沈阳航空航天大学

硕士学位论文

汉语术语语义分析技术研究及其应用

姓名：陈小芳

申请学位级别：硕士

专业：计算机软件与理论

指导教师：张桂平

201101

摘 要

汉语语义分析是自然语言处理的核心技术之一，也是对汉语的深层理解。汉语语义分析效果的提高对于信息检索和机器翻译都具有推动作用。随着科技的发展，术语也不断涌现，所以对术语语义分析的研究也有着重大的意义。

本文对术语的特点进行研究，实现了汉语术语语义分析系统。该系统包括依存分析和语义分析两个部分，这两部分通过机器学习的方法实现。在语义分析的基础上实现了汉语术语翻译系统，并取得了较好的效果。具体内容如下：

首先，对大量的汉语术语进行分析，发现术语大部分为名词性短语，而且术语用词的重复性低。针对上述两个特点，在依存分析阶段，本文选择适合术语的特征，利用支持向量机（SVM）训练得到依存分析模型，从而有效识别出了术语内部的依存关系。所选特征包括基本特征，互信息特征和知网第一义原特征。

其次，提出了一种汉语术语语义分析方法。本文定义了 14 种语义关系，并利用 CRF 训练得到语义分析模型，该模型可以有效识别出两个词之间的语义关系。由于术语所涉及的语义关系范围较窄，所以该模型对于易混淆的类别分类能力较差。本文对于易混淆的类别采用 SVM 训练分类器，对 CRF 模型输出的 2-best 结果中的两个语义关系进行识别，确定词对最终语义关系。

最后，将语义分析技术应用到了术语翻译。首先对术语进行依存分析，根据依存分析的结果抽取出结构化的短语，再利用传统的 GROW-DIAL-FINAL 方法抽取非结构化短语；之后利用提取的调序模板对源语言进行调序；最后利用摩西对已经调序的术语解码。

实验结果表明语义分析方法的有效性，在大类语义关系和小类语义关系上正确率分别达到 77.13%和 69.05%。将语义分析结果应用到术语翻译，使翻译的效果有所提高。

关键词： 依存分析；语义分析；SVM；CRF；术语翻译；

Abstract

Chinese semantic analysis is one of key technologies of natural language processing, and contributes to well comprehend Chinese. The improvement of Chinese semantic analysis will play an important role in information retrieval and machine translation. As the development of technology, terms appear continuously, so the semantic analysis of terms is of great significance.

Based on the research of the term, the paper introduces a term semantic analysis system. The system includes two parts: dependency analysis and semantic analysis. The two parts are based on the method of machine learning. Based on the semantic analysis, the Chinese term translation system is realized, and the specific contents are as follows:

Firstly, we analyze a lot of terms, and find out most of terms are noun phrase and low recurrence rate of terms, so we choose the proper features for terms in the dependency analysis stage based on the two features and we train the support vector machine (SVM) dependency model to identify the dependency relationships within terms.The proper features include basic feature, mutual information and the first sememe of words in hownet,

Secondely, this paper proposes an approach for Chinese term semantic analysis. Firstly, we define 14 semantic relationships and then train CRF model to identify the semantic relationships between two words. But the range of semantic relationship within terms is so finite that the semantic model cannot comprehensively identify the confusion categories. So we train SVM model to solve this problem. After the CRF model outputs 2-best semantic results, we use SVM model to identify the final result in 2-best results.

At last, the result of the semantic analys is applied to term translation. In the first stage, we extract the constituent pharses based on the result of the dependency analysis and extract the non-constituent phrases by the method of GROW-DIAL-FINAL, and then abstract the ranking template. On the basis of the ranked source language, we use Moses to decode the ranked terms.

Experimental results show that the method is effective, and the accuracy of semantic analysis reaches 77.13% and 69.05% respectively in parent-category and sub-category. the result of the semantic analys is applied to term translation and the translation result are better than before.

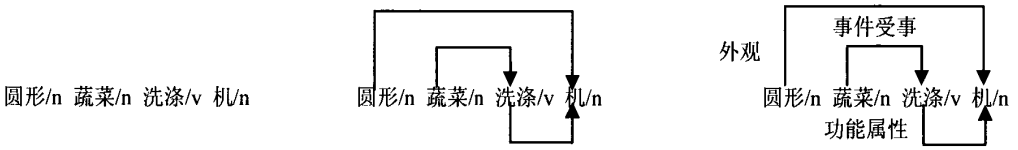
Keywords: Dependency analysis; Semantic analysis; SVM; CRF; Term translation

第1章 引言

随着专利文献的不断涌现，术语随之增多，对术语的正确识别与分析成为自然语言处理的重要组成部分。组成术语的词之间的语义分析对机器翻译和信息检索都有着重要作用。

1.1 术语语义分析过程

语义分析是找出结构意义及其集合意义，从而确定语言所表达的真正含义或概念。本文语义分析的过程首先将经过分词和词性标注的术语进行依存分析，之后确定具有依存关系的两个词的语义关系。过程如图 1.1 所示。



