的算法的效果也造成了一定的影响。

为了了解这个短语对齐算法的效率，我们将实验中在汉语和英语的词法分析、句法分析和在词语和短语对齐上所花费的时间做一个对比：

表 5.8 语料库处理各阶段所花费的时间

汉语词法分析时间	54 秒
汉语句法分析时间	370 秒
英语词法分析时间	48 秒
英语句法分析时间	8718 秒
词语对齐时间	502 秒
短语对齐时间	107 秒

我们知道，假定句子的长度为 n，词法分析的时间一般是 O(n)，而句法分析的时间一般是 $O(n^3)$。从上面的结果看到，短语对齐的时间跟汉语和英语的词法分析时间基本上在一个数量级上，而比汉语和英语的句法分析时间都短得多[1]。符合我们前面的分析，即本文给出的短语对齐算法的时间复杂度为 O(n)。

5.3.5 实验结果的进一步分析

通过对实验结果的进一步分析，我们可以初步得到以下结论：

1、 本章提出的短语结构对齐算法可以在线形时间复杂度下实现高效的短语对齐，短

[1] 以上数据中，由清华大学周强博士等人开发的汉语句法分析器由于采用了特别的优化算法[周强 1999]，所以速度相当快。

语对齐的正确率和召回率比词语对齐的正确率和召回率都有较大的提高，这表示短语对齐确实可以在较大程度上避免词语对齐错误所造成的干扰；

2、 在源语言和目标语言句子的词法、句法分析正确，词语对齐结果基本正确（可以有少量错误）的情况下，这个搜索算法可以得到很好的短语对齐结果，并排除掉词语对齐中的错误；

3、 词法分析、句法分析的错误会对短语对齐的结果造成较大的干扰。这种错误对短语对齐造成的影响有以下几种可能：

(1) 如果句法分析得到的短语是一个正确的短语，只是短语标记错了，那么短语对齐基本上都不会出错。

(2) 如果句法分析得到的短语是不是一个正确的短语，而且与正确的短语有很严重的错位，那么对于这个短语一般不会产生任何短语对齐的结果。因为严重的句法分析错位会使得在短语对齐中很难找到满足交叉约束的短语对。这样就使得短语对齐的召回率下降，但并不会对短语对齐的正确率造成影响，对词语对齐的指标也没有多大影响。

(3) 如果句法分析得到的短语是不是一个正确的短语，而且与正确的短语有少许的错位，那么就很有可能会产生一个错误的短语对齐结果。因为这时候很有可能把错位处的词语对齐当作错误的词语对齐排除掉，从而仍然可以得到一个错误的短语对齐的结果。这种会使得词语对齐和短语对齐的所有指标同时下降，造成很糟糕的后果。

实际上，词法分析和句法分析的正确率越高，这个短语对齐搜索算法的效果就会越好。现有词法分析和句法分析算法的不成熟（如汉语句法分析的正确率和召回率只能达到 77%，这在目前已经是很高的水平了），也严重影响了这个本算法的效果。前面我们介绍了，如果我们在词法分析和句法分析阶段保留多个结果，就可以很大程度上克服词法分析和句法分析导致的短语对齐错误。由于我们的实验中采用了独立的词法分析和句法分析系统，因此没有能做到这一点。在以后的实验中我们可以在这方面加以改进。

4、 在本文的算法中，词语对齐的错误在短语对齐中可以被排除掉，这是有一定条件的。可以想象，如果所有的词语对齐都是错误的，当然不可能产生正确的短语对齐结果。要在短语对齐的过程中排除掉错误的词语对齐，除了要求有正确的句法分析结果外，还要求词语对齐的错误比例不能太高，分布上也不能太集中。也可以这么理解，这个算法本身提供了一种比较好的择优汰劣的机制，一些错误的词语对齐会通过交叉约束机制所造成的竞争，被周围正确的词语对齐所淘汰。但是，如果错误的词语对齐比较集中的话，那么它们可能就不会与其他正确的词语对齐发生交叉约束的冲突，也就很难排除掉。

5、 造成短语对齐错误的另一个原因就是空的词语对齐，正如前面的例子所示。空的词语对齐会造成多个短语都可以和另一种语言的同一个短语对齐，从而引起对齐的歧义。空的词语对齐有两种可能：一种是应该对齐到空的词语，比如说一些虚词；另一种是有对译词、但没有被词语对齐程序所召回的词语。这两种可能性都存在。解决这一问题的办法是在短语对齐的评分机制中引入其他特征，比如说短语标记的转换概率等等。

5.4 小结

本章中，我们给出了一个双语短语结构对齐的搜索算法。

算法的关键是定义了一个局部对齐数据结构，通过对子树局部对齐的归并得更大的句法树的对齐。通过一个柱形搜索（Beam Search）策略进行搜索。

算法执行的效率很高，时间复杂度是 $O(N^2L)$，其中 N 是柱形搜索的宽度（每次保留的中间结果数），L 是句子的长度。

这个算法的最大特点是可以在很大程度上避免词语对齐错误对短语结构对齐的干扰，这样我们就可以在词语对齐阶段尽可能提高词语对齐的召回率，以确保短语对齐有较高的覆盖率。

这个算法还可以通过以下方法加以改进，这也是我们下一步的工作目标：

1．部分词语对齐的处理策略：我们看到，词语的部分对齐问题目前还没有很好解决，解决这一问题将对短语对齐的效果有较大的改善；

2．对齐算法与句法分析算法的一体化：输出 N-Best 的句法分析结果，通过词语对齐和短语对齐，选择总体最优的句法分析和对齐结果；

3．局部对齐评分方法的改进：现在的局部对齐评分方法中，对于对齐的结点，使用了一个奖励因子。可以尝试考虑其他的评分方法。比如说，引入短语标记之间的转换概率，或者考虑短语长度之间的转换概率，等等；

4．建立基于短语结构的统计翻译模型：统计翻译模型（如 IBM Model 1～5）既可以用于翻译，也可以用于对齐。建立统计翻译模型是解决对齐问题的比较理想的办法。因为建立统计翻译模型后，可以通过迭代训练调整参数，以达到最优的对齐效果。

第6章　短语结构转换模板的提取与应用

获得短语结构对齐的语料库以后，我们希望能够在这个基础上提取翻译模板。

翻译模板（Template），或者翻译模式（Pattern），是一个被广泛使用，但又没有统一定义的概念。本章中，我们将提出“短语结构转换模板”的定义，并给出其提取和应用的算法，为下一步在短语结构转换模板的基础上建立统计翻译模型打下基础。

6.1 基于模板的机器翻译概述

很多人的工作中都使用了翻译模板或者翻译模式的概念。不过这些概念在不同的场合下都有不同的含义，这里我们做一个简单的归纳和总结。

一些比较典型的基于模板（或者基于模式，下面模板）的机器翻译工作包括：

1. CMU 的泛化的基于实例的机器翻译方法[Brown 1996,2000]：这是对双语例句的一种简单抽象，变量通常是人名、城市名、时间等等，通过维护一个词表来对变量的取值进行控制；
2. 土耳其 Bilkent 大学的工作[Güvenir 1998]：他们提出了一种通过两对互译的实例互相比较来获得翻译模板的方法；通过比较两个实例句子中的相同部分和不同部分，将相同部分作为常量，不同部分作为变量，再建立源语言和目标语言变量之间的对应关系，从而得到翻译模板；
3. Takeda 的基于模式的机器翻译上下文无关语法[Takeda 1996]：将通常的上下文无关语法通过中心词约束进行细化，并在两种语言之间的上下文无关语法之间通过链接约束建立对应关系，文中还提供了一种增量式的学习方法，可以从实例库中学习翻译模板；
4. Kaji 的基于实例的机器翻译[Kaji 1992]：通过对双语的句法树进行短语对齐，然后抽取翻译模板，作为机器翻译的实例库来使用；
5. Och 的对齐模板技术：[Och 1998,1999]的工作试图在 IBM 的词语对齐模型基础上引入组块对齐模型。在组块内部，采用了一种基于类的模板描述方法。模板中只有变量，没有常量，所有变量都是通过词语自动聚类得到类别，这些类别实现了对词表的一个划分。两种语言模板之间的对应关系通过一个取值为 0 或 1 的矩阵来表示。

翻译模板一般来说具有以下共同的特点：

1. 由两个单语模板组成；
2. 单语模板是由一些变量或者常量构成的字符串；常量就是具体词，变量可以用词串进行替换；
3. 两个单语模板的变量之间必须存在对应关系，这种关系通常是一一对应的；
4. 模板可能含有约束条件。

各种不同的模板的区别主要体现在以下一些方面：

1. 对于变量的定义不同，具体来说有一下几种：
   b) 通配符：不做任何限制，可以和任何字符串匹配[Güvenir 1998]；
   c) 词语的有限集合：通过单语语料库进行词语聚类得到的类别，或者人工指定的词语有限集合[Brown 1996,2000][Och 1998,1999]；

d) **词类或短语类：通常是一些比较具体的实词类标记，如人名、城市名、一般地名，或者名词单元（**np**）、动词短语（**vp**）等等**[Kaji 1992][Takeda 1996]**。**

2**．翻译模板对应的语言单位的层次不同：**

e) **句子级：一个翻译模板对应于一个完整的句子；模板之间不能嵌套或递归**[Brown 1996,2000][Güvenir 1998]**；**

f) **组块级：翻译模板分为两层，一层是句子级，一层是组块级，句子级模板中的变量可以用一个组块级模板进行替换；也就是说，模板之间允许一层嵌套**[Och 1998,1999]**；**

g) **短语级：模板之间可以任意嵌套甚至递归**[Kaji 1992][Takeda 1996]**。**

3**．对模板的约束条件不同。**

6.2 短语结构转换模板定义

我们采用的模板和前面的一些做法都有所不同，我们称之为“短语结构转换模板（Phrase Structure Transduction Template，PSTT**）”。**

下面我们给出一个短语结构转换模板的定义：

1. **短语结构转换模板由两个带对齐关系的句法子树组成；**
2. **两棵句法子树的根结点互相对齐；**
3. **句法子树的所有非根非叶结点对齐到空；**
4. **句法子树的叶节点或者对齐到对应句法子树的叶节点，或者对齐到空；**
5. **如果句法子树的叶节点对齐到空，那么它必须是一个终结符结点（具体词）。**

模板的具体书写形式为：

1. **两棵句法子树之间用⇔分隔；**
2. **句法子树中每个结点用该结点的句法标记表示；**
3. **对于非叶节点，句法标记的右边有一对圆括号，圆括号内自左向右依次记录其每个子节点；（这样句法子树就表示为一个嵌套的括号结构）；**
4. **对于对齐的叶结点，在句法标记的右面加一个冒号和一个序号（正整数），两棵句法子树中序号相同的叶节点互相对齐，同一棵子树中不能有序号相同的叶节点。**

从形式上看，短语结构转换模板主要有以下特点：

1. **短语结构转换模板所描述的是子树到子树的转换，而不是序列到序列的转换。**
2. **模板标记可以是词类标记或者短语标记；**
3. **模板之间可以嵌套。**

6.3 短语结构转换模板举例

我们先来看从一个例子。根据下面三个对齐结果：

$$\left\{\begin{array}{l}\text{vp(v(拟定) n(计划))}\\ \text{vp(v(v(draw up)) np(t(our) n(plan)))}\end{array}\right\}\quad\left\{\begin{array}{l}\text{v(拟定)}\\ \text{v(draw up)}\end{array}\right\}\quad\left\{\begin{array}{l}\text{n(计划)}\\ \text{n(plan)}\end{array}\right\}$$

我们就可以得到一条短语结构转换模板：

vp(v:1 n:2) ⇔ vp(v:1 np(t(one’s) n:2))

这个模板完成以下两棵句法子树的转换：

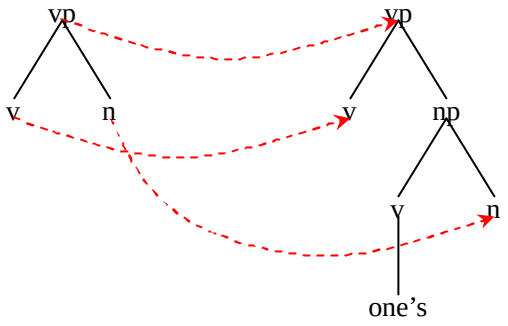


图 6.1 短语结构转换模板示意图

上面的例子中，双箭头⇔左边的是源语言（汉语）句法子树，右边是目标语言（英语）句法子树。冒号后面的数字 1 和 2，表示结点之间的对应关系，这里表示汉语的动词 v 译成英语的动词短语 vp，汉语的名词 n 译成英语的名词 n。

直观地理解，我们可以可以把一棵子树看成一个区域，这个图形的边缘是根结点和叶结点，图形的内部是非根非叶结点。图形的边缘有三类结点：第一类是唯一的一个根结点，第二类是若干个词语结点，第三类是若干个短语结点。一个短语结构转换模板由两棵树子树组成，而这两棵子树构成的区域的边缘结点之间存在某种对应关系：根结点和根结点对应，短语结点和短语结点一一对应，词语结点不和其他结点对应。

6.4 短语结构转换模板的提取

短语结构转换模板的抽取并不复杂。还是以前面的例子来说明：

汉语：我们可以比照其它工厂的做法拟定计划。

英语：We can draw up our plan in the light of the experience of other factories.

通过短语结构对齐，我们得到以下对齐的词语和短语：

r(**我们**) ？ r(we)

v(**可以**) ？ aux(can)

b(**其他**) ？ t(other)

n(**工厂**) ？ n(factories)

u(**的**) ？ p(of)

v(**拟定**) ？ v(draw up)

n(**计划**) ？ n(plan)

np(**其他工厂**) ？ np(other factories)

np(**其他工厂的**) ？ pp(of other factories)

np(**其他工厂的做法**) ？ np(the experience of other factories)

vp(**比照其他工厂的做法**) ？ pp(in the light of the experience of other factories)

vp(**拟定计划**) ？ vp(draw our plan)

vp(**比照……拟订计划**) ？ vp(draw our plan in … of other factories)

vp(**可以比照……拟订计划**) ？ vp(can draw our plan in … of other factories)

dj(**我们可以比照……拟订计划**) ？ s(we can draw our plan in … of other factories)

根据上面的对齐结果，我们可以抽取得到的以下模板：

dj(r:1 vp:2) ⇔ s(r:1 vp:2)

vp(v:1 vp:2) ⇔ vp(aux:1 vp:2)

vp(vp:1 vp:2) ⇔ vp(vp:2 pp:1)

vp(v(**比照**) np:1) ⇔ pp(p(in) np(np(t(the) n(light)) pp(p(of) np:1)))

np(ap:1 n(**做法**)) ⇔ np(np(t(the) n(experiences)) pp:1)

ap(np:1 u:2) ⇔ pp(p:2 np:1)

np(b:1 n:2) ⇔ np(t:1 n:2)

vp(v:1 n:2) ⇔ vp(v:1 np(t(one's) n:2))

r(**我们**) ⇔ r(we)

v(**可以**) ⇔ aux(can)

b(**其他**) ⇔ t(other)

n(**工厂**) ⇔ n(factories)

u(**的**) ⇔ p(of)

v(**拟定**) ⇔ v(draw up)

n(**计划**) ⇔ n(plan)

直观地理解，模板抽取的过程就是对源语言的短语结构树和目标语言的短语结构树分割成一些对应的子块，每一对对应的子块就构成了一个短语结构转换模板。下图给出了源语言句法树和目标语言句法树被分割后的情况：

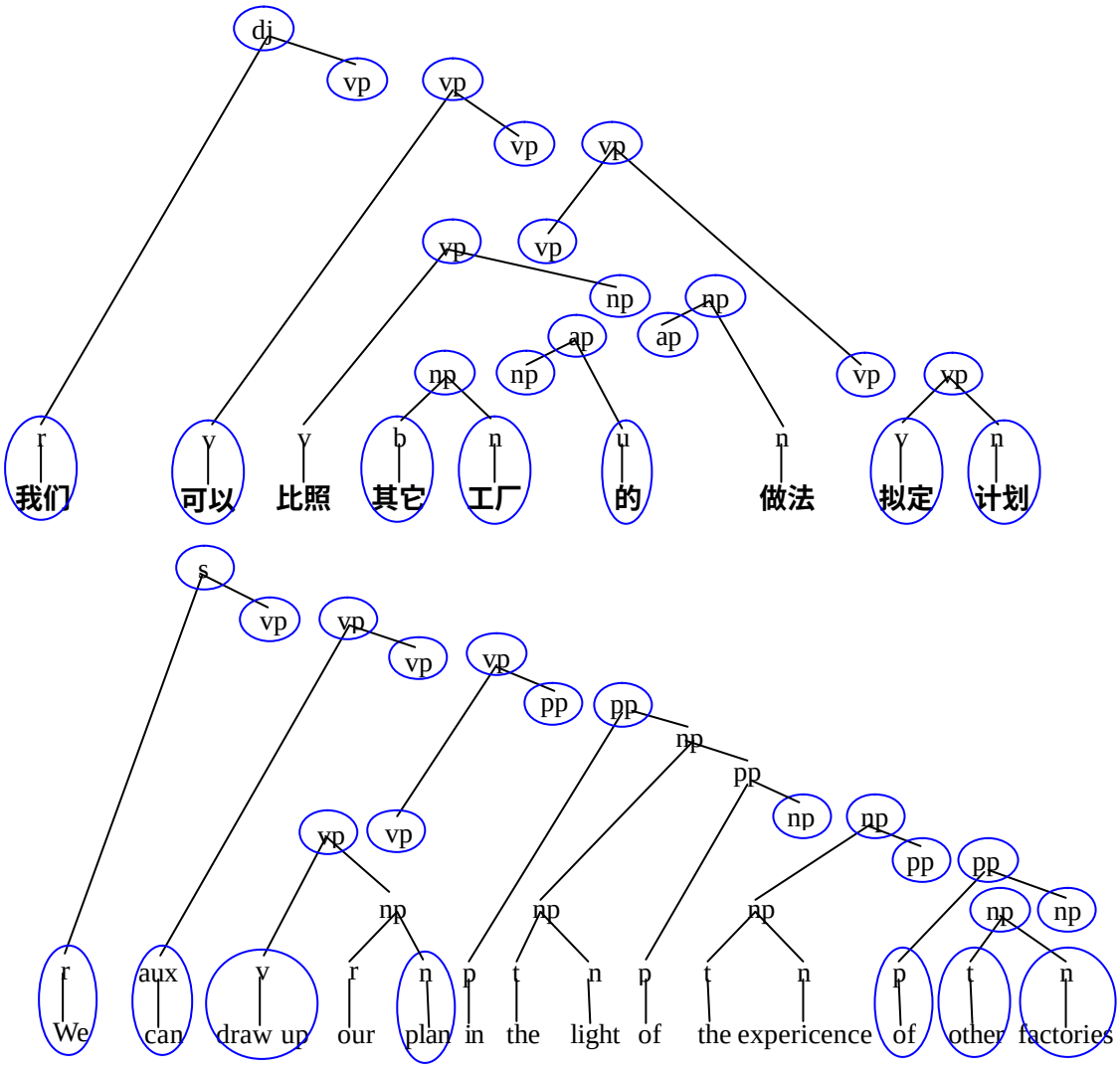


图 6.2 **短语结构对齐的句法树的分块**

我们可以看到，在源语言的短语结构树中，除了根结点和叶结点以外，每一个对齐的结点都被复制并分割成了两个结点，一个结点作为上层模板的叶结点，只保留到父结点的边，另一个结点作为下层模板的跟结点，只保留到所有子结点的边。这些对齐的结点构成了被分割得到的子块的边缘结点，而子块内部的结点都是没有对齐的结点。目标语言的短语结构树经过类似的分割，得到相同数量的子块，而且这些区域与源语言树分割产生的子块是一一对应的，根据每一对对应的源语言子块和目标语言子块就可以得到一个短语结构翻译模板。

抽取的算法直观描述如下：从源语言树每一个对齐的结点开始，进行一个自顶向下的深度优先搜索，与一般深度优先搜索不同的是，每次搜索到一个对齐的结点时就不再往下搜索，而是记录下一个编号，得到一棵源语言子树。然后对目标语言树进行一个对应的搜索，得到一棵目标语言子树。这两棵子树就构成了一个模板。

算法具体描述如下所示：

算法：转换模板抽取算法

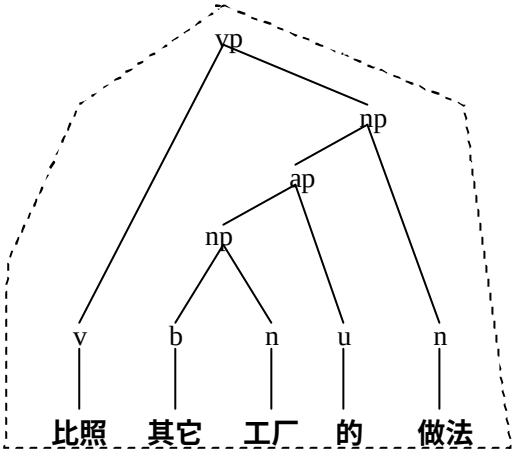
```
输入：
    对齐的句法树库
输出：
    转换模板集合以及每一个模板出现的频度
过程：
    将对齐结点序号 Number 赋值为 0；
    将模板字符串 Template 清空；
    对于源语言短语结构树中每一个对齐的结点 SrcRoot，执行以下操作：
        从 SrcRoot 开始，按照深度优先的顺序，对搜索到的每一个源语言结点 SrcNode 执行以下操作：
            向 Template 末尾输出 SrcNode 的句法标记；
            如果 SrcNode 是一个对齐的结点，那么：
                将对齐结点计数 Number 递增（自加 1，即 Number++）；
                向 Template 末尾输出冒号(:)和 Number 值；
                将二元组(SrcNode,Number)加入到一个结点序号对应表 NodeNumberTable 中；
            如果该结点是一个非对齐的叶结点，那么：
                向 Template 末尾输出左括号；
                向 Template 末尾输出叶结点所对应的单词；
                向 Template 末尾输出右括号；
            如果该结点是一个非对齐的中间结点，那么
                向 Template 末尾输出左括号；
                向 Template 输出 SrcNode 的每一个子结点（递归调用）；
                向 Template 末尾输出左括号；
        向 Template 末尾输出⇔符号；
        从 SrcRoot 的对应结点 TgtRoot 开始，按照深度优先的顺序，对搜索到的每一个目标语言结点 TgtNode 执行以下操作
            向 Template 末尾输出 TgtNode 的句法标记；
            如果 TgtNode 是一个对齐的结点，那么：
                在结点序号对应表 NodeNumberTable 中，查找 TgtNode 对应的 SrcNode 的序号 Number ；
```

向 Template 末尾输出冒号(:)和 Number 值；
如果该结点是一个非对齐的叶结点，那么：
向 Template 末尾输出左括号；
向 Template 末尾输出叶结点所对应的单词；
向 Template 末尾输出右括号；
如果该结点是一个非对齐的中间结点，那么
向 Template 末尾输出左括号；
向 Template 输出 TgtNode 的每一个子结点（递归调用）；
向 Template 末尾输出左括号；
在模板库中查找 Template
如果找到，那么给 Template 对应的频度递增（自加 1）
否则，那么将 Template 加入到模板库中，并将其频度设置为 1
输出模板库，返回。

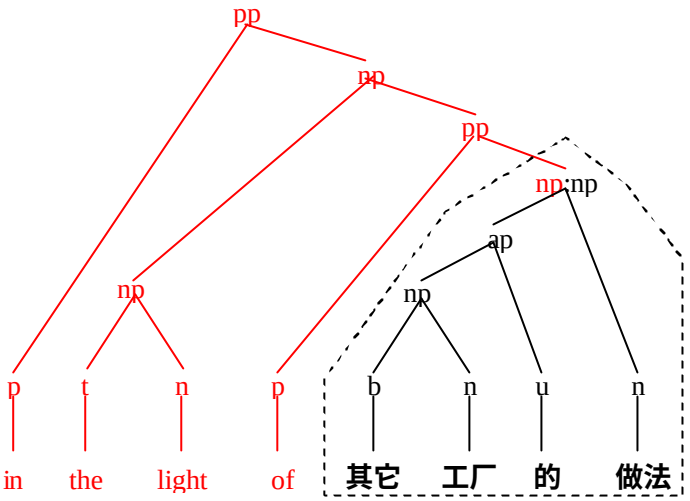
6.5 短语结构转换模板的应用——基于模板的转换

利用模板库，我们可以把一棵源语言的短语结构树转换成目标语言的短语结构树。转换算法的设计有很多种办法，既可以自顶向下进行，也可以自底向上进行。在进行过程中，既可以采用深度优先搜索策略，也可以采用广度有限或其他搜索策略。

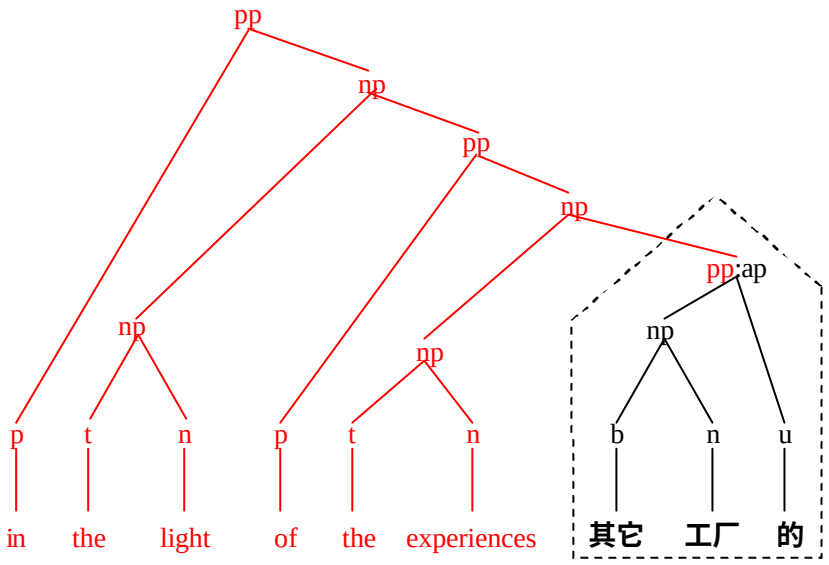
基于模板的转换算法并不复杂，我们这里不详细给出。下图我们通过一个例子来演示一棵汉语树通过转换模板自顶向下转换成英语树的过程：



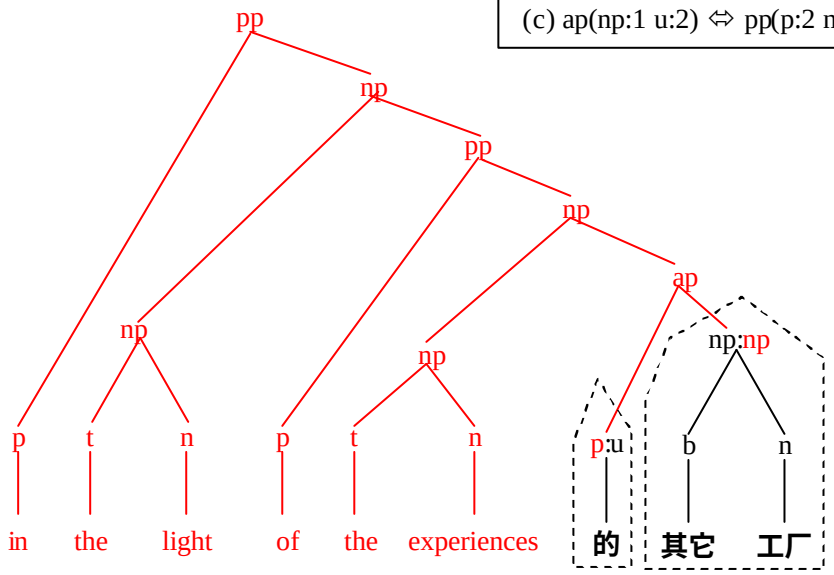
(a) vp(v(**比照**) np:1) ⇔ pp(p(in) np(np(t(the) n(light)) pp(p(of) np:1)))



(b) np(ap:1 n(**做法**)) ⇔ np(np(t(the) n(experiences)) pp:1)



(c) ap(np:1 u:2) ⇔ pp(p:2 np:1)



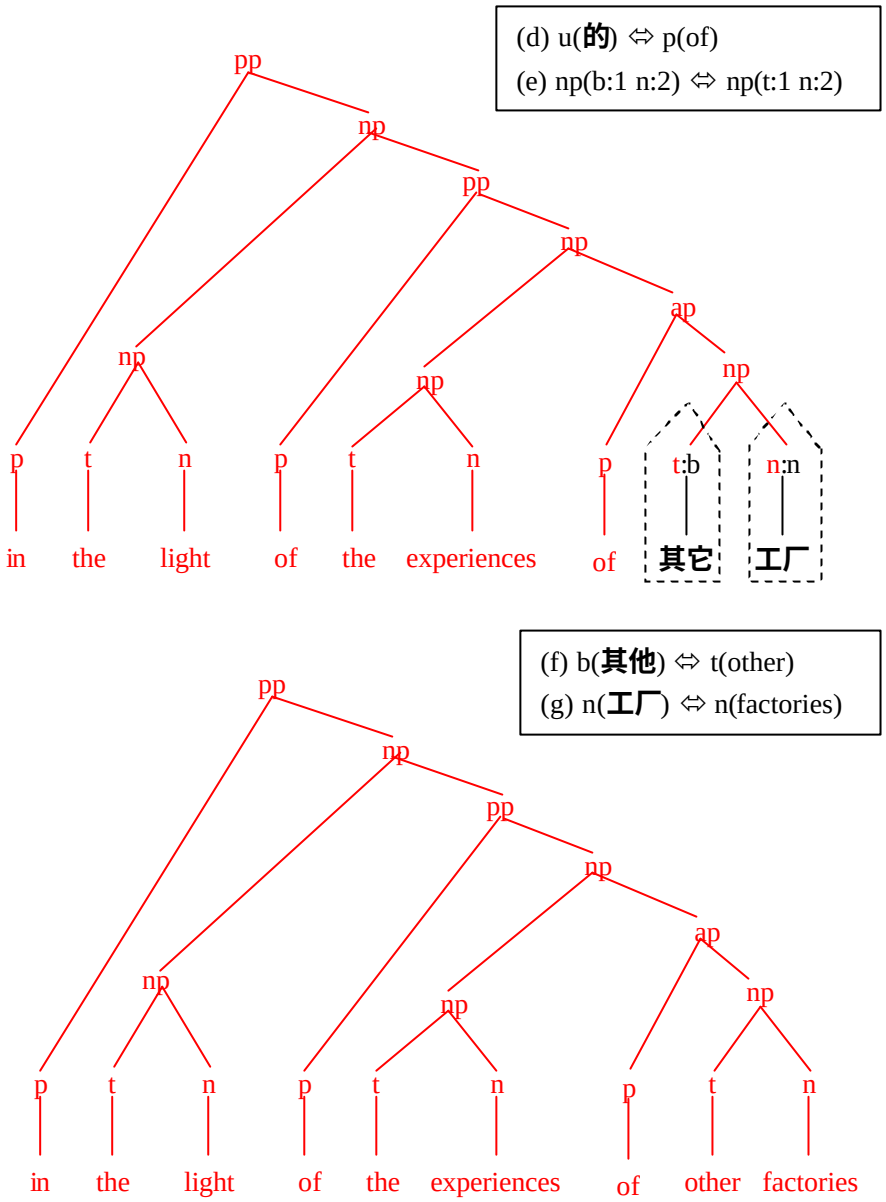


图 6.3 应用短语结构模板自顶向下进行进行句法树转换的过程示意图

注意这里有些生成的英语词是复数。单复数的问题在汉英机器翻译中是难点之一。采用基于模板的方法，可以解决一部分这类问题。这个问题我们不在这里做详细的讨论。

当模板库的规模很大的时候，我们在转换过程中会遇到大量的模板的选择问题（歧义）。而且这种歧义的组合会随着句子长度的增加呈指数增长。为了获得一个最好的翻译结果，我们需要解决两个问题：