（a）词性标注后的术语　（b）术语依存分析结果　（c）术语语义分析结果

图 1.1 术语语义分析过程

1.2 课题的提出

术语是通过语音或文字来表达或限定科学概念的约定性语言符号[1]。分为单词术语和多词术语，且大部分为多词术语。分析组成术语的词之间的语义关系对理解术语的意思和该领域的科学发展都有着至关重要的作用。

宗成庆指出：自然语言处理的最终目的应该是在语义理解的基础上实现相应的操作[2]。一个自然语言处理系统如果有语义的参与能够大大提高系统的性能，但是语义分析对于计算机来说是一个前所未有的挑战，如何能模拟人的思维方式使计算机能够像人一样理解句子所表达的语义至今还是一个难题。

汉语语义分析大体可以分为两个过程：依存分析过程和确定具有依存关系的两个词之间的语义关系过程。其中依存分析已经成为自然语言处理研究的核心任务，也是目前研究的难点问题。但是对汉语的依存分析还很难满足工业应用的需求，而且研究的重点都是句子级的，对名词短语特别是术语的研究还很少。这种状况间接影响了术语自动翻译技术和专利检索技术的发展。

术语翻译和长句子翻译是影响专利翻译的两个重要因素。将术语翻译单独提出研究是必要的，如果术语翻译的效果提高，那么整个专利翻译乃至整个机器翻译的效果也会随之提高。术语集中体现和负载了一个学科领域的核心知识，术语翻译的准确率高低直接影响着科学的发展，也能促进使用不同语言的专业知识的交流。

1.3 本文的研究意义

1.3.1 术语语义分析

语义分析是自然语言处理的底层技术，可以应用于许多上层技术当中。例如：机器翻译，信息检索等。

目前的机器翻译还停留在基于句法[3-5]的层次上，这种方法解决了基于短语翻译的全局排序和短语非连续性问题。但句法层次的翻译还没能理解一个句子或者一个短语所表达的意思，这种状况阻碍了翻译过程中调序和选词，也直接影响到了翻译的效果。例如："冷却 塔"在翻译过程中如果能够分析出"冷却"是说明"塔"的功能，那么我们在选词的过程就会选择"cooling tower"而不是"cool tower"，因为英文中一般用动名词来表示另一个词的功能；另外，汉语和英语中修饰词的语序也有差异，所以得到术语的语义结构对汉英翻译系统中源语言调序也有帮助。

传统的信息检索是对关键词的匹配，当关键词为专业术语时，所能匹配的文档数量远远达不到人们的需求，如果从语义层次上对关键词进行理解，可以提高信息检索系统的性能。

1.3.2 术语自动翻译

术语的自动翻译技术除了应用于机器翻译以外还可以应用到其他领域，例如：双语词典构建，跨语言信息检索等。

术语自动翻译作为机器翻译的子任务，可以帮助机器翻译建立专业词典，从而提高机器翻译的效率。在专利翻译过程中，术语的翻译是亟待解决的难题，如果术语翻译性能有所改善，必将大幅度改善专利翻译的结果。

跨语言的信息检索是用户用一种语言提问，检索出用另一种语言书写的信息，也就是一种不受语言限制进行检索的问题。术语自动翻译技术可以提供准确的翻译词对从而提高跨语言信息检索的效果。

术语翻译和长句子翻译是影响专利翻译的两个重要因素。将术语翻译单独提出研究是必要的，如果术语翻译的效果提高，那么整个专利翻译乃至整个机器翻译的效果也会随之提高。术语集中体现和负载了一个学科领域的核心知识，术语翻译的准确率高低直接影响着科学的发展，也能促进使用不同语言的专业知识的交流。

1.3 本文的研究意义

1.3.1 术语语义分析

语义分析是自然语言处理的底层技术，可以应用于许多上层技术当中。例如：机器翻译，信息检索等。

目前的机器翻译还停留在基于句法[3-5]的层次上，这种方法解决了基于短语翻译的全局排序和短语非连续性问题。但句法层次的翻译还没能理解一个句子或者一个短语所表达的意思，这种状况阻碍了翻译过程中调序和选词，也直接影响到了翻译的效果。例如："冷却 塔"在翻译过程中如果能够分析出"冷却"是说明"塔"的功能，那么我们在选词的过程就会选择"cooling tower"而不是"cool tower"，因为英文中一般用动名词来表示另一个词的功能；另外，汉语和英语中修饰词的语序也有差异，所以得到术语的语义结构对汉英翻译系统中源语言调序也有帮助。