1. **评分问题；如何比较两个转换哪一个更合理？如果我们采用统计方法来解决这个问题，那么这就是一个统计模型问题。也就是说，如何建立一个统计模型，来估计一种转换的概率？**
2. **搜索问题：如何在尽可能短的时间内找到最合理的转换？**

这是本项研究今后将要进行的主要工作。本文不对此展开讨论。

6.6 实验结果

6.6.1 实验语料的来源及规模

模板提取的实验语料来自于 5.3 节所介绍的实验所产生汉英双语短语结构对齐的语料库。语料库全部来自都是一些著名新闻机构发布的正式新闻，总规模为 8009 个汉英句子对。详情请参见 5.3.2 小节。

6.6.2 实验结果分析

通过在这个双语语料库上进行汉英双语的词法分析、句法分析、词语对齐、短语对齐和模板抽取，我们得到了一个很大的模板库，总共有 52,551 个模板。这些模板实际上又可以分为两类，一类是词语模板，刻画了词语之间的转换关系，总共有 27813 个，另一类是短语模板，刻画了短语结构之间的转换关系，总共有 24738 个。下面我们对这些模板的频度、长度、可信度和覆盖度分别进行分析。

6.6.2.1 模板频度分析和长度分析

首先我们对模板库中模板的频度和长度的分布情况做一个统计分析，以了解整个模板库的概况。

模板库中的模板出现的频度分布如下：

表 6.1 模块的频度分布

模板频度	词语模板数	短语模板数	合计
1	20,156	23,342	43,498
2	3,320	752	4,052
3	1,308	224	1,532
4	779	106	885
>=5	2,250	314	2,564
总数	27,813	24,738	52,551

我们可以看到，模板的分布是非常分散的。绝大部分的模板都只出现过一次，对于短语模板来说，这一现象更为明显：94.4%的模板都只出现过一次。

下面给出出现频度最高的 10 个词语模板和短语模板：

表 6.2 频度最高的词语模板

源文模板	译文模板	频度
v|说	VBD|said	880
p|在	IN|in	669
ns|中国	NNP|China	625

nt|新华社	NNP|“Xinhua News Agency”	419
d|将	MD|will	372
ns|美国	NNP|US	333
q|美元	NNS|dollars	217
n|经济	JJ|economic	209
nt|法新社	NNP|AFP	208
n|政府	NN|government	198

表 6.3 频度最高的短语模板

源文模板	译文模板	频度
vp_BH(wkz|" (" vi|完 wky|") ")	S("("|"(" VP(NN|End ")"|"-rrb"))	280[1]
pp_JB(p:1 np_DZ:2)	PP(IN:1 NP:2)	162
vp_PO(v:1 np_DZ:2)	VP(VB:1 NP:2)	158
np_DZ(n:1 n:2)	NP(JJ:1 NNS:2)	112
dj_ZW(np_DZ:1 vp_ZZ:2)	S(NP:1 VP:2)	86
np_DZ(a:1 n:2)	NP(JJ:1 NNS:2)	83
sp_FW(np_DZ:1 f:2)	PP(IN:2 NP:1)	80
mp_DZ(m:1 q:2)	NP(CD:1 NNS:2)	76
ap_ZZ(d:1 a:2)	ADJP(RB:1 JJ:2)	71
dj_ZW(np_DZ:1 vp_PO:2)	S(NP:1 VP:2)	71

词语模板的作用，类似于一个概率词典。我们这里主要关心的是能够反映两种语言句法结构对应关系的短语模板，因此后面的分析主要都是针对短语模板的。

我们定义短语模板的长度为源语言模板和目标语言模板的结点数之和。

这里我们对于模板长度的意义做一个直观的说明：假设一个模板的源语言模板对应的子树为 SrcTree，目标语言模板对应的子树为 TgtTree，又假设 SrcTree 和 TgtTree 都含有 n 个叶结点，且都是二叉树，那么 SrcTree 和 TgtTree 将分别含有 2n－1 个结点，于是这个模板的长度将是：2×（2n－1）。反过来说，如果模板长度等于 20，那么这个模板的源语言子树和目标语言子树所含有的叶结点数大约为 5.5。考虑到实际的句法数未必都是二叉树，实际的模板中叶结点数可能略多，也就是 6～7 个结点。也就是说，长度为 20 的模板所表示的，一般相当于 6～7 个结点所组成的序列的结构转换关系。

下面我们来看一下短语模板的长度分布情况：

表 6.4 短语模板的长度分布

模板长度	模板数
10 以下	8,332
11～15	6,006
16～20	2,764
21～25	1,462
26～30	979
31 以上	5,195

[1] 这个模板出现次数最高，是因为语料库中含有宾州大学树库的双语对照文本，其中每一篇文章的结尾都是“（完）”，而对应的译文就是“（End）”。

我们看到，相当一部分的模板都是很长的，长度大于 20 的模板所占的比例约为 31%。理论上，模板长度越长，能够被匹配上的可能性也就越小，但一旦匹配上，翻译的准确率也就越高，当然这是在模板本身正确的基础上。

另外，模板的长度和频度是有关系的。我们看看下表：

表 6.5 短语模板长度和频度的关系

模板长度	模板总数	频度为 1 的模板数	频度为 2 的模板数	频度大于 2 的模板数
21 以上	7,636	7,616	13(0.17%)	7(0.09%)
11～20	8,770	8,590	133(1.52%)	47(0.54%)
10 以下	8,332	7,136	605(7.26%)	591(7.09%)

从这个表里我们可以看到，在长模板中，高频的模板非常少。在长度为 21 以上的模板中，99.7%的模板都是频度为 1 的模板。在长度为 11～20 的模板中，将近 98%的模板都是频度为 1 的模板。也就是说，绝大部分高频模板都是短模板。

6.6.2.2 模板可信度分析和覆盖度分析

所谓可信度，就是一个模板在翻译中值得信赖的程度。所谓覆盖度，就是这个模板库对于各种源语言结构的转换形式的覆盖程度。我们这里所说的可信度和覆盖度指的都是主观的判断而不是客观的指标。在没有找到合适的模板可信度客观评价指标和评价方法之前，我们主要采用主观判断的方法来对一个模板做出评价。评价的方式也主要是定性的评价，而不是定量的评价。

由于模板的形式是非常直观的，所以人类专家很容易判断一个模板是否可信。

模板的可信度和模板的频度有直接的关系。从上面给出的 10 个最高频的词语模板和短语模板来看，它们的可信度非常高，可以说都是非常常见的转换模式。一般而言，模板出现的频度越高，这种模板可信度也就越高。根据我们对所获取模板的初步分析，出现两次以上的模板可信度是非常高的。有一句西方谚语说：“幸福的家庭都是相似的，而不幸的家庭则各有各的不幸”。这是很有道理的。把这句话套用到这里也非常合适，我们可以说：“正确的模板都是相似的（所以会重复出现），而错误的模板则各不相同（所以出现频度都很低）。”

不过，我们并不能因此反过来认为频度为 1 的模板都是不可信的。实际上，频度为 1 的模板中虽然有不少错误，但还是有相当数量的模板是正确的模板。如果简单地抛弃所有频度为 1 的模板是非常不明智的。为了更好地利用这些模板，我们有必要从从各方面对这些模板进行更加深入细致的分析。

模板的长度跟模板的可信度也有一定的关系。一般来说，模板越长，包含的结点数越多，含有错误的可能性也越大。产生长模板的原因主要有以下两种：

1、 两种语言的结构差异，使得一些句法结构无法在较小的语言单位上进行对齐，而只能在更大的语言单位上对齐。这种语言结构的差异，又可以细分成三种情况：
   a) 两种语言的结构差异是真实的，产生的模板虽然很长，但却是正确的；
   b) 由句法分析错误所造成的语言结构差异，这种差异是不真实的，会导致错误的模板；
   c) 由词语对齐和短语对齐错误所造成的结构对齐的错位，也会造成虚假的语言结构差异，这种差异也是不真实的，同样会导致错误的模板。

2、词语对齐和短语对齐的缺失。由于一些应该对齐的词语或短语没有对齐，导致它只能同时出现在更大的短语模板中。

在上面的原因中，由原因 1a)和原因 2 导致的长模板都是可信的。由原因 2 导致的长模板虽然可信，但是有冗余。由原因 1b)和 1c)导致的模板则是错误的模板。

我们看两个例子。

第一个例子是汉语句法分析错误（上面的原因 1b）所造成的短语结构错误，从而形成了一个长度为 26 的模板：

dj_ZW(np_DZ(rzv|这 np_DZ(mp_DZ:1 n:2)) ⇔

vp_PO(v:3 dj_ZW(n:4 vp_JY:5))) S(NP(NP:1 PP(IN|of NP(JJ:2 NN:3 NNS:4))) VP(MD|will VP:5))

用图形表示，句法树的差异能看得比较清楚：

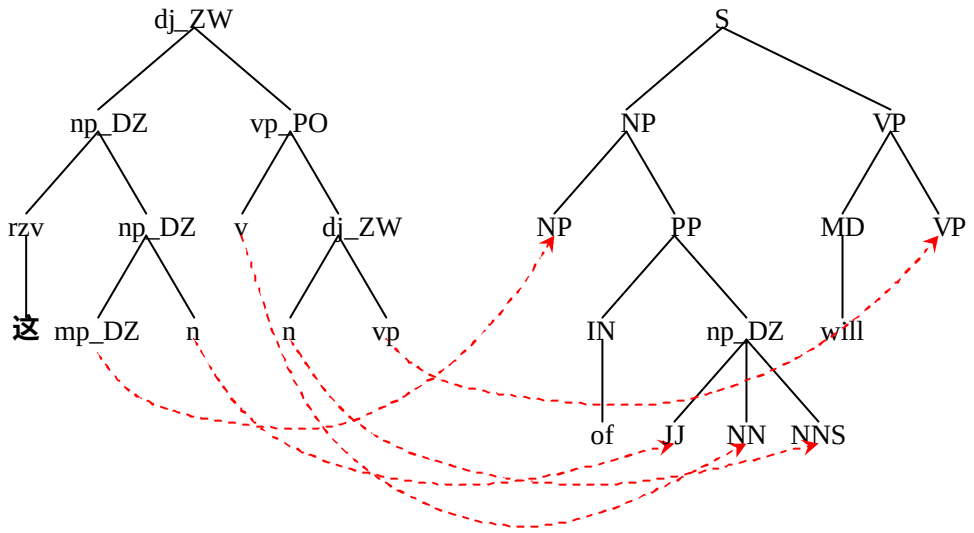


图 6.4 句法分析错误导致的长模板

第二个例子是由于词语对齐缺失（上面的原因 2）所造成的：

dj_ZZ(pp_JB(p|在 sp_FW:1) wd|"，" dj_ZW(np_DZ(np_DZ(nsf|塔吉克斯坦 n:2) nrf|"安德烈·阿布杜瓦利耶夫") vp_ZZ:3)) ⇔

S(PP:1 ","|"," NP(NP(NNP|Tajikistan POS|"'s") NN:2 NNP|Andrew NNP|Abudwaliev) VP:3)

实际上，大多数长模板都是由上面的多种原因共同造成的，其中都存在着或多或少的错误，由前面的分析已知，这样的模板一般都只出现一次。

模板的可信度还可以从其他一些特征来进行推测。比如说，有些模板中有几个实词都对齐到空，这种模板一般可信度都比较低。例如：

np_DZ(ap_ZZ(pp_JB(p:1 n|信用) ap_ZZ(d|较 a|高)) ude1|的 n:2)⇔

VP(TO:1 VP(VB|price PP(IN|in NN:2)))

通过多模型可信度分析，有助于我们在使用这些模板进行机器翻译时对模板选择并对翻译结果的可靠性进行估计。

我们希望，从语料库中抽取得到的短语模板，应该尽可能反映每一种句法结构的可能的转换形式。也就是说，我们希望短语模板的覆盖率尽可能要高。通过对模板库的初步分析发现，这些模板对于汉语句法结构和英语句法结构的对应关系具有较高的覆盖率。虽然我们这个语料库中只含有 8009 个句子对，数量并不多，不过，我们所获得的模板还是基本上覆盖了常见汉语句法结构到英语句法结构的主要转换形式。

下面我们以汉语中“形容词＋的＋名词”这种形式的名词短语的转换模板为例，对模板库的可信度和覆盖度做一个初步的分析。

表 6.6 汉语“形容词＋的＋n”形式的名词短语转换模板

汉语模板	英语模板	频度	可信度
np_DZ(a:1 ude1|的 n:2)	NP(JJ:1 NNS:2)	20	高
np_DZ(a:1 ude1|的 n:2)	NP(DT|a JJ:1 NN:2)	19	高
np_DZ(a:1 ude1|的 n:2)	NP(JJ:1 NN:2)	6	高
np_DZ(a:1 ude1|的 n:2)	NP(DT|an JJ:1 NN:2)	4	高
np_DZ(a:1 ude1|的 n:2)	NP(DT|the JJ:1 NN:2)	2	高
np_DZ(a:1 ude1|的 n:2)	VP(VB:1 NN:2)	1	中
np_DZ(a:1 ude1|的 n:2)	NP(DT|An JJ:1 NN:2)	1	高
np_DZ(a:1 ude1|的 n:2)	NP(DT|this JJ:1 NN:2)	1	高
np_DZ(a:1 ude1|的 n:2)	NP(DT|these JJ:1 NNS:2)	1	高
np_DZ(a:1 ude1|的 n:2)	NP(DT|The VBG:1 NN:2)	1	高
np_DZ(a:1 ude1|的 n:2)	NP(JJ:1 NN:2 NNS|aftereffects)	1	低
np_DZ(a:1 ude1|的 n:2)	NP(JJ:1 NN|township NNS:2)	1	低
np_DZ(a:1 ude1|的 n:2)	NP("PRP$"|its JJ:1 NN:2)	1	高
np_DZ(a:1 ude1|的 n:2)	NP("PRP$"|your JJ:1 NN:2)	1	高
np_DZ(a:1 ude1|的 n:2)	ADJP(JJ:1 PP(IN|of NN:2))	1	中
np_DZ(a:1 ude1|的 n:2)	NP(ADJP(RB|utterly JJ:1) NN:2)	1	中
np_DZ(a:1 ude1|的 n:2)	NP(DT|a JJ:1 JJ|joint NN:2)	1	中
np_DZ(a:1 ude1|的 n:2)	NP(NN:2 PP(IN|of NN:1))	1	高
np_DZ(a:1 ude1|的 n:2)	NP(DT|the JJ:1 NNP|United NNPS|Nations NN:2)	1	低
np_DZ(a:1 ude1|的 n:2)	NP(DT|a ADJP(JJ|practical CC|and JJ:1) NN:2)	1	低
np_DZ(a:1 ude1|的 n:2)	NP(NP(JJ:1 NNS|types) PP(IN|of NNS:2))	1	中
np_DZ(a:1 ude1|的 n:2)	NP(NP(DT|the VBG|operating NNS:2) PP(IN|in NP(JJ:1 NN|condition)))	1	低
np_DZ(a:1 ude1|的 n:2)	NP(NP(DT|the NN|ceremony) PP(IN|of NP(DT|the JJ:1 NN|ending)) PP(IN|of NP(DT|the NN:2)))	1	低

我们看到，在模板库中，关于这种汉语结构，我们得到了 23 个转换模板。其中，18 个模板只出现一次，有 3 个模板出现次数在 10 次以内，还有两个模板分别出现 19 次和 20 次。通过可信度分析，我们发现，其中 6 个模板可信度为“低”，5 个模板可信度为“中”，12 个模板的可信度为“高”。频度在两次以上的模板可信度均为“高”。这里可信度为“高”，表示该模板在很多情况下都是正确的。可信度为“中”表示该模板只有在非常特别的情况下才有可能正确，而且即使模板并不符合这种特定的情况，按照这种模板翻译得到的译文应该还是可懂的，不至于影响对译文的理解。可信度为“低”表示该模板是错误的，在任何情况下都不应该使用该模板进行翻译。

为了对这些短语结构转换模板的覆盖度有一个更直观的了解，我们将这些模板和一个基于规则的机器翻译系统[刘群 1997]中所使用的同一类型汉语句法结构的转换规则进行比较。这个规则库中的规则是对通过 4000 多个各种句型的汉语句子进行人工调试所产生的。下面给出这个规则库中的相应规则（为了节省篇幅，这里省略了规则中大部分合一约束和转换条件）：

表 6.7 一个机器翻译规则库中汉语“的字结构＋np”形式的名词短语的转换规则[1]

```
%% np->ap !np :: $.内部结构=组合定中,$.定语=%ap,$.中心语=%np, %ap.内部结构=的字,…
        => NP(!NP/np AP/ap)          /*比这再大的困难*/
        => NP(!NP/np C<who> CS/ap) %NP.NCASE=COMM
        => NP(!NP/np C<whom> CS/ap) %NP.NCASE=COMM /*我喜欢的女孩  */
        => NP(!NP/np C<whose> CS/ap) %NP.NCASE=COMM
        => NP(!NP/np C<whose> CS/ap) %NP.NCASE=COMM /*父亲是烈士的那个女孩*/
        => NP(!NP/np CS/ap)   /*他所说的话*/
        => NP(!NP/np D/ap) %D.DNEND=Yes /*这里的交通*/
      => NP(!NP/np P<of> NP/ap) %%NP.NCASE=COMM,… /*该协会的会员*/
      => NP(!NP/np P<of> NP/ap) %%NP.NCASE=COMM,…/*该协会的会员的总数*/
        => NP(!NP/np P<of> NP/ap) %%NP.RSUBC=RPOSS,… /*他的一个朋友*/
        => NP(!NP/np P<of> VP/ap ) %NP.NCASE=COMM,… /*旅行的最后一站*/
        => NP(!NP/np PP/ap) %NP.NCASE=COMM     /*修改后的系统*/
        => NP(!NP/np VP/ap) %VP.FORM=VN /*安装在桌子上的灯*/
       => NP(AP/ap  !NP/np)
        => NP(NP/ap !NP/np) %NP.NCASE=POSS,%NP.RSUBC=~RPOSS~RPERS
        => NP(T/ap !NP/np) %T.TNSUB=%NP.NSUBC… /*~他的<自行车的质量>*/
        => NP(T/ap !NP/np) %T.TSUBC=TPOSS /*谁的手套*/
        => NP(VP/ap  !NP/np)   %VP.FORM=VN /*所说的问题*/
      => NP(VP/ap  !NP/np)  %VP.FORM=VG  /*活着的诗人*/
        => NP(VP/ap !NP/np) %VP.FORM=VG /*计算与绘图的工作量*/
        => NP(VP/ap !NP/np) %VP.FORM=VG /*撒谎的孩子  游泳的经验*/
        => NP(VP/ap !NP/np) %VP.FORM=VN /*丢的那本书*/
```

注意，这条规则所覆盖的汉语句法结构实际上比上面的模板更广泛。因为这里的 ap 可以是任何形式的“的字结构(ap)[2]”，而不仅仅包含了“a＋的”的形式。因此，这条规则看上去虽然复杂，但是真正适合“a＋的＋n”结构的转换形式并不多。这里我们把规则中不适合这种结构的转换形式标为灰色。注意，有些没有被标为灰色的转换模式，虽然后面所附的例子并不是“a+的+n”结构，但这种转换形式是可以用于这种结构的。如果只考虑规则中 ap 为“a＋的”的情形，也就是说，不考虑规则中标记为灰色的转换形式，那么这条规则所给出的转换形式几乎已被上面的模板完全覆盖。实际上，上面的模板所给出的转换形式比这条规则给出的转换形式要丰富得多，例如，模板中出现了各种限定词被插入到英语短语中的情况，而人工撰写的规则却基本没有考虑这种情况。

由此我们可以看到，虽然我们所使用的双语语料库并不大，不过从中获得的短语结构转换模板所涵盖的两种语言结构的对应转换形式还是相当丰富的。通过对更多常见的汉语句法结构转换模板的分析，我们认为这个模板库已经覆盖了常见汉语句法结构到英语句法结构的

[1] 这条规则所使用的语法形式是该系统中专门定义的，由于这种形式比较复杂，在这里难以用简单的话解释清楚。要了解其具体含义，可以参见[詹卫东 2000]。这里只针对这个例子做一些简单的说明。例如：“NP(!NP/np C<who> CS/ap)”表示汉语 np 转换成英语 NP，汉语 ap 转换成英语子句 CS，而整个英语短语为一个 NP+who+CS 构成的 NP。“%NP.NCASE=COMM”表示这个英语结构中出现的第一个 NP 是普通格（不是所有格）。英语词性 T 是限定词，D 是副词。“%VP.FORM=VG”表示该 VP 的形式是现在分词。“%%NP.RSUBC=RPOSS”表示英语短语中第二个 NP 的中心词是一个名词性物主代词。

[2] 在汉语语言学的著作中，“的字结构”一般都认为是 np 而不是 ap。不过在机器翻译系统中，由于“的字结构”做主宾语的情况比较少见，一般都把“的字结构”处理为 ap，这样可以减少很多不必要的歧义。对于“的字结构”做主宾语的情况可采用单独的规则处理。

大部分转换形式。

6.7 小结

本章给出了一种“短语结构转换模板”的定义和自动抽取算法。

与其他形式的翻译模板相比，短语结构转换模板有很多突出的优点。具体来说表现在：

1. 短语结构转换模板具有很强的表达能力。这种模板不仅既可以反映词与词之间的转换关系，而且可以反映两种语言深层句法结构之间的对应关系。
2. 这种模板在刻画两种语言句法结构对应关系的时候，完全保持了两种语言句法结构树的原貌。通过短语结构转换模板，可以将一棵完整的源语言短语结构树转换成一棵完整的目标语言短语结构树，而不需要对两种语言的语法做任何修改。
3. 模板形式非常直观，这使得人类语言专家很容易理解这种模板，甚至直接编写或者修改模板（虽然这并不是我们的初衷）。
4. 可以从语料库中自动获取。这使得我们可以在大规模语料库的基础上，全自动地构造一个机器翻译系统，而无需花费大量精力去人工构造知识库。

我们看到，短语结构转换模板从形式上说，可以认为是一种规则，具有一般规则直观、表达能力强等优点，同时，这种模板又可以直接从双语语料库中直接获取，而不依赖于人工编写知识库。采用短语结构转换模板进行机器翻译，有可能实现规则方法和统计方法的有效结合，从而真正实现基于语言深层结构转换的统计机器翻译，使得机器翻译的水平在现有基于表层结构的统计机器翻译基础上有一个较大的提高。

本章中，我们给出了一种从短语对齐的语料库中抽取短语结构转换模板的算法。实验表明，我们从不太大（含 8009 个句子对）的语料库中抽取到大量的短语结构转换模板，其中词语模板和短语模板大约各占一半。其中，大部分模板具有较高的可信度，而且这种可信度可以通过模板的频度、长度等信息做出基本准确的判断。这些模板对于汉语句法结构和英语句法结构的对应关系具有较高的覆盖度，可以覆盖常见的汉语句法结构的大部分转换模式。

第7章　微引擎流水线机器翻译系统结构

本文第二章综述中介绍了常见的多引擎机器翻译技术，主要的两种多引擎系统结构是系统级多引擎结构和部件级多引擎结构。

在系统级多引擎机器翻译结构中，各个翻译引擎的颗粒度非常大，引擎之间的结合非常松散，一个翻译引擎无法引用另一个翻译引擎的中间结果，这限制了整个系统性能的提高，使得系统无法充分利用多种算法互相结合的优势。因此，采用这种方法的系统性能提高是比较有限的。不过，系统级多引擎机器翻译结构有一个突出的优点，就是其易扩充性（scalablity）。在这种结构下，各个翻译引擎的程序接口完全相同，添加和删除新的翻译引擎变得非常简单，这使得程序的扩充变得非常容易。

而在部件级多引擎机器翻译结构中，特别是在很多部件都采用多引擎方式的时候，由于各个引擎（部件）实现的功能不尽相同，引擎之间的通讯关系变得非常复杂，不同的引擎很难采用统一的程序接口，这使得程序的扩充变得非常困难。在 Verbmobil 系统中，为了解决翻译引擎之间通讯的问题，采用了多黑板结构，不过，由于黑板的设定是比较随意的，黑板数量很多，整个系统的复杂程度依然很高，模块的划分还不是非常清晰，系统的可扩充性也不是很好。

为了克服现有这两种多引擎机器翻译系统结构的缺点，我们定义并实现了一种特定的多引擎机器翻译系统结构，我们称之为“微引擎流水线”结构。这种结构本质上也是一种部件级的多引擎机器翻译系统结构。与一般的部件级多引擎结构不同之处在于，我们为每一个部件级的机器翻译引擎（我们称为微引擎）定义了统一的几种类型的接口，并给出了清晰的微引擎调度算法，即总体翻译算法。微引擎的增加、减少和修改都变得非常简单，并且一个微引擎的调整不会对其他微引擎的算法和总体翻译算法造成干扰，这样，系统的扩充变得非常容易。

7.1 微引擎流水线的基本思想

微引擎流水线的基本思想，就是希望在部件级多引擎结构上，实现系统模块接口的归一化，使得系统能够方便地进行增加、删除功能模块和调整功能模块的调用顺序，尝试各种不同算法的组合，以便寻找最优的机器翻译算法，提高系统的总体性能。

在微引擎流水线结构中，所有的功能模块被划分成四种类型：识别器（Recognizer）、选择器（Selector）、转换器（Transferror）和生成器（Generator），统称为微引擎（Micro-Engine）。每一种类型的微引擎都具有完全相同的接口。

整个系统共享一个公共的数据结构：线图。

识别器的作用是从源语言句子中识别出短语，选择器的作用是排解歧义，删除错误的短语结点。每个识别器可以采用比较单一的算法以识别相应的短语，例如可以用一个有限状态机识别器来识别数词、时间词、网址等结构固定的格式；用规则识别器可以识别一般的短语；用复句识别器根据关联词的搭配来识别复句等等。选择器用于排除识别器产生的歧义，例如可以设计专门的选择器用于排除汉语的 V＋N＋N 歧义。

多个识别器和选择器排列成一个分析流水线，系统依次调用每个识别器和选择器，往线图中添加或删除结点，直到完成输入句子的分析。识别器和选择器共同完成对源文句子的分析。

转换器的作用是将一个识别器识别出来的短语转换成目标语言短语。转换器不构成流水线形式，转换器的选择由产生该源语言结点的识别器来指定。每一种转换器只针对特定类型的短语进行转换，例如汉语的人名翻译可以采用拼音转换器，数词翻译需要特定的数词转换器，而这些转换器分别采用不同的算法。由于短语是嵌套的，所以一个转换器可以调用其他转换器以实现其构成成分（子结点）的转换。

转换时，对源文结构树上的每个结点，调用相应的转换器，将源文结构树转换成译文结构树。

生成器的作用是对目标语言句子结构进行调整，使之符合目标语言的语法结构。每一个生成器都对整个目标语言结构树进行操作。每一个生成器实现较为单一的操作，比如说一个生成器专门处理动词的时态，另一个生成器专门处理助动词的位移，还有一个生成器专门处理目标语言单词的形态变化。多个生成器构成一条生产流水线，系统依次调用生产流水线上的每个生成器，以完成对译文结构树的调整，使之复合目标语言的语法。

7.2 微引擎流水线的系统结构

整个微引擎流水线结构如下图所示：

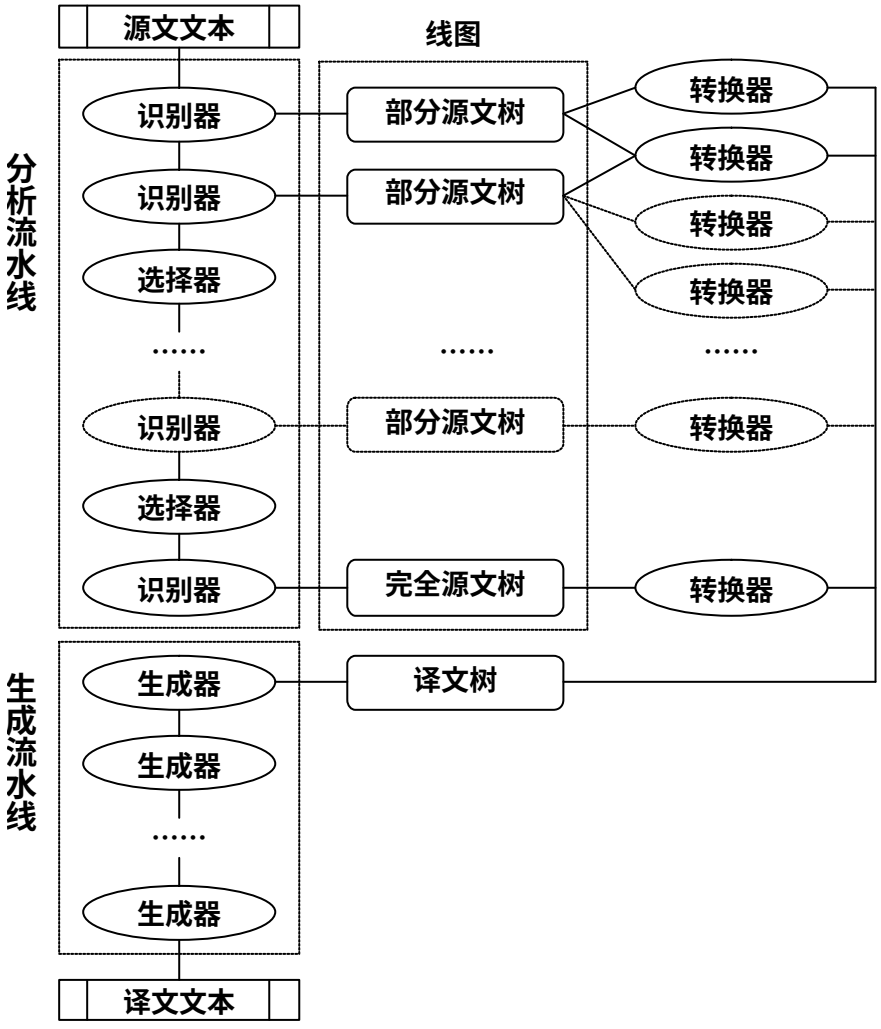


图 7.1 微引擎流水线机器翻译系统结构图

一个微引擎流水线的程序模块结构由以下一个七元组构成：

{ R，S，T，G，Σ，Ω，Ψ}

其中：

R：是一个“识别器（Recognizer）”的集合；

S：是一个“选择器（Selector）”的集合；

T：是一个“转换器（Transferror）”的集合；

G：是一个“生成器（Generator）”的集合；

Σ：是一个向量：$\sigma_1\sigma_2\cdots\cdots\sigma_n$，其中$\sigma_i \in R \cup S$，$1 \le i \le n$；