传统的信息检索是对关键词的匹配，当关键词为专业术语时，所能匹配的文档数量远远达不到人们的需求，如果从语义层次上对关键词进行理解，可以提高信息检索系统的性能。

1.3.2 术语自动翻译

术语的自动翻译技术除了应用于机器翻译以外还可以应用到其他领域，例如：双语词典构建，跨语言信息检索等。

术语自动翻译作为机器翻译的子任务，可以帮助机器翻译建立专业词典，从而提高机器翻译的效率。在专利翻译过程中，术语的翻译是亟待解决的难题，如果术语翻译性能有所改善，必将大幅度改善专利翻译的结果。

跨语言的信息检索是用户用一种语言提问，检索出用另一种语言书写的信息，也就是一种不受语言限制进行检索的问题。术语自动翻译技术可以提供准确的翻译词对从而提高跨语言信息检索的效果。

术语的不断涌现给词典构建带来了新的挑战，使得术语自动翻译技术成为了术语词典编纂关键技术。

1.4 本文的主要工作

本文的主要工作是研究汉语术语语义分析方法以及该方法在汉英机器翻译系统中的应用。首先研究汉语术语的特点，选择合适的特征对术语进行依存分析，之后再进行语义分析实现了汉语术语语义分析系统，并将语义分析结果应用到了术语翻译的短语抽取和源语言调序过程中。本文的工作包括以下三个方面。

术语的依存分析：首先对术语进行规则处理，之后再用支持向量机（SVM）模型对不满足已定义规则的词对进行处理。通过对大量的术语观察研究得到了影响术语依存分析的因素，我们从中选择了基本特征，互信息以及词语在知网中的第一义原作为支持向量机的特征，该方法的优点是加入知网中的第一义原特征从语义层次上进行判别，有效缓解了数据稀疏问题。

术语的语义分析：本文将术语内部的语义关系分为 14 个大类，并将其中的“宿主-属性”关系又进一步细分为 7 个小类。在语义分析过程中，首先，利用已定义的规则处理分词错误；之后，在基于条件随机场（CRF）的语义分析模型基础上加入了 SVM 后处理模型，对 CRF 输出的不可靠语义关系进行校正。

术语翻译：术语翻译方法包含三个阶段，短语抽取，源语言调序和摩西解码。在短语抽取阶段，我们加入了在句法层次上具有依存关系的词串作为短语，而不仅仅是传统方法中将连续的词串作为短语。在调序阶段，我们将术语语义分析结果和词对齐结果相结合，从训练语料中提取出调序模型，并利用该调序模型对源语言进行调序。

1.5 本文的组织结构

论文主要包括以下几章内容。

第一章：引言。主要介绍了术语语义分析的研究背景以及研究意义，然后介绍了本文的主要工作，最后介绍本文的组织结构。

第二章：相关研究。首先介绍汉语术语的特点，之后详细阐述依存分析，语义分析，机器翻译以及术语翻译的研究现状，并指出各种方法的优缺点。

第三章：基于统计和规则相结合的汉语术语语义分析方法。主要介绍了基于 SVM 的依存分析方法和基于 CRF 和 SVM 结合的语义分析方法。并对实验结果进行了分析。

第四章：基于语义分析的汉语术语翻译方法。详细介绍两种短语抽取的方法：结构化的短语抽取和非结构化的短语抽取，其次介绍了调序模板的提取，再次介绍摩西解码方法，最后给出了实验结果，以及和摩西系统的比较，指出该方法的优点与不足。

第五章：系统的设计与实现。首先介绍系统的整体框架，之后介绍了系统各个模块的具体实现方法。

最后是本文的总结部分，并指出该系统的不足以及下一步工作。

第 2 章　相关研究

术语是非通用领域概念的抽象，也代表非通用领域的核心知识，和通用领域的词语相比具有自身的特点。本章首先对术语所具有的特点进行介绍，之后对目前的语义分析技术和术语自动翻译技术所采用的方法进行总结。

2.1　术语的特点

从术语的外部特征看，术语具有领域性，结合紧密型和语言完备性三个特点。领域性指术语是在特定的领域中使用的，也就是说在某一个领域出现频繁而在其他领域很少出现；结合紧密性说明术语是一种半固定或者固定的短语或词；语言完备性指出术语是语言学上成立的词语。

从语言学角度来看，术语的组成结构有以下几个特点：

术语用字固定：据统计术语的常用字在 2000 个左右，而且和通用领域常用字有所不同。例如术语中最常出现的字为“器”，“机”，“电”等；而很少出现“是”，“些”，“和”等在通用领域经常出现的字。

中心词特点：汉语术语的一个明显特征就是中心词一般为最后一个词，且中心词一般为名词。

语义关系特点：通用领域词语之间的语义关系比较复杂，和通用领域的词语相比，组成术语的词之间的语义关系种类较少，最常出现的语义关系为“属性-宿主”关系，“事件-受事”关系。