Ω：是一个向量：$\omega_1\omega_2\cdots\cdots\omega_m$；其中$\omega_j \in G$，$1 \le j \le m$；

Ψ：是一个 R? 2^T的映射，即对任意 $r \in R$，存在唯一的 $t \subset T$（T 的子集），使得 $t = \psi(r)$；

Σ是一个由识别器和选择器组成的流水线，称为分析流水线；

Ω是一个由生成器组成的流水线，称为生成流水线；

Ψ是一个识别器到转换器的映射关系表，对于任何一个给定的识别器，对应的有一组转换器将该识别器识别出来的源语言结点转换成目标语言结点。

7.3 微引擎流水线的公共数据结构

微引擎流水线的公共数据结构总共包括两类：一类是线图（Chart）结构，一类是句法树结构。其中句法树结构又分为源文句法树和译文句法树。

线图结构如下图所示：

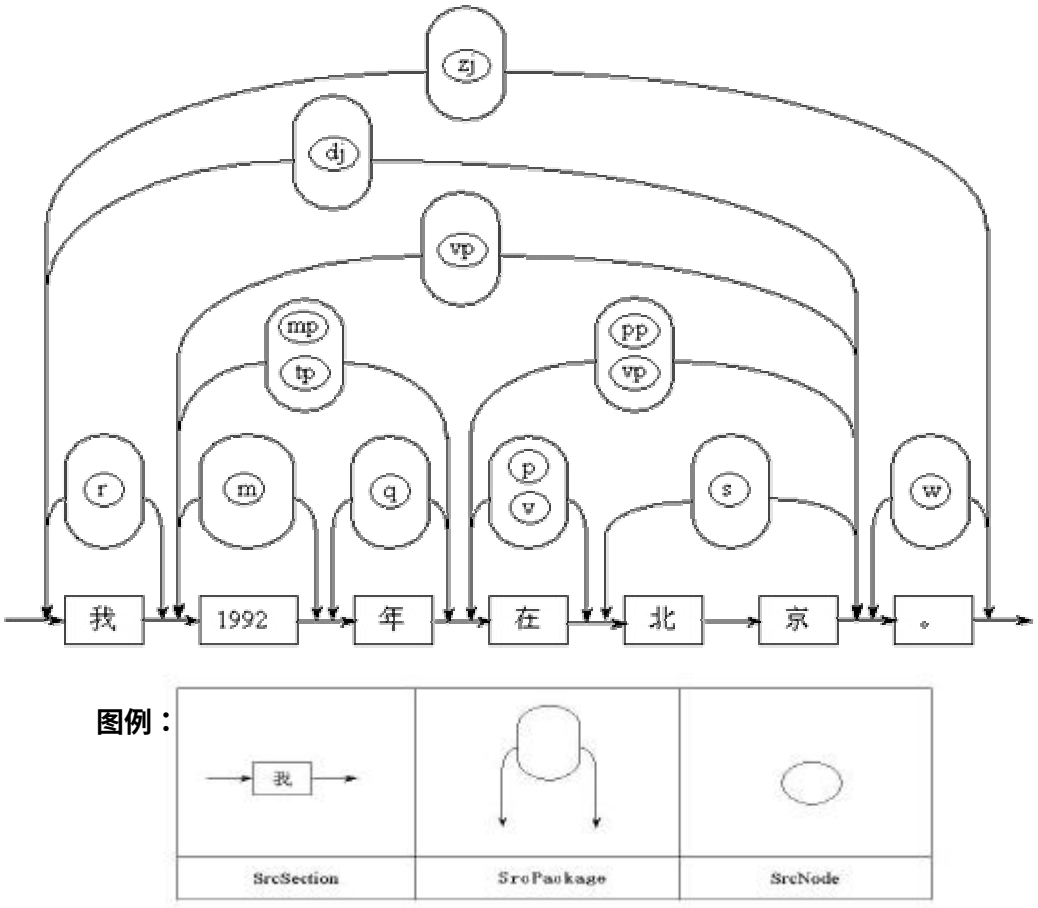


图 7.2　微引擎流水线的公共数据结构：线图

对应的数据结构定义如下：

表 7.1 线图中所使用的数据说明

SrcSection	线图结构中相邻两个词语结点之间的一段文本， 是句子中可能构成词的最小单位， 一般是一个汉字，或一串数字，或一个外文单词。
SrcSectionTable	是一个由 SrcSection 组成的数组。
SrcNode	线图中的弧，或句法树中的结点， 可以是一个词，也可以是一个短语或句子； 每一个 SrcNode 由首尾两个 SrcSection 确定其位置。
SrcPackage	位置相同的所有 SrcNode 构成一个结点包， 存放在一个 SrcPackage 的数据结构中。
SrcPackageTable	是一个由 SrcPackage 组成的数组，通过函数： getPackage(int first,int last) 可以存取任何指定位置的 SrcPackage， 参数 first 和 last 用于指定首尾 SrcSection 序号。
SrcChart	由一个 SrcSectionTable 和一个 SrcPackageTable 组成。

同时，各 SrcNode 通过子结点关系构成源文句法树，如下图所示：

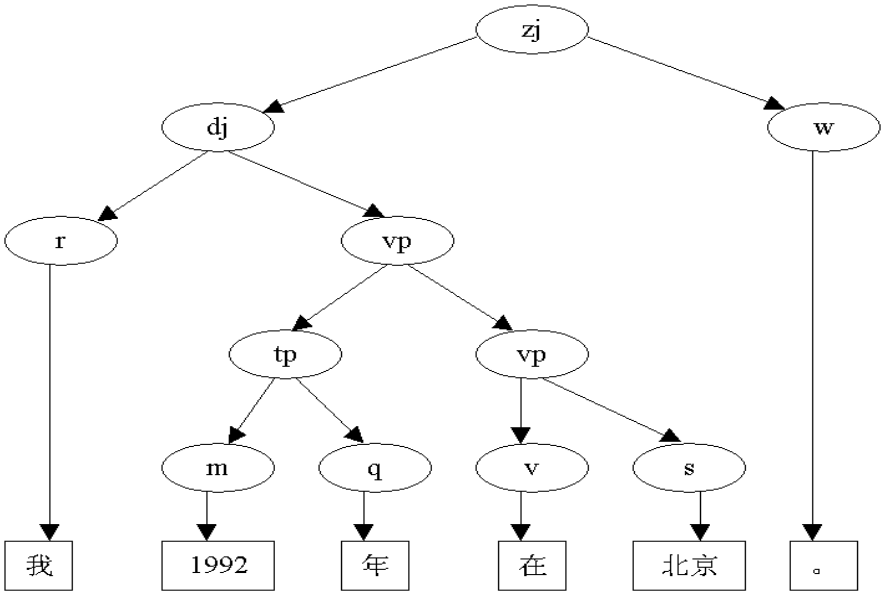


图 7.3 与线图对应的句法树结构

源文句法树经过转换生成就得到了译文句法树。其形式与源文句法树类似，这里不再给出图例说明。

7.4 各种微引擎的程序接口和功能说明

1. 识别器（Recognizer）

识别器需要实现两个函数：

函数：初始化（Initialize）

输入：Chart 结构（SrcChart）

输出：无

说明：为识别操作准备初始数据，每个识别器只需执行一次。

函数：识别（Recognize）

输入：Chart 结构（SrcChart）

输出：一个源文结点（SrcNode）和一个转换器（Transferror）指针

说明：从 Chart 结构中识别出一个结点，并指明该结点应该由哪个转换器进行转换。此操作被反复执行。

2. 选择器（Selector）

选择器需要实现两个函数：

函数：初始化（Initialize）

输入：Chart 结构（SrcChart）

输出：无

说明：为选择操作准备初始数据，每个选择器只需执行一次；

函数：选择（Select）

输入：Chart 结构（SrcChart）

输出：一个源文结点表（list<SrcNodes>）

说明：从 Chart 结构中选择一些结点放入输出的源文结点表中，凡是不在该表中的结点将在后续的操作中不再有效（被剪枝）。要注意的是，选择器并不要求输出唯一的结果，暂时无法解决的歧义结点完全可以都保留下来，留给以后处理；

3. 转换器（Transferror）

转换器需要实现两个函数：

函数：初始化（Initialize）

输入：一个源文结点（SrcNode）

输出：无

说明：为转换操作准备数据。由于转换所需的与结点有关的数据都存放在 SrcNode（或其派生类）中，因此此操作需对每个 SrcNode 执行一次。

函数：转换（Transfer）

输入：一个源文结点（SrcNode）

输出：一个译文结点（TgtNode）

说明：对以输入的源文结点为根结点的源文子树进行转换，得到一个译文子树，并输出译文子树的根结点。此函数可通过递归调用，实现对其子孙结点的转换。

4. 生成器（Generator）

生成器需要实现两个函数：

函数：初始化（Initialize）

输入：一个译文结点（TgtNode）

输出：无

说明：为生成操作准备数据。此操作需对每个 TgtNode 执行一次。

函数：生成（Generate）

输入：一个译文结点（TgtNode）

输出：另一个译文结点（TgtNode）

说明：对以输入的译文结点为根结点的译文子树进行某种特定类型生成操作，并输出所得的新译文子树的根结点。

7.5 微引擎调度算法

对于微引擎流水线结构的机器翻译系统来说，作为总体翻译算法的微引擎调度算法是固定的，不需要修改。对于翻译系统的调整主要体现在微引擎的实现算法和流水线的安排上。

整个微引擎调度翻译算法分为分析、转换、生成三个步骤。

1. **分析算法**

```
BEGIN
    REPEAT 依次从分析流水线中取一个微引擎
    WHILE 该微引擎不为空
        IF 该微引擎是识别器
        THEN
            调用该识别器的初始化函数
            REPEAT 调用该识别器的识别函数
                IF 识别出的结点覆盖整个输入文本
                THEN 返回成功，将该结点置为源文根结点
                ELSE 将识别出的结点加入的 Chart 结构中
                ENDIF
            ENDREPEAT
        ELSE
            调用该选择器的初始化函数
            调用该选择器的选择函数
            根据返回的结点表重新构造 Chart 结构，
                删除不在表中的结点
        ENDIF
    ENDREPEAT
    返回失败
END
```

2. **转换算法**

```
BEGIN
    取源文根结点的识别器
    取该结点对应的转换器
    调用该转换器的初始化函数，对源文结构树根结点进行转换初始化
    调用该转换器的转换函数，对源文结构树进行转换，转换过程中递归调用
    子结点对应的转换器的转换函数，进行子结点的转换；
    返回得到的译文结点，作为译文结构树根结点
END
```

3. **生成算法**

```
BEGIN
    REPEAT  依次从生成流水线中取一个生成器
    WHILE  该生成器不为空
        调用该生成器的初始化函数，对译文结构树根结点进行生成初始化
        调用该生成器的生成函数，对译文结构树进行生成
        将返回的译文结点作为新的译文结构树根结点
    ENDREPEAT
    返回译文根结点
END
```

可以看到，在分析过程中，各个识别器依次对输入文本进行处理，由于各个识别器采用的算法不同，各种算法取长补短，尽可能得到一个较好的分析结果。而选择器用于对过多的识别结果进行剪枝排歧，以减少搜索的空间。

在转换算法中，采用的方法是自顶向下的一遍扫描。不同的转换器用于对不同类型的识别器产生的结点进行转换。

在生成算法中，采用的方法是自顶向下的多遍扫描。每一个生成器都要对整个译文结构树进行一次扫描。

7.6 面向新闻领域的汉英机器翻译系统

7.6.1 研究背景

基于已有的机器翻译技术和上述的微引擎流水线结构，我们实现了一个“面向新闻领域的汉英机器翻译系统（NOCEMT）”，这里对这个系统的情况做一个简单的介绍。

这个系统的前身是一个纯规则的基于合一语法的汉英机器翻译系统[刘群，1997]，其基本情况如下：

1. 词典规模：含有丰富句法语义和英语对译信息的核心词典约 5 万词；
2. 规则库规模：全局规则和局部规则总数约 1000 条；
3. 语法规则形式：自定义的类似于词汇功能语法（LFG）的“上下文无关规则骨架＋合一约束”的形式；
4. 合一约束：基于自定义的特征网络表示形式的合一运算；
5. 训练情况：使用了 3000 多个典型句子进行人工调试；
6. 测试情况 在 1998 年 863 专家组组织的一次评测中译文质量分数为 70 分(百分制)，测试文本主要为一些简短的规范汉语句子。

“面向新闻领域的汉英机器翻译系统”在原先系统的基础上进行了较大规模的改动，主要的改动包括以下方面：

1. 采用我们提出的微引擎流水线机器翻译系统：在这个结构下，原先的机器翻译系统被拆分成若干个微引擎，并和新增加的微引擎结合成了一个整体；
2. 合一运算和规则执行机制的改变：原来规则执行方式类似于程序设计语言的解释执行方式，规则执行时直接对字符串进行匹配。在新的系统中，规则执行方式类似于程序设计语言的编译执行方式，规则事先经过处理变成一种内部数据结构，特别是所有的字符串都变成了一个编号，字符串的比较变成了整数的比较，系统运行的整体性能比原来有了较大的提高（3～4 倍）；

3. 新的汉语词法分析器：利用本文前面介绍的基于层叠隐马尔可夫模型的汉语词法分析器，汉语词法分析的正确率有了大幅度提高，特别是较好地解决了汉语未定义词识别的问题；
4. 命名实体的翻译：对一些特定的命名实体类型，如人名、地面、数词等等，实现了专门的翻译程序，效果较好；
5. 扩充词典：补充我们本文前面介绍的机器翻译扩充词典，词典规模从原来的 5 万词左右上升到 20 多万词，另外我们还补充了一个小规模的人工构造的短语词典，包含大约 5 千个汉英对照短语，较大程度上解决词典覆盖率问题。

可以看到，在上面这些工作中，我们没有对原有的规则系统进行进一步的补充和调试。这一方面是因为我们没有足够的人力资源投入，另一方面也是因为我们看到规则库的调试要耗费太多的人力物力，而这种大规模而繁琐枯燥的工作在高校和研究机构是比较难以开展的。因此我们希望能够将基于统计的方法用于机器翻译知识的自动获取，这也是本文开展短语结构对齐、短语结构转换模板自动提取和基于模板的统计翻译模型研究的一个主要原因。不过由于时间和进度的关系，到目前为止，上述的研究成果还没有反映到目前的系统中来。下面我们仅就目前已经完成的情况介绍一下系统的实现方案和测试结果。

7.6.2 系统实现方案

现有的“面向新闻领域的机器翻译系统”完全采用了前面所说的微引擎流水线机器翻译系统结构。其具体实现方案如下：

分析流水线依次由以下识别器构成：

1. 词法分析识别器
   一个融汉语词语切分、未定义词识别、词性标注为一体的词法分析模块，采用本文前面介绍的基于层叠隐马尔可夫模型的汉语词法分析算法；
2. 扩充词典识别器
   采用[刘群 2002]介绍的一种人机互助方法，利用数十部供人使用的词典（或词表），我们构造了一部的大规模的机器翻译扩充词典，规模为汉语词条 21 万，汉英对照词条 41 万。对于词法分析识别器没有识别出的词或短语，可以利用这部词典作为补充；
3. 短语库识别器
   利用一个人工构造短语库识别句子中的短语，含短语 5000 余条；
4. 基于规则的识别器
   从原有的机器翻译系统中拆分出来的基于合一语法规则的句法分析引擎；
5. 软失败识别器
   在无法将输入句子分析得到一个完整的结点时，使用软失败识别器将已有的结点尽可能组合成一个覆盖整个句子的结点。

转换器包括以下几种：

1. 核心词典转换器
   根据核心词典进行转换；
2. 扩充词典转换器
   根据扩充词典进行转换；
3. 中国人名转换器
   中国人名的翻译；
4. 中国地名转换器