结构特点：组成术语的词的形式大部分为“名词+名词”，“形容词+名词+名词”，且长度一般不超过 6 个词。

2.2　术语分析方法

术语作为现代汉语词汇集的一个有机组成部分，一方面有自己独特的构成方式，另一方面遵从“现代汉语词汇构成”这样一个大规律。对术语的分析可以从较浅层次进行，例如组成术语的部件描述；也可以从较深层的语义进行分析。在进行语义分析之前，首先要确定术语的句法结构，其次要有一个语义分类体系来支撑。本章将从依存分析和语

义分析两个方面进行介绍。

2.2.1 依存分析方法

依存文法是1959年在《结构句法基础》一书中被法国的著名语言学家特思尼耶尔提出的[6]。他认为一个句子中核心词为动词，其他词都直接或者间受到这个动词的支配，该动词本身不受其它词支配。之后，1970年美国语言学家罗宾逊提出了四条依存公理：1)一个句子中只有一个成分是独立的；2)其他成分直接依存于某一成分；3)任何一个成分都不能依存于两个或两个以上的成分；4)如果A成分直接依存于B成分，而C成分在句子中位于A和B之间的话，那么，C或者直接依存于A成分，或者直接依存于B成分，或者直接依存于A和B之间的某一成分。依存语法描述的是句子中词与词之间直接的句法关系。这种句法关系是有方向性的，通常是一个词支配另一个词，或者说，一个词受另一个词的支配(也即依存关系)。

在汉语中，研究者为汉语设置了不同种类的依存关系，周明、黄昌宁等曾设计了44种依存关系[7]。其中包括：主谓关系（SUBJ），表语补足语关系（PRD），限定关系（DET）等。图2.1表示了“那是一本书”的依存句法树。

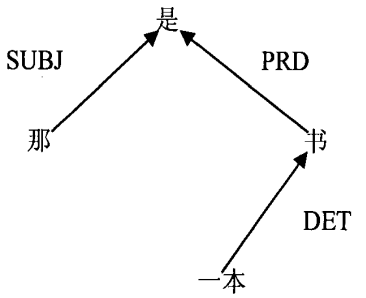


图 2.1 依存句法树

目前对依存分析方法的研究主要分为基于规则的方法和基于统计的方法。基于规则的方法主要利用语言学知识人为定义规则来处理依存分析问题，但是单纯规则的方法很难覆盖所有的语言现象，而且规则之间可能出现冲突问题。基于统计的方法通过大量的训练语料和统计工具的学习能力实现，一般采用支持向量机[8-10]，最大熵[11]，条件随机场[12]，在线学习[13]等方法。

1 基于规则的依存分析方法

基于规则的方法是利用语言学家对语言现象的认识，以知识为主体的方法。通过人

义分析两个方面进行介绍。

2.2.1 依存分析方法

依存文法是1959年在《结构句法基础》一书中被法国的著名语言学家特思尼耶尔提出的[6]。他认为一个句子中核心词为动词，其他词都直接或者间受到这个动词的支配，该动词本身不受其它词支配。之后，1970年美国语言学家罗宾逊提出了四条依存公理：1)一个句子中只有一个成分是独立的；2)其他成分直接依存于某一成分；3)任何一个成分都不能依存于两个或两个以上的成分；4)如果A成分直接依存于B成分，而C成分在句子中位于A和B之间的话，那么，C或者直接依存于A成分，或者直接依存于B成分，或者直接依存于A和B之间的某一成分。依存语法描述的是句子中词与词之间直接的句法关系。这种句法关系是有方向性的，通常是一个词支配另一个词，或者说，一个词受另一个词的支配(也即依存关系)。

在汉语中，研究者为汉语设置了不同种类的依存关系，周明、黄昌宁等曾设计了44种依存关系[7]。其中包括：主谓关系（SUBJ），表语补足语关系（PRD），限定关系（DET）等。图2.1表示了“那是一本书”的依存句法树。

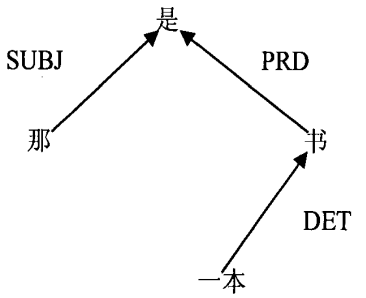


图 2.1 依存句法树

目前对依存分析方法的研究主要分为基于规则的方法和基于统计的方法。基于规则的方法主要利用语言学知识人为定义规则来处理依存分析问题，但是单纯规则的方法很难覆盖所有的语言现象，而且规则之间可能出现冲突问题。基于统计的方法通过大量的训练语料和统计工具的学习能力实现，一般采用支持向量机[8-10]，最大熵[11]，条件随机场[12]，在线学习[13]等方法。

1 基于规则的依存分析方法

基于规则的方法是利用语言学家对语言现象的认识，以知识为主体的方法。通过人

为制定规则并建立知识库，利用一些制约条件来解决句法歧义问题。2000 年，郭艳华等提出了一种基于规则的多级分析方法[14]。该方法首先对句子进行谓词识别，找出句子的中心词；之后再识别句子中其他词之间的依存关系。微软亚洲研究院也提出了一种基于规则的短语级别的依存方法，并在当时取得了较好的效果。尽管规则的方法在一定条件下表现出了它的优越性，但是规则本身也存在着很大的缺陷。首先，规则的获取十分复杂，主要依赖于语言学家的知识，并且需要大量的人力劳动；其次，规则难以覆盖所有的语言现象，所以在分析过程中难免会导致依存分析的效果差。随着统计方法的出现，特别是各种统计工具的出现，人们开始将统计的方法应用到了依存分析系统中，并取得了规则无法达到的效果。

2　基于统计的依存分析方法

基于统计的方法已经成为了依存分析的主流方法，目前有监督的依存分析[15-17]达到了较好的效果，但是有监督的依存分析在建立树库时耗费时间和人力，所以人们又开始关注无监督[18]的和半监督[19]的依存分析方法。这些方法通常采用产生式分析模型，判别式分析模型和决策式分析模型。

（1）产生式依存分析模型

产生式依存分析模型[8][20-23]是在 PCFG 的基础上产生并发展的，这种方法采用联合概率模型生成一系列依存句法树并赋予其概率分值，然后采用相关算法找到概率打分最高的分析结果作为最后输出，属于全局搜索算法，所以具有较高的计算复杂度。

产生式模型主要包括以词汇概率依存模型和依存生成概率模型。1996 年 collins 利用词汇概率依存模型实现了一个依存分析系统，该方法在训练阶段通过极大似然估计在树库中统计出任意两个词之间存在依存关系的概率；在测试阶段通过寻找具有依存关系的词对的依存概率的乘积最大的依存树来实现依存分析。同年，Eisner 利用依存生成概率模型进行实验，对于每棵候选依存树 T，整棵树的生成概率定义树中所有节点生成概率的乘积，之后寻找概率最大的依存树。

除了上述提到的两种方法，文献[8]提出了一种产生式模型和 SVM 相结合的依存分析方法。该方法首先利用产生式方法进行依存分析，将分析的结果进行人工校对，得到一组由错误结果构成的样本负例和一组经过人工校验的样本正例，用于 SVM 分类器训练。采用的三种产生式模型分别为：（1）N-gram 模型，（2）结构化最优型，（3）混合

模型。实验结果表明第三种分析方法取得了较好的结果。文献[23]将结构信息加入了依存分析模型，并利用动态规划算法实现了一个高效的依存分析系统。

（2）判别式依存分析模型

判别式方法利用条件概率模型，避开了联合概率模型所要求的独立性假设。该方法和传统的寻找最佳的依存方法不同，它将以往的方法转化成最优路径的搜索过程[24-26]。可以运用许多运筹学方法和机器学习方法，并且和产生式依存分析模型相比，在可计算性上具有优势，复杂度也有所降低。

判别式依存分析方法通常采用最大生成树模型和状态转移模型。最大生成树方法的主要思想为：给定一个包含 N 个词的句子，任意两个词之间都可能存在依存关系共有 N*(N-1)种可能的依存边，只是依存的强弱不同，将这种依存强弱表示为完全图中边的分数，于是该任务就变成寻找完全图中最大生成树的过程。状态转移模型是将分析的任意时刻作为一个状态，依据该状态下的特征做出某种决策，从而进入新的状态的过程。这种方法在决策式依存分析中也经常用到。

（3）决策式依存分析模型

决策式依存分析模型是逐步取一个预分析的词，为每次输出的词产生一个分析结果，直到句子结束为止。每次只保留一个分析结果所以该方法大大降低了算法的复杂度。该方法是对完全句法分析和部分句法分析的折中，它具备了部分句法分析鲁棒、有效和确定的特性，而且又能像完全句法分析那样得到完整句子的依存分析结果。这种方法的缺点在于不是寻找全局最优解，所以错误率难免会上升。目前研究者对决策式的研究侧重在了如何减小错误率上。

目前比较常用的决策式算法有 Niver 算法[27]，Yamada 算法[28]等。Niver 算法的基本思想是建立一个栈和一个队列，其中栈中存放当前所有具有依存关系的子序列；队列中存放还没有分析的词序列。利用栈顶和队列头部确定当前分析器的状态，决定进行哪种操作；其中操作包括规约、移进和两个词具有依存关系操作。任意状态下，有四种转移动作 left、right、reduce 和 shift。在一次扫描的过程中，分类器根据栈顶，队首以及其他特征预测要执行上述哪种动作，并转移到新的状态，这一过程直到入队的队列为空为止。这种方法的优点为对近距离依存关系识别的正确率高，而且可以任意加入特征，灵活性较高。