中国地名的翻译；

5. 译名转换器
   译名的翻译，将外来词的汉语音译名转换成符合英语构词习惯的英语词，此模块由清华大学的合作研究小组开发；
6. 数词转换器
   一个完整的汉语数词的翻译模块，将汉语数词翻译成阿拉伯数字，由本课题组其他成员开发；
7. 短语库转换器
   根据短语库进行翻译；
8. 基于规则的转换器
   从原有汉英机器翻译系统中拆分出来的基于规则的句法结构转换引擎，采用基于合一的规则形式，可以将一棵汉语句法树转换成英语句法树；
9. 软失败转换器
   对软失败识别器产生的结点进行转换。

生成流水线依次由以下生成器构成：

1. 基于规则的结构生成器
   从原有的汉英机器翻译系统中拆分出来的基于规则的英语结构生成引擎，采用基于合一的语规则形式进行英语句法结构的调整，主要进行英语局部词序的调整和根据动词时态和语态添加助动词，以及进行英语助动词的位移；
2. 基于规则的词语生成器
   从原有的汉英机器翻译系统中拆分出来的基于规则的英语词语生成引擎，生成英语词语的变形，包括动词的分词形式和动名词形式、名词复数形式、形容词比较级和最高级等等。

可以看到，目前的微引擎流水线结构还是比较简单的，主要是在已有的基于规则的机器翻译系统基础上扩充整理而成，还没有把基于统计的句法分析和翻译模块加进来，也还没有用到选择器。

7.7 实验结果及分析

下表是我们利用 NIST 的测试程序[Papineni 2001][NIST 2002]。NIST 测试程序是一种基于 n 元语法的机器翻译自动评测程序，这种评测程序将系统得到的译文与四个人工翻译的参考译文进行匹配，计算匹配率。引入对于匹配的片段，引入信息量来衡量该片段的重要程度。NIST 测试结果的评分值跟所提供的测试语料库的规模有一定关系，因此其测试结果只具有相对比较意义，不具有绝对意义。

我们的评测采用 NIST 提供的测试语料，主要内容为新华社新闻稿，含 800 多个句子。下面给出了在各种不同的微引擎组合情况下，对机器翻译系统产生的译文的进行测试所得到的结果：

表 7.2 各种不同微引擎的组合实验结果

对词法分析识别器使用扩充词典转换器	使用短语识别器和短语转换器	使用扩充词典识别器和扩充词典转换器	使用基于规则的识别器和基于规则的转换器	运行时间（毫秒）	NIST 评分
	√	√	√	3293172	5.8697
√	√	√		103609	5.4669

√	√		√	962594	5.7706
√	√			97250	5.2793
√		√	√	3386672	5.8351
√		√		43110	5.4073
√			√	666516	5.6695
√				27515	5.1449
	√	√	√	3397828	5.4332
	√	√		96797	5.0443
	√		√	906187	5.3243
	√			97937	4.8343
		√	√	3224968	5.3866
		√		52516	4.9869
			√	658859	5.2121
				36172	4.6993

上表中，第一列如果没有选中（对应单元格为空），表示只将切分标注产生的词送到核心词典转换器进行转换，而不送到扩充词典转换器，如果选中（对应单元格为 v），表示将切分标注产生的词,在通过核心词典转换器进行转换失败时,送到扩充词典转换器进行转换；第二列表示是否使用短语库识别器和短语库转换器 ;第三列表示是否使用扩充词典识别器和扩充词典转换器；第四列表示是否使用基于规则的识别器和基于规则的转换器。最后两列分别是翻译时间（单位为毫秒）和 NIST 评分。

从上面的结果可以看到，仅使用一部核心词典和汉语切分标注程序，结果评分即可达到 4.6993。仅加入规则引擎，结果评分可达到 5.2121，仅加入扩充词典和短语库，不使用规则引擎，结果评分可达到 5.4669，加入全部微引擎，结果评分可达到 5.8697。这个结果告诉我们，通过扩充词典导致的翻译效果改善甚至会大于加入翻译规则导致的翻译效果改善，这一方面说明词典的重要作用，另一方面也说明目前的系统中深层翻译知识还比较缺乏，更说明通过大规模双语语料库获取翻译知识的迫切性。而综合使用所有引擎,翻译效果有很大提高，说明多引擎方法还是非常有效的。

通过这个实验可以看到，在微引擎流水线的机器翻译结构下，可以方便地实现翻译微引擎的自由组合，通过对不同组合情况下的结果进行评分，很容易了解各个微引擎在系统中所起到的作用，这特别有利于对各种翻译算法进行取舍和调整。

7.8 小结

微引擎流水线的机器翻译体系结构具有以下优点：

1. 在机器翻译的每个阶段，都可以采用多个算法不同的微引擎，以达到取长补短的效果；
2. 句法分析阶段，将分析器和选择器的功能分开；分析器只关注可能性，无需考虑与其他结点的冲突；选择器专门处理结点间的冲突，这种分工有助于在系统设计中对排歧问题进行更全面考虑，更有益于歧义冲突的解决；
3. 每个微引擎的编程接口非常清晰，微引擎设计者在了解自己的任务分配的情况下，可以着重关注于算法本身，而无需考虑与其他模块的交互，这样就把复杂的机器翻译问题分解成了一系列可以独立处理的小问题，化繁为简，有助于对机器翻译

系统进行团队式开发，有利于探索机器翻译中的新算法和新思路；

4. **采用面向对象的编程方法，开发一个新的微引擎只需在已有的微引擎基类的基础上派生出新的子类，可靠性高，易于实现；**
5. **整个的机器翻译算法是固定的，任何时候都无需改变；通过设计新的微引擎和调整微引擎流水线的结构，可以实现对翻译系统功能的任意裁减，以产生不同的输出结果（如产生切分标注结果的输出、句法分析结果的输出等）。**

第8章 总结及今后的工作

机器翻译是困难的，而汉英机器翻译尤为困难。虽然全自动高质量的机器翻译也许是一个在现有计算条件下很难达到的目标，但作者还是相信，现有的机器翻译方法还远远没有充分发挥现代计算机系统所提供的强大的计算能力。作者认为，通过充分利用人类专家知识库、在大规模语料库基础上高效获取语言翻译知识、建立合适的反映语言深层结构对应关系的模型，可以把现有的机器翻译水平向前推进一大步。本文就是作者在这种指导思想下，对汉英机器翻译中的一些关键技术所开展的研究及取得的成果。

在汉语词法分析方面，本文提出了一种基于层叠隐马尔可夫模型的汉语词法分析一体化算法，对于汉语词法分析所面临的主要问题，包括词语切分、未定义词识别、词性标注等，都提出了相应解决方法，并将这些方法有效地结合成了一个完整的算法框架，使汉语词法分析的总体性能达到了一个较高的水平，为汉英机器翻译奠定了一个很好的基础。

为了解决基于实例的机器翻译中相似例句匹配等问题，本文提出了一种基于《知网》的词汇语义相似度计算方法，这种方法充分利用了《知网》里所蕴涵的丰富的人类专家知识，无需通过大规模的语料库训练，就可以准确细致地刻画出词语之间的语义相似度，取得了较好的效果。

双语语料库的短语结构对齐是获取两种语言深层结构对应关系的前提。本文对此问题提出了一种双语短语结构对齐的柱形搜索算法。这种方法尽可能地避免了词语对齐错误对短语对齐所带来的困扰，使得短语对齐的正确率和召回率比词语对齐的相应指标有较大的提高。同时这个算法的时间复杂度是线性的，效率很高。

我们定义了一种短语结构转换模板，并在双语短语结构对齐语料库的基础上，实现了这种模板的自动抽取算法。这种模板的优点在于一方面比较直观，可以反映两种语言深层句法结构之间的对应关系，另一方面，这种模板又可以从大规模语料库中直接抽取。实验表明，我们给出的模板抽取算法可以从不大的语料库中抽取到大量的短语结构转换模板，这些模板对于汉语句法结构和英语句法结构的对应关系具有较高的覆盖度，可以覆盖常见的汉语句法结构的大部分转换模式。模板中虽然含有一定数量的错误，不过模板的可信度可以通过模板的频度、长度等信息做出大致的判断。

为了实现多种翻译算法在一个系统中的有效融合，我们定义了一种微引擎流水线机器翻译系统结构。在这种结构下，各种翻译模块被定义为微引擎。不同类型的具有统一的接口，微引擎之间在系统中组织成流水线的形式。通过微引擎的增删和流水线结构的调整可以快速地实现多种算法的各种组合方式，有利于开发者通过频繁测试选择最优方案，而 无需修改整个系统的算法。

在下一步的工作中，我们将主要在短语结构转换模板的基础上，建立反映两种语言深层结构对应关系的统计翻译模型，并在此基础上实现一个基于统计的汉英机器翻译系统。

附录：汉语词性标记集 ICTPOS

0. 说明

汉语词性标记集 ICTPOS 主要用于汉语词法分析器、句法分析器和汉英机器翻译系统。本标记集主要参考了以下词性标记集：

0. 北大《人民日报》语料库词性标记集；
1. 北大 2003 版词性标记集；
2. 清华大学汉语树库词性标记集；
3. 教育部语用所词性标记集（国家推荐标准草案 2002 版）；
4. 美国宾州大学中文树库（ChinesePennTreeBank）词性标记集；

本词性标记集主要以北大《人民日报》语料库的词性标记集为蓝本，并参考了北大《汉语语法信息词典》中给出的汉语词的语法信息。

本标记集在制定过程中主要考虑了以下几方面的因素：

1. 有助于提高汉语词法分析器的切分和标注正确率；
2. 有助于提高汉语句法分析器的正确率；
3. 有助于汉英机器翻译系统进行翻译；
4. 易于从北大《人民日报》语料库词性标记集进行转换；
5. 对于语法功能不同的词，在不造成词法分析和句法分析歧义区分困难的情况下，尽可能细分子类。

基于以上考虑，我们在标注过程中尽量避免那些容易出错的词性标记，而采用那些不容易出错、而对提高汉语词法句法分析正确率有明显作用的标记。例如，在动词的子类中，我们参考了宾州大学中文树库的做法，把汉语动词“是”和“有”分别做成单独的标记，而没有采用“系动词”的标记。因为同样是“是”这个动词，其句法功能很多，作“系动词”只是其中一种功能，而要区分这些功能是非常困难的，会导致词法分析的正确率下降。

在名词子类中，我们区分了“汉语人名”、“日语人名”和“翻译人名”，这不仅仅是因为这三种人名要采用不同的参数进行训练与识别，而且在汉英机器翻译中也要采用不同的分析算法进行翻译。又如，我们把表示时间的“数词＋‘年’”（如“1995 年”）合并成一个时间词，而表示年头的“数词＋‘年’”分别标注为“数词”和“量词”，这是因为我们通过实验发现这种区分在词法分析阶段通过统计方法可以达到较高的正确率，而且这种区分对于后续的句法分析和机器翻译有非常重要的作用。

对于某些词类（助词和标点符号），基本上是一个封闭集，而这些词类中各个词的语法功能相差很大，在这种情况下，我们尽可能地细分其子类。

另外，与其他词性标记集类似，在我们的标记体系中，小类只是大类中一些有必要区分的一些特例，但小类的划分不满足完备性。

1. 名词

名词分为以下子类：

n 名词
  nr 人名
    nr1 汉语姓氏
    nr2 汉语名字

- nrj **日语人名**
- nrf **音译人名**
- ns **地名**
- nsf **音译地名**
- nt **机构团体名**
- nz **其它专名**
- nl **名词性惯用语**
- ng **名词性语素**

2. **时间词**
   - t **时间词**
   - tg **时间词性语素**

3. 处所词

- s **处所词**

4. 方位词

- f **方位词**

5. 动词

- v **动词**
  - vd **副动词**
  - vn **名动词**
  - vshi **动词“是”**
  - vyou **动词“有”**
  - vf **趋向动词**
  - vx **形式动词**
  - vi **不及物动词（内动词）**
  - vl **动词性惯用语**
  - vg **动词性语素**

6. 形容词

- a **形容词**
  - ad **副形词**
  - an **名形词**
  - ag **形容词性语素**
  - al **形容词性惯用语**

7. 区别词

- b **区别词**
  - bg **区别词性语素**
  - bl **区别词性惯用语**

8. 状态词

- z **状态词**

9. 代词

- r 代词
  - rr 人称代词
  - rz 指示代词
    - rzt 时间指示代词
    - rzs 处所指示代词
    - rzv 谓词性指示代词
  - ry 疑问代词
    - ryt 时间疑问代词
    - rys 处所疑问代词
    - ryv 谓词性疑问代词
  - rg 代词性语素

10.数词

- m 数词
  - mq 数量词

11.量词

- q 量词
  - qv 动量词
  - qt 时量词

12.副词

- d 副词

13.介词

- p 介词
  - pba 介词“把”
  - pbei 介词“被”

14.连词

- c 连词
  - cc 并列连词

15.助词

- u 助词
  - uzhe 着
  - ule 了 喽
  - uguo 过
  - ude1 的 底
  - ude2 地
  - ude3 得
  - usuo 所

udeng **等 等等 云云**
uyy **一样 一般 似的 般**
udh **的话**
uls **来讲 来说 而言 说来**
ujl **极了**
uzhi **之**
ulian **连 （“连小学生都会”）**
uqj **起见**

16.叹词

e **叹词**

17.语气词

y **语气词**

18.拟声词

o **拟声词**

19.前缀

h **前缀**

20.后缀

k **后缀**

21.字符串

x **字符串**
xx **非语素字**
xu **网址** URL

22.标点符号

w **标点符号**
wkz **左括号，全角：（ 〔 ［ ｛ 《 【 〖 〈 半角：**([{ <
wky **右括号，全角：）〕 ］ ｝ 》 】 〗 〉 半角：**)] { >
wyb **半角引号，半角：**"'
wyz **左引号，全角：“ ‘ 『**
wyy **右引号，全角：” ’ 』**
wj **句号，全角：。**
ww **问号，全角：？ 半角：**?
wt **叹号，全角：！ 半角：**!
wd **逗号，全角：， 半角：**,
wf **分号，全角：； 半角：** ;
wn **顿号，全角：、**
wm **冒号，全角：： 半角：** :
ws **省略号，全角：…… …**

wp **破折号，全角：——　　——　　———　　半角：**---　----
wb **百分号千分号，全角：％ ‰　　半角：**%
wh **单位符号，全角：￥　＄　£　　°　℃　半角：**$

参考文献

[Al-Onaizan 1999] Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith and David Yarowsky (1999). Statistical Machine Translation: Final Report, Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Speech and Language Processing, Baltimore, MD.

[Alshawi 1998] Alshawi, H., Bangalore, S. and Douglas, S., Automatic Acquisition of Hierarchical transduction models for machine translation, Proc. 36th Conf. Association of Computational Linguistics, Montreal, Canada, 1998.

[Agirre 1995] Agirre E. and Rigau G., A proposal for word sense disambiguation using conceptual distance, in International Conference "Recent Advances in Natural Language Processing" RANLP'95, Tzigov Chark, Bulgaria,.

[Berger 1994] Berger, A., P. Brown, S. Della Pietra, V. Della Pietra, J. Gillett, J. Lafferty, R. Mercer, H. Printz, L Ures, The Candide System for Machine Translation, Proceedings of the DARPA Workshop on Human Language Technology (HLT)

[Berger 1996] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, A maximum entropy approach to natural language processing, Computational Linguistics, 22(1):39-72,March 1996.

[Brill 1995] Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging, Computational Linguistics, Dec. 1995

[Brown 1990] Peter F. Brown, John Cocke, Stephen A Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin, A Statistical Approach to Machine Translation, Computational Linguistics,1990

[Brown 1991] Brown, P. F., Lai, J. C., and Mercer, R. L., Aligning Sentences in Parallel Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91), pages 169-176, Berkeley, CA, 1991.