2.2.2 语义分析方法

目前对语义的研究还处于初级阶段，主要采用《知网》和《同义词词林》两种语义资源。文献[29]根据同义词林把词语分成18个语义类和59种语义关系，其中18种语义类来自于《同义词词林》；59种依存关系的定义来自《知网》中的定义的76种“动态角色”。手工标注各个词之间的语义关系，形成了一个大型的语义依存库。文献[30]在此基础上实现了一个语义分析集成系统，构建了结构化的语言模型。该系统能够自动分析句子中词之间的语义依存关系，词义正确率和语义依存标注正确率分别达到90．85%和75．84%，并将这个系统应用到语音识别，达到了很好的效果。文献[31]提出词汇的语义倾向性判别，利用《知网》中对汉语词汇的定义和描述，建立由褒贬倾向较强烈的词汇组成的种子集，并结合上下文环境因素的影响，采用一种度量方法获取种子词与普通词之间的语义倾向相似度，识别普通词的褒贬倾向。文献[32]提出了一种基于语义分析的中文文本特征值提取方法，并给出了具体算法。与传统特征值提取方法相比，该方法降低特征向量维数的目的。通过实验证明了特征值提取方法比基于出现概率的特征值提取更加能改善文本分类的性能。

2.3 术语翻译方法

随着科技的发展，术语随之增多，术语翻译也成为机器翻译的重要组成部分。对术语的翻译一般结合术语的特点并采用机器翻译的方法来实现。下面本文将分别介绍机器翻译和术语翻译的研究现状。

2.3.1 基于机器翻译的方法

机器翻译是 1949 年美国 Warren Weaver 在《翻译》中提出的。经过几十年的发展机器翻译已经加深了人们对知识和语言等问题的理解，也促进了科学的发展。目前机器翻译的方法主要包括基于规则的方法和基于统计的方法。

1　基于规则的机器翻译

基于规则的机器翻译系统[33-36]是对语言语句的词法、语义进行分析、判断和取舍，然后重新排列组合，最后生成等价的目标语言的过程。在上个世纪 90 年代以前统计机器翻译方法还没有兴起,基于规则的机器翻译一直占有主导地位。该技术发展到了今天已经取得了很大的成就，规则库在不断扩大，覆盖的语言现象也更加全面。传统的规则

2.2.2 语义分析方法

目前对语义的研究还处于初级阶段，主要采用《知网》和《同义词词林》两种语义资源。文献[29]根据同义词林把词语分成18个语义类和59种语义关系，其中18种语义类来自于《同义词词林》；59种依存关系的定义来自《知网》中的定义的76种“动态角色”。手工标注各个词之间的语义关系，形成了一个大型的语义依存库。文献[30]在此基础上实现了一个语义分析集成系统，构建了结构化的语言模型。该系统能够自动分析句子中词之间的语义依存关系，词义正确率和语义依存标注正确率分别达到90．85%和75．84%，并将这个系统应用到语音识别，达到了很好的效果。文献[31]提出词汇的语义倾向性判别，利用《知网》中对汉语词汇的定义和描述，建立由褒贬倾向较强烈的词汇组成的种子集，并结合上下文环境因素的影响，采用一种度量方法获取种子词与普通词之间的语义倾向相似度，识别普通词的褒贬倾向。文献[32]提出了一种基于语义分析的中文文本特征值提取方法，并给出了具体算法。与传统特征值提取方法相比，该方法降低特征向量维数的目的。通过实验证明了特征值提取方法比基于出现概率的特征值提取更加能改善文本分类的性能。

2.3 术语翻译方法

随着科技的发展，术语随之增多，术语翻译也成为机器翻译的重要组成部分。对术语的翻译一般结合术语的特点并采用机器翻译的方法来实现。下面本文将分别介绍机器翻译和术语翻译的研究现状。

2.3.1 基于机器翻译的方法

机器翻译是 1949 年美国 Warren Weaver 在《翻译》中提出的。经过几十年的发展机器翻译已经加深了人们对知识和语言等问题的理解，也促进了科学的发展。目前机器翻译的方法主要包括基于规则的方法和基于统计的方法。

1　基于规则的机器翻译

基于规则的机器翻译系统[33-36]是对语言语句的词法、语义进行分析、判断和取舍，然后重新排列组合，最后生成等价的目标语言的过程。在上个世纪 90 年代以前统计机器翻译方法还没有兴起，基于规则的机器翻译一直占有主导地位。该技术发展到了今天已经取得了很大的成就，规则库在不断扩大，覆盖的语言现象也更加全面。传统的规则

方法主要依靠语言学家总结的经验进行的，而现在更加注重自动从大规模语料库中获取规则。这些改变都大大提高了基于规则方法的性能。

基于规则的方法虽然在某些方面已经取得了较好的效果，但是由于语言是随着科技的发展而随时变化的，尽管已经有足够大的规则库，还是难以覆盖新出现的规则。而且自然语言中存在着大量的例外情况，当规则库比较庞大的时候可能产生很多冲突。规则的调试需要专家知识，非常耗时，并且很难保证修改后的规则不会带来新的冲突。由于上述原因，基于统计的机器翻译随之兴起。

2 基于统计的机器翻译

机器翻译从基于词的翻译模型发展到基于短语的翻译模型，之后又出现了基于句法的翻译模型。基于词的翻译模型主要是 IBM 提出的 5 个数学模型，该模型中由于对齐的粒度小，因此歧义词是一个普遍存在的现象；这使得机器翻译从基于词的翻译转向了基于短语的翻译，增大了对齐粒度，而且实现了句子的局部调序。虽然基于短语的翻译比基于词的翻译在效果上有了提高，但是在长距离调序和短语的连续性方面仍然存在这问题。之后人们将句法知识引入到统计机器翻译，并取得了很好的效果。该方法主要分为两类，基于形式化语法的方法和基于语言学语法的方法。前者基于纯形式文法，例如 SCFG(上下文无关文法)、同步替换文法、ITG 模型等；后者符合语言学家定义的语法规范，如句法分析、依存分析等。无论是基于短语的统计机器翻译还是基于句法的统计机器翻译，调序和短语的抽取都是其中的关键问题之一。由于不受限调序是一个 NP 问题，所以 Richard Zens 等提出了一种受限调序的模型，对调序模型进行减支，大大提高了调序模型的效率。之后又从句法层次上对句子进行调序，实现了长距离调序。目前对调序的研究还停留在句法层次上，没能考虑到句子中词与词之间的语义关系。短语抽取的过程大部分都是抽取连续的词串，而不是语言学意义上的短语，这也给机器翻译带来负面影响。

（1） 基于词的机器翻译

基于词的机器翻译（word-based statistical machine translation）技术是在上个世纪 90 年代发展起来的，主要的翻译模型为 IBM 的 5 个数学模型。公式 2.1 是这 5 个模型翻译模型的基础。

$$P(S,A/T)=P(m/T)\prod_{j=1}^{m}p(a_j/a_1^{j-1},s_1^{j-1},m,T)P(s_j/a_1^{j},s_1^{j-1},m,T) \quad (2.1)$$

由于公式 2.1 右侧参数太多，不能保证参数之间相互独立，所以，要对前提进一步限制。其中模型 1 满足如下三个条件：首先，假设$P(S,A/T)$与源语言的句子长度以及目标语言无关；其次，$p(a_j/a_1^{j-1},s_1^{j-1},m,T)$只依赖于目标语言的句子长度；最后，假设$P(s_j/a_1^{j},s_1^{j-1},m,T)$只依赖于$S_j$和$t_{a_j}$。模型 1 的方法容易实现，但是该方法的表现力有限。所以在此基础上又提出了模型 2，模型 2 引入了对位概率（alignment probabilities），在翻译的过程中考虑到了从源语言到目标语言之间的位置变化。模型 3 又在模型 2 的基础上增加了约束条件。该模型是三个集合：繁衍概率集合，翻译概率集合，扭曲概率集合。模型 4 对模型 3 的扭曲概率进一步完善。模型 5 消除了模型 4 中可能出现的零概率的情况。模型 5 虽然能力强大，当并为被广泛使用，相反模型 2，3，和 4 更容易被人们所接受。

（2） 基于短语的机器翻译

基于词的机器翻译存在这明显的缺陷，比如词和词之间都是孤立的，没有考虑上下文信息，这样导致了大量的错误。因此人们想到将词和词捆绑起来形成短语，然后再进行翻译，可以大大提高翻译的效率。这样就产生了基于短语的机器翻译（phrase-based statistical machine translation）。

基于短语的机器翻译[37-39]的流程为：首先，把训练语料中所有对齐短语以及它所对应的翻译概率作为翻译词典；其次，将预翻译的句子进行短语切分，找出最合理的切分方法，之后进行词典匹配；最后，将获得的目标短语进行重排序。由上述过程可以看出基于短语的机器翻译要解决三个问题：短语划分，短语的调序，短语翻译。研究者们分别对以上问题进行了研究，并使基于短语的机器翻译在大规模语料库上取得了很好的效果。2004 年，Philipp Koehn[37]开发的法老系统对短语机器翻译影响很大，效果远远好于基于词的翻译模型，推出以后很快成为研究者们的基准。文献[38]提出了一种调序模型，传统的调序模型对处理紧邻的短语，而对于长距离调序效果不好，该方法缓解了短语调序中的长距离调序所带来的错误，使翻译效果有了明显改善。

（3） 基于句法的机器翻译

由于基于短语的机器翻译存在缺陷，近年来基于句法的机器翻译（syntax-based statistical machine translation）成为研究者们关注的热点。目前，部分基于句法的机器翻译效果已经超过了基于短语的机器翻译。该方法可以分为两类：基于形式化语法的统计机器翻译和基于语言学句法的统计机器翻译。