[Brown 1993] Peter. F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, Vol 19, No.2, 1993

[Brown 1995] Ralf Brown and Robert Frederking 1995. Applying Statistical English Language Modeling to Symbolic Machine Translation. In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95), pages 221-239, Leuven, Belgium.

[Brown 1996] Ralf D. Brown, "Example-Based Machine Translation in the Pangloss System". In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), p. 169-174. Copenhagen, Denmark, August 5-9, 1996.

[Brown 2000] Ralf D. Brown. "Automated Generalization of Translation Examples". In Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000), p. 125-131. Saarbrücken, Germany, August 2000.

[Charniak 2000] Charniak E., A maximum-entropy- inspired parser. In Proceedings of the 2000

Conference of the North American Chapter of the Assocation for Computational Linguistics. ACL, pp.132-139, New Brunswick N J, 2000.

[Chen 2002] Keh-Jiann Chen and Wei-Yun Ma. 2002. Unknown Word Extraction for Chinese Documents. In COLING-2002: The 19th International Conference on Computational Linguistics Vol. 1, pages 169–175.

[Church 1990] Church K.W. and Hanks, Word association norms, mutual information and lexicography. In: Computational Linguistics 16(1):22~29, 1990

[Dagan 1995] Dagan I., Marcus S., et al., Contextual Word Similarity and Estimation from Sparse Data, ACL95

[Dagan 1999] Dagan I., Lee L. and Pereira F., Similarity-based models of word cooccurrence probabilities, Machine Learning, Special issue on Machine Learning and Natural Language

[Frederking 1994] Robert Frederking, Sergei Nirenburg, Three Heads are Better than One, Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP-94), pages 95-100,Stuttgart, Germany

[Fung 1995] Fung P., Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In: proceedings of the 33th Annual Meeting of tha Association for Computational Linguistics, Boston, USA., 1995

[Furuse & Iida 1992] Furuse, Osamu & Hitoshi Iida: 1992, `Cooperation between transfer and analysis in example-based framework', in Proceedings of the Fifteenth [sic] International Conference on Computational Linguistics (COLING-92), Nantes, France, pp. 645-651.

[Gale 1993] Gale, W. A., and Church, K. W., A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(2): 75-102, 1993.

[Gauch 1995] Gauch S. and Chong M. K., Automatic Word Similarity Detection for TREC 4 Query Expansion, Proc. of TREC-4: The 4th Annual Text REtrieval Conf., Nov. 1995, Gaithersburg, MD, pp. 527-536.

[Grishman 1994] Grishman, R., Iterative Alignment of Syntactic Structures for a Bilingual Corpus. Proc. Of 2nd Workshop for Very Large Corpora (WVLC-94), pp.57-68, 1994

[Güvenir 1998] Halil Altay Güvenir, Ilyas Cicekli, Learning Translation Templates from Examples, Information Systems, Information Systems, Vol. 23 (6), pp. 353-363 , 1998

[Haruno 1996] Haruno M., Ikehara S. and Yamazaki T., Learning bilingual collocations by word-level sorting. In: COLING96 (pp. 525~530) , 1996

[Hatzivassiloglou 1995] Vasileios Hatzivassiloglou, Kevin Knight, Unification-Based Glossing, Proc. 14th Int. Joint Conf. Artificial Intelligence, 1995

[Hockenmaier 1998] Hockenmaier, J. and Brew, C., Error-driven learning of Chinese word VHJ P HQWDWLRQ Q- * XR. 7 / XDQG- ; XHGLWRUVWK3 DFLILF &RQIHUHQFHRQ Language and Information, pp. 218-229, Singapore. Chinese and Oriental Languages Processing Society, 1998

[Hogan 1998] Christopher Hogan, Robert E. Frederking, An Evaluation of Multi-engine MT Architecture, in: Machine Translation and the Information Soup, pages 113-123, Third Conference of the Association for Machine Translation in Americas (AMTA' 98), Langhorne, PA. USA, October

[Hutchins 1994] John Hutchins, Research methods and system designs in machine translation: a ten-year review, 1984-1994, International conference 'Machine translation: ten years on', Cranfield University, England, 12-14 November 1994. Proceedings edited by Douglas

Clarke and Alfred Vella. Cranfield: Cranfield University Press, 1998. Paper no.4

[Imamura 2001] Kenji Imamura, Hierarchical phrase alignment harmonized with parsing, in Proc. of NLPRS 2001, Tokyo. 2001

[Kaji 1992] Kaji, H., Kida, Y., and Morimoto, Y. , Learning Translation Templates from Bilingual Texts. COLING-92, pp. 672-678.

[Ker 1997] Sue J. Ker, Jason S. Chang, A Class-based Approach to Word Alignment, Computational Linguistics, Vol. 23, No. 2, Page 313-343, 1997

[Kishore 2001] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Technical Report, (keyword = RC22176), 2001, http://domino.watson.ibm.com/library/CyberDig.nsf/home.

[Kitano 1993] Kitano, Hiroaki: 1993, `A comprehensive and practical model of memory-based Machine Translation', in Proceedings of the 13th International Joint Conference on Artificial Intelligence, Chambery, France, pp. 1276-1282.

[Knight 1997] Kevin Knight, Automating knowledge acquisition for machine translation. AI Magazine 18(4).

[Knight 1998] Kevin Knight and Yaser Al-Onaizan, Translation with finite-state devices. In Proc. AMTA-98.

[Knight 1999] Kevin Knight, A Statistical Machine Translation Tutorial Workbook. unpublished, prepared in connection with the JHU summer workshop, August 1999. (available at http://www.clsp.jhu.edu/ws99/projects/mt/wkbk.rtf).

[Li 1995] LI Xiaobin, Szpakowicz S., and Matwin S., A WordNet-based algorithm for word sense disambiguation, in Proceedings of the Twelth International Joint Conference on Artificial Intelligence (IJCAI-95)

[Li 2002] LI Sujian, ZHANG Jian，HUANG Xiong, BAI Shuo，and LIU Qun,Semantic Computation in a Chinese Question-Answering System，Journal of Computer Science&Technology(JCST),Vol.17, No.6, 2002.

[Liu 1998] Qun Liu, Shiwen Yu, TransEasy: A Chinese-English Translation System based on hybrid approach, Third Conference of the Association for Machine Translation in the Americas (AMTA-98), Langhorne, PA, USA, Oct. 1998, In: David Farwell, et al, Eds., Machine Translation and the Information Soup, Lecture Notes in Artificial Intelligence Vol. 1529, Springer, pp514-517, 1998

[Lü 2001] Yajuan Lü , Ming Zhou, Sheng Li, Changning Huang, Tiejun Zhao. Automatic translation template acquisition based on bilingual structure alignment. International Journal of Computational Linguistics and Chinese Language Processing. 6(1), pp. 1-26.

[Lü 2002] Yajuan Lü , Sheng Li, Tiejun Zhao, Muyun Yang, Learning Chinese Bracketing Knowledge Based on a Bilingual Language Model, Conference on Computational Linguistics, August 2002, Taipei

[Marcus 1993] Mitchell P. Marcus, Beatrice Santorni, etc., Building a Large Annotated Corpus of English:The Penn Treebank, Computational Linguistics, Vol.19, No.2, 1993.

[Marshall 1983] Marshall I., Choice of Grammatical Word-class without Global Syntactic Analysis: Tagging Words in the LOB Corpus. Computers and the Humanities 17, 139-50.

[Matsumoto 1993] Matsumoto, Y., Ishimoto, H., and Utsuro, T., Structural Matching of Parallel Texts, ACL-93, pp. 23-30.

[Melamed 1996] Melamed I. D., Automatic Construction of Clean Broad-Coverage Translation

Lexicons. In: Conference of the Association for Machine Translation in Americas, Montreal, Canada, 1996

[Melamed 2000] I. Melamed, Models of translational equivalence among words. Computational Linguistics, 26(2), 2000

[Meyers 1996] Meyers, A., Yanharber, R., and Grishman, R., Alignment of Shared Forests for Bilingual Corpora. Proc. Of COLING-96, pp460-465.

[Nagao M. 1984] Nagao M., A framework of a mechanical translation between Japanese and English by analogy principle, *in* `Artificial and Human Intelligence: edited review papers at the International NATO Symposium on Artificial and Human Intelligence sponsored by the Special Programme Panel held in Lyon, France, October, 1981', Elsevier Science Publishers, Amsterdam, chapter 11, pp. 173-180, 1984

[Ng & Lua 2002] Hong I Ng and Kim Teng Lua, 2002, A Word Finding Automation for Chinese Sentence Tokenization, submitted to ACM Transaction of Asian Languages Processing. (Can be downloaded from: http://cslp.comp.nus.edu.sg/luakt/paper/publication.html)

[Nie 1999] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand, Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.74-81, August 15-19, 1999.

[NIST 2002] NIST Research Report, Automatic evaluation of Machine Translation Quality: Using n-gram Co-occurence Statistics, Research Report for NIST MT Evaluation, http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf

[Och 1998] Franz Josef Och and Hans Weber. Improving statistical natural language translation with categories and rules. In Proc. Of the 35th Annual Conf. of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics, pages 985-989, Montreal, Canada, August 1998.

[Och 1999] F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. In Proc. of the Joint SIGDAT Conf. On Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20-28, University of Maryland, College Park, MD, June 1999.

[Och 2001] Franz Josef Och, Hermann Ney. What Can Machine Translation Learn from Speech Recognition? In: proceedings of MT 2001 Workshop: Towards a Road Map for MT, pp. 26-31, Santiago de Compostela, Spain, September 2001.

[Och 2002] Franz Josef Och, Hermann Ney, Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, ACL2002

[Papineni 1997] K. A. Papineni, S. Roukos, and R. T. Ward. 1997. Feature-based language understanding. In European Conf. on Speech Communication and Technology, pages 1435-1438, Rhodes, Greece, September.

[Papineni 1998] K. A. Papineni, S. Roukos, and R. T. Ward. 1998. Maximum likelihood and discriminative training of direct translation models. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, pages 189-192, Seattle, WA, May.

[Papineni 2001] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, IBM Research, RC22176 (W0109-022) September 17, 2001

[Peng 2001] Peng, F. and Schuurmans, D., Self-supervised Chinese Word Segmentation, In:

Proceedings of the Fourth International, Symposium on Intelligent Data Analysis (IDA-2001), 2001.

[Rabiner 1986] L.R. Rabiner and B.H. Juang, An Introduction to Hidden Markov Models. IEEE ASSP Mag., Pp.4-166, Jun. 1986

[Rabiner 1989] L. R.Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of IEEE 77(2): pp.257-286, 1989

[Rayner 1997] Manny Rayner, David Carter, Hybrid Language Processing in the Spoken Language Translator, Proceedings of ICASSP-97, pages 107-110, Munich, Germany.

[Resnik 1997] P. Resnik and I. Melamed, Semi-automatic acquisition of domain-specific translation lexicons. In proceedings of The Fifth Conference on Applied Natural Language Processing,.pp.340-347, Washington DC, USA, 1997

[Ronald 1995] Ronald A. Cole, et al., eds., Survey of the State of the Art in Human Language Technology, 1995, http://cslu.cse.ogi.edu/HLTsurvey

[Sato & Nagao 1990] Sato, S. & Nagao, M. Toward memory-base translation, *in* H. Karlgren, ed., `Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)', Vol. 3, Helsinki: Helsinki University, pp. 247-252, 1990

[Shen 1997] Dayang Shen, Maosong Sun and Changning Huang. 1997. The application & implementation of local statistics in Chinese unknown word identification. In COLIPS, Vol. 8. (in Chinese).

[Smadja 1996] Smadja F., McKeown K.R. and Hatzivassiloglou V., Translation collocations for bilingual lexicons: a statistical approach. In: Computational Linguistics 22(1):1~38, 1996

[Sproat 1994] R. Sproat, C. Shih, W. Gale, N. Chang, A Stochastic Finite-State Word Segmentation Algorithm for Chinese, Proc. of 32nd Annual Meeting of ACL, New Mexico, 1994

[Sproat 1990] Richard Sproat and Chilin Shih. 1990. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, 4:336-35

[Sproat 2003] Richard Sproat, Thomas Emerson. The First International Chinese Word Segmentation Bakeoff, The 2nd SIGHAN Workshop attached with the ACL2003, 2003.7, pp.133-143

[Steven 1988] Steven J. DeRose, Grammatical Category Disambiguation by Statistical Optimization, Computational Linguistics Vol.14, No.1: 31-39, 1998

[Sumita & Iida 1991] Sumita, Eiichiro & Hitoshi Iida: 1991, `Experiments and prospects of example-based Machine Translation', in Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91), Berkeley, CA, USA, pp. 185-192.

[Sun 1997] Sun Maosong; Shen Dayang; Huang Changning, CSeg&Tagl.0: A Practical Word Segmenter and POS Tagger for Chinese Texts, Fifth Conference on Applied Natural Language Processing, 1997

[Sun 2002] Sun J., Gao J. F., Zhang L., Zhou M Huang, C.N. Chinese Named Entity Identification Using Class-based Language Model, Proc. of the 19th International Conference on Computational Linguistics, Taipei, 2002, pp 967-973

[Takeda 1996] Koichi Takeda, Pattern-Based Context-Free Grammars for Machine Translation, Proc. of 34th ACL, pp. 144-- 151, June 1996

[Turcato 1999] Turcato, D., McFetridge, P., Popowich, F. & Toole, J. (1999), A unified

example-based and lexicalist approach to machine translation, *in* `Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '99)', Chester.

[Vogel 1996] S. Vogel, H. Ney, and C.Tillmann, HMM-based word alignment in statistical translation. In COLING-96, pp.836-841, 1996.

[Wahlster 2000] Wolfgang Wahlster, Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System, In Wolfgang Wahlster eds., Verbmobil: Foundations of Speech-to-Speech Translation, pp 3-21, Springer, 2000, ISBN 3-540-67783-6

[Wang 1998a] Y. Wang and A. Waibel. Modeling with Structures in Statistical Machine Translation. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Montreal, Canada. August 1998.

[Wang 1998b] Ye-Yi Wang, Grammar Inference and Statistical Machine Translation, Ph.D Thesis, Carnegie Mellon University, 1998

[Watanabe 2000] Watanabe H., Kurohashi S.., Aramaki E., Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation. COLING-2000.

[Watanabe 2002] Taro Watanabe, Kenji Imamura, Eiichiro Sumita, Statistical Machine Translation Based on Hierarchical Phrase Alignment, proceedings of TMI 2002

[Wu 1995] Dekai Wu. Stochastic Inversion Transduction Grammars, with Application to Segmentation, Bracketing, and Alignment of Parallel Corpora. 14th Intl. Joint Conf. On Atifical Intelligence, pp1328-1335, Montreal, Aug, 1995. IJCAI-95

[Wu 1997] Dekai Wu, Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora, Computational Linguistics Vol.23 No.3 1997.

[Wu 1998] Andi Wu, Zixin Jiang, Word Segmentation in Sentence Analysis, International Conference on Chinese Information Processing, Beijing, pp. 169-180, 1998.

[Wu 2000] Wu, Andi, J. and Z. Jiang, 2000. Statistically-enhanced new word identification in a rule-based Chinese system, in Proceedings of the 2nd Chinese Language Processing Workshop, pp. 46-51, HKUST, Hong Kong.

[Xia 2000] Fei Xia, The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0), October 17, 2000

[Xue 2002] Nianwen Xue and Susan P. Converse, Combining Classifiers for Chinese Word Segmentation, First SIGHAN Workshop attached with the 19th COLING, pp.63-70., 2002.8

[Yamada 2001] K. Yamada and K. Knight, A Syntax-Based Statistical Translation Model, in Proc. of the Conference of the Association for Computational Linguistics (ACL), 2001

[Ye 2002] Ye S.R, Chua T.S., Liu J. M., An Agent-based Approach to Chinese Named Entity Recognition, Proc. of the 19th International Conference on Computational Linguistics, Taipei, 2002, pp 1149-1155

[Zhang 1999] Min Zhang, Key-Sun Choi, Multi-Engine Machine Translation: Accomplishment of MATES/CK System, Proceedings of TMI99, pages:228-238

[**白硕** 1998] **白硕，论语义重心偏移，**http://www.nlp.org.cn/docs/download.php?doc_id=74

[**柏晓静** 2002] **柏晓静，常宝宝，詹卫东，吴拥华，构建大规模的汉英双语平行语料库，黄河燕主编，机器翻译研究进展（全国机器翻译研讨会论文集），**pp.124-131**，电子工业出版**

社，2002.11
[常宝宝 1998] 常宝宝，刘颖，刘群，汉英机器翻译中的冠词处理研究，中文信息学报，1998 年第 3 期
[董振东 1999] 董振东，董强，知网，http://www.keenage.com
[杜飞龙 1999]《知网》辟蹊径，共享新天地——董振东先生谈知网与知识共享，《微电脑世界》杂志，1999 年第 29 期
[冯志伟 2001] 冯志伟，机器翻译研究的历史和现状，在新加坡的讲学演稿，2001（可从以下网址下载：http://naxun.sjtu.edu.cn/articles/fengzhiwei/Mt.ppt）
[李建华 2000] 李建华，王晓龙，中文人名自动识别的一种有效方法，高技术通讯，High Technology Letters，Vol.10，No.2，P.46-49，2000
[李涓子 1999]，汉语词义排歧方法研究，清华大学博士论文
[刘群 1997] 刘群，詹卫东，常宝宝，刘颖，一个汉英机器翻译系统的计算模型与语言模型，第三届全国智能接口与智能应用学术会议，吴泉源，钱跃良主编，智能计算机接口与应用进展，第 253-258 页，电子工业出版社，1997.8
[刘群 1998] 刘群，俞士汶，汉英机器翻译的难点分析，International Conference on Chinese Information Processing，黄昌宁主编，1998 中文信息处理国际会议论文集，第 507-514 页，清华大学出版社，1998.11
[刘群 2002] 刘群，张彤，汉英机器翻译系统扩充词典的建造，黄河燕主编，机器翻译研究进展（全国机器翻译研讨会论文集），pp.25-33，电子工业出版社，2002.11
[鲁松 2001] 鲁松，自然语言中词相关性知识无导获取和均衡分类器的构建，中国科学院计算技术研究所博士论文
[马红妹 2002] 马红妹，汉英机器翻译中汉语上下文语境的表示与应用研究，国防科学技术大学博士论文，湖南长沙，2002
[清华大学 1998] 汉语语料库词性标注规范，清华大学计算机系智能技术与系统国家重点实验室技术资料，1998.10
[宋柔等 1993] 宋柔，朱宏，潘维桂，尹振海，基于语料库和规则库的人名识别法，陈力为主编，计算语言学研究与应用，北京语言学院出版社，1993
[沈达阳 1995] 沈达阳，孙茂松，中文地名的自动辨识，见：陈力为、袁琦主编，计算语言学进展与应用，清华大学出版社，68—74 页，1995
[孙健 2002]孙健，基于统计方法的短语识别和句法结构歧义消解的研究，北京邮电大学工学博士学位论文，2002
[孙茂松 1997] 孙茂松, 黄昌宁, 邹嘉彦, 陆方, 沈达阳. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义，计算机研究与发展，第 34 卷第 5 期，第 332-339 页，19
[孙茂松 1993] 孙茂松，张维杰，英语姓名译名的自动辨识，陈力为主编，计算语言学研究与应用，北京语言学院出版社，1993
[孙茂松 1995] 黄昌宁，高海燕等，中文姓名的自动辨识，中文信息学报，Vol.19, No.2，1995
[孙茂松 1999] 孙茂松，左正平，邹嘉彦，1999, 高频最大交集型歧义切分字段在汉语自动分词中的作用，中文信息学报，Vol.13，No.1，1999
[孙茂松 2000] 孙茂松，左正平，黄昌宁，2000，汉语自动分词词典机制的研究实验，中文信息学报，Vol.14，No.1，2000
[孙茂松 2001] 孙茂松 邹嘉彦，汉语自动分词研究评述（A Review and Evaluation on Automatic Segmentation of Chinese），见：当代语言学（Contemporary Linguistics），第 3 卷，第 1 期，22-32 页，2001 年，北京
[王斌 1999] 王斌，汉英双语语料库自动对齐研究，中国科学院计算技术研究所博士论文，

1999

[王惠 1998] 王惠，詹卫东，刘群，《现代汉语语义词典》的概要及设计，中文信息处理国际会议论文集，清华大学出版社，1998

[王伟 2002] 王伟，机器翻译中的对齐技术研究，北京邮电大学博士论文，2002

[俞士汶 1991] 俞士汶等，机器翻译译文质量自动评估系统，中国中文信息学会 1991 年会论文集, PP314~319

[俞士汶 1998] 俞士汶，朱学锋，王惠，张芸芸，现代汉语语法信息词典详解，北京：清华大学出版社，第 1 版，1998 年 4 月

[俞士汶 1999] 俞士汶主编，现代汉语语料库加工——词语切分与词性标注规范与手册，北京大学计算语言学研究所，1999 年 3 月，http://icl.pku.edu.cn/research/corpus/spec.htm

[俞士汶 2000] 俞士汶、朱学锋、段慧明，大规模现代汉语标注语料库的加工规范，《中文信息学报》，2000 年 6 期, pp. 58-64

[语用所 2002] 教育部语言文字应用研究所计算语言学室“语料库加工”课题组，信息处理用现代汉语词类及词性标记集规范（征求意见稿），2002-04-08

[詹卫东 2000] 詹卫东，面向中文信息处理的现代汉语短语结构规则研究，清华大学出版社，广西科学技术出版社，2000

[张俊盛 1992] 张俊盛，陈舜德，郑萦，多语料库做法之中文姓名辨识，中文信息学报，Vol.16, No.3，1992

[张艳丽 2001] 张艳丽，黄德根等，统计和规则相结合的中文机构名称识别. 自然语言理解与机器翻译,清华大学出版社.2001.p233-p239

[郑家恒 1993] 郑家恒，刘开瑛，自动分词系统中姓氏人名处理策略探讨，见：陈力为主编，计算语言学研究与应用，北京语言学院出版社，1993

[周强 1999] 周强，黄昌宁，基于局部优先的汉语句法分析方法，软件学报，Vol.10，No.1，pp1-6，1999

[周锡令 2003] 周锡令，对朱德熙著《语法答问》中一个论断的质疑，2003-7-3，见：http://www.nlp.org.cn/docs/download.php?doc_id=301

[朱德熙 1985] 朱德熙，语法答问，商务印书馆，1985.7

作者在攻读博士学位期间发表的论文

按发表时间排序：

Qun Liu, A Chinese-English Machine Translation System Based on Micro-Engine Architecture, International Conference on Translation and Information Technology, Hong Kong, 2000

Qun Liu, Baobao Chang, Weidong Zhan, Qiang Zhou, A News-oriented Chinese-English Machine Translation System, International Conference on Chinese Computing (ICCC2001), Singapore, 2001

李素建，刘群，白硕，统计和规则相结合的汉语组块分析技术，计算机研究与发展, 2002(4)，**第** 381-385 **页**

张华平，刘群，基于 N-**最短路径方法的中文词语粗分模型**,**《中文信息学报》第** 16 **卷第** 5 **期**, 2002 **年**

刘群，汉语词法分析与句法分析技术综述，第一届学生计算语言学研讨会专题报告，2002.8

Huaping Zhang, Qun Liu, Hao Zhang, Xueqi Cheng, Automatic Recognition of Chinese Unknown Words Based on Role Tagging, 19th International Conference on Computational Linguistics,SigHan Workshop, 2002.8.

刘群，李素建，基于《知网》的词汇语义相似度计算，Computational Linguistics and Chinese Language Processing，Vol.7, No.2, August 2002, pp.59-76

刘群，张浩，白硕，中文信息处理开放资源平台，《语言文字应用》，No.4，Nov.2002，pp.50-56

刘群，张彤，汉英机器翻译系统扩充词典的建造，黄河燕主编，机器翻译研究进展（全国机器翻译研讨会论文集），pp.25-33，**电子工业出版社，**2002.11

刘群，统计机器翻译综述，中文信息学报，Vol.17，No.4，pp.1-12，2003.7

Hong-Kui Yu, Hua-Ping Zhang, Qun Liu, Recognition of Chinese Organization Name Based on Role Tagging, in Maosong Sun, Tianshun Yao, Chunfa Yuan, eds., Advances in Computation of Oriental Languages, Proceedings of 20th International Conference on Computer Processing of Oriental Languages, Tsinghua University Press, August 2003, pp. 79-87**（中文）**

Hua-Ping Zhang, Qun Liu, Hong-Kui Yu, Xue-Qi Cheng, Shou Bai, Chinese Named Entity Recognition Using Role Model, Computational Linguistics and Chinese Language Processing, Vol.8, No.2, August 2003, pp. 29-60

Hua-Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang and Hong-Kui Yu, Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, proceedings of 2nd SigHan Workshop, July 2003, pp. 63-70

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong and Qun Liu, HHMM-based Chinese Lexical Analyzer ICTCLAS, proceedings of 2nd SigHan Workshop, July 2003, pp.184-187

刘群，基于微引擎流水线的机器翻译系统结构，计算机学报（已录用，将于 2004 **年** 5 **月发表）**

刘群，张华平，俞鸿魁，程学旗，基于层次隐马模型的汉语词法分析，计算机研究与发展（已录用，将于 2004 **年** 6 **月发表）**

张华平，刘群，基于角色标注的中国人名自动识别研究，计算机学报，Vol.27, No.1, pp.85-91, 2004.1

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

按照学校要求提交学位论文的印刷本和电子版本；

学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务；

学校可以采用影印、缩印、数字化或其它复制手段保存论文；

在不以赢利为目的的前提下，学校可以公布论文的部分或全部内容。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日

致谢

在本文即将完成之际，我要向所有对我提供过支持和帮助的人表示感谢！

首先，我要真诚地感谢我的导师俞士汶教授。在 1992 年我硕士毕业的时候，俞老师就审阅了我的硕士论文，并对我的文章提出了很多细致中肯的建议，我现在仍记忆犹新。在以后十余年的学习和工作中，俞老师一直从各方面给了我大力的支持和帮助。从 1999 年开始攻读博士研究生后，俞老师更是对我悉心指导、严格要求，让我能时时看到自己的不足之处，督促我在学业上不断取得进步。俞老师为人谦和的处事风格、淡薄名利的治学态度、严谨扎实的工作作风，是值得我永远学习的榜样。

我要感谢给我提供学习机会的北京大学计算语言学研究所。计算语言所在俞老师的领导下，形成了一种非常好的研究氛围，每个人既有自己的研究自由，同时又在为这个集体贡献自己的力量。每周的讨论班，使我收益多多。研究所的陆俭明老师和朱学锋老师都是我景仰的前辈，他们都在我的学习、生活工作中给我提供过指导和帮助，对此我心存感激。詹卫东博士、常宝宝博士、于江生博士、孙斌博士、王厚峰博士、胡俊峰博士、陈玉忠博士都是我多年的合作伙伴、同事和朋友，我们经常在一起就一些学术问题所进行的探讨和思想的碰撞，不仅使我学到了很多知识，更促使我对一些问题进行深入的思考。在这几年里，我和詹卫东博士、常宝宝博士、王厚峰博士、周强博士（清华大学）、陈玉忠博士、王惠博士、吴云芳博士、段慧明老师、柏晓静同学等人共同完成了由孙茂松教授（清华大学）和俞老师领导的国家重点基础研究子课题“面向新闻领域的汉英机器翻译系统”，在共同的研究过程中，我们互相学习、互相促进，不仅增长了知识，也锻炼了能力，增进了友谊。

我要感谢我的工作单位中国科学院计算技术研究所。我读博士学位的这五年，我有幸亲眼目睹了计算所在以李国杰院士为所长的新一届班子领导下从衰弱到转型到重新奋起的整个过程，我个人所在的部门也从老计算所的二室到新计算所的软件室、数字化室，这种经历对我来说是非常宝贵的。计算所的研究氛围和北大有很大差别，这里更强调大规模的工程项目的组织和管理，在这两种差异很大、同时又都非常有效的研究环境中工作，与这么多优秀的人共事，不仅使我开阔了眼界，也使我各方面的能力得到了实实在在的锻炼。高庆狮院士是我国计算机界最早的中科院院士之一，正是高院士在 1980 年代后期在计算所组织开展的机器翻译研究，把我带入了这个研究领域。当时我被保送到计算所读硕士研究生的时候，我报的导师就是高庆狮院士。后来由于某些原因我没有能成为高院士的硕士研究生，我觉得非常遗憾。这几年高院士回到计算所以后，一直对我非常关心，多次找我长谈，在学习和工作上给我很多的指导和督促，对此我感激不尽。白硕研究员是我尊敬的学者，他渊博的知识、跳跃的思维、对学术问题深刻的洞察力，让我非常佩服，和白老师每一次讨论，都让我受益匪浅。特别我在初到软件室的这段时间里得到了他多方面大力帮助，使我很快融入到一个新的环境之中。我还要感谢数字化室和软件室的各位领导和老师，包括钱跃良老师、李锦涛研究员、林守勋研究员、程学旗研究员、郭莉副研究员，是他们给我提供了一个良好的工作环境，使我在完成本职工作的同时，还能抽出时间来完成我的学业。我特别要感谢我所在的自然语言处理课题组的所有同事和研究生，如王斌博士、李素建博士、孙健博士、周立新博士、张华平、骆卫华、邹刚、邓丹、俞鸿魁、侯宏旭、熊德意、刘洋、张浩、李继锋、张健、张奕涛、张彤、王长胜等等，这几年我们在一起摸爬滚打，完成了一项又一项的研究工作，本文的很多研究成果都和他们的工作有直接和间接的关系。特别是张华平和我一起研究并实现了基于层叠隐马尔可夫模型的汉语词法分析系统，他的很多创造性工作是这项研究取得成功的不可或缺的因素。张浩的句法分析研究、邓丹的词语对齐研究、俞鸿魁的机构名识别研究

和他开发词语对齐、短语对齐辅助校对工具，都对本论文的研究工作有很大的帮助。

我还要感谢参加过我的综合考试、开题报告、预答辩评审的各位老师，包括高庆狮院士、冯志伟教授、陆俭明教授、周锡令教授、黄河燕研究员、孙斌副教授、王厚峰副教授、常宝宝博士，他们对我每一阶段的工作都提出了很多中肯而宝贵的意见，帮助我不断改进自己的研究工作。在此，我还要预先感谢我的博士论文的匿名评审人和参加我的博士答辩的答辩委员，谢谢他们为此付出的辛苦劳动。

最后，我要感谢我的家人对我无私的支持和帮助。我的父母不仅仅对我有养育之恩，更用他们的行动教给了我做人的道理。在这五年里，我的儿子刘思诺已经从一个呱呱坠地的婴儿长成了一个四岁多的“大哥哥”，他给我和我们这个家庭带来的欢乐是任何人都不能比拟的。我的妻子辛利不仅要忙于工作，还替我承担了大部分家务和抚养教育儿子的责任，对此我只能表示愧疚和感激。我希望我的这篇文章没有辜负我的家人为我付出的辛苦，我更希望今后有更多的时间来陪伴他们，弥补我对他们感情的亏欠。

需要感谢的人太多，无法一一列举。最后，再一次真诚的感谢所有关心过我、帮助过我的老师、同事、同学和朋友们！