基于形式化语法的统计机器翻译模型,只用到某种形式化的语法体系,但是该语法中并不包含任何语言学知识,如一些语言学标记和关系等[40]。该模型包括ITG模型[41]和层次短语模型[42]。其中，ITG模型是由吴德恺提出的，其本质是上下文无关文法（SCFG）的一种简化。ITG模型有两种假设，语言的调序只存在两种可能性，逆序和保序。该模型理论简单，所以在当时被广泛的应用。另一种方法为层次短语模型，它是由Chiang提出的。该模型定义了两条“glue”规则，作用在于可以对源语言短语切分再顺序合并它们所对应的译文。

基于语言学语法的统计机器翻译在翻译的过程中加入了丰富的语言学知识，该方法也包含了两类:一类是基于短语结构树的[43]，另一类是基于依存树的[44-45]。基于短语结构树的模型又可以分为树-串，串-树和树-树三类；在翻译过程中选择哪种方法还要根据实际问题决定。目前，树-串和串-树都有比较成熟的翻译系统，而树-树的翻译还处于探索阶段。基于依存树的方法逐渐引起了人们的注意，短语结构树相比，依存树是词汇化的而且依存语法本身体现了一种语义上的关系。

2.3.2 基于语言学特征的翻译方法

对术语的翻译主要有两种方法：基于统计的术语翻译方法[46-47]和基于网络的术语翻译方法。基于网络的方法利用网络双语资源较丰富的特点来解决一部分翻译中未登录词问题，但同时该方法忽略了术语内部词与词之间存在的关系。基于统计的术语翻译一般采用噪声信道模型，对数线性模型等。

1 基于统计的术语翻译

基于统计的术语翻译一般采用机器翻译方法的同时再加入术语自身的特点，比如：术语领域特征和术语组成结构特征等。文献[46]利用术语的形态学信息进行翻译，将术语进行进一步切分，例如：“superconducting quantum”将被切分成“super|conduct|ing|quantum”，减小术语的粒度，在一定程度上解决了数据稀疏问题。文献[47]在日汉术语翻译的过程中加入了领域信息，日汉字形信息和相对条件熵信息，之后采用对数线性模

型将特征融合并赋予不同的权重，并取得了较好的效果。

2 基于网络的术语翻译

随着科技的发展，网络已经成为人们获取知识的重要渠道。文献[48]提出一种基于语义预测的汉英术语方法。该方法在语义预测的过程中，采用网络作为资源，通过网页处理和术语翻译挖掘来预测术语的翻译；最后再对译文进行排序，得到较好的翻译结果。基于网络的翻译方法相对于传统的方法在新词翻译上取得了较好的效果。

2.4 本章小结

本章首先介绍术语的特点，之后分别介绍了依存分析，语义分析，机器翻译以及术语翻译的研究现状，并通过分析介绍了各种机器翻译方法的优缺点。

第 3 章　基于统计和规则相结合的汉语术语语义分析方法

本章主要针对汉语术语中的多词术语进行语义分析。由于术语自身的特点导致对术语的语义分析有以下难点：首先，术语中存在着大量的未登录词，导致现有的工具对术语分词和词性标注效果比通用领域句子的效果要差，这也直接影响到下一步的语义分析效果；其次，术语用词的重复性低，所以在训练语料有限的情况下，数据稀疏问题比较严重；再次，目前没有标注语义关系的熟语料，这也给术语分析带来了一定的困难。针对上述分析本章提出了适合术语的语义分析方法。在依存分析阶段，选用基本特征，互信息和词语在知网中的第一义原作为支持向量机（SVM）的特征，训练依存分析模型。这种方法的优点是加入知网中的第一义原特征从语义层次上进行判别，有效缓解了数据稀疏问题。在语义分析阶段，我们定义了 14 种语义关系，两种句法层次的依存关系。并利用已定义的规则对分词错误处理，在基于条件随机场（CRF）的语义分析模型基础上加入了 SVM 后处理模型，对 CRF 输出的不可靠语义关系进行校正。实验表明，利用基于统计和规则相结合的方法对术语进行语义分析是有效的。

3.1　基于 SVM 的依存分析方法

传统的统计方法只有在样本趋向无穷大时，其性能才有理论的保证。但是实际应用中样本总是有限的。而支持向量机方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折中，其推广能力明显优于一些传统的学习方法。

本文利用 SVM 的分类能力，将词对之间的关系分成两种类别。存在依存关系的词对作为正例，不存在依存关系的词对作为负例。一个词数为 N 的术语，正确的依存关系为 N-1 个。而“非依存关系”的词对确远远超过了这个数目。如果用穷举法把所有的负例都考虑进来，会出现偏置问题。基于以上考虑，本文对负例进行处理，在所有负例中随机选择与正例个数相当的数目，再选择适合术语的特征，利用多项式核函数将线性不可分问题转化成线性可分问题。

3.1.1 SVM 模型

支持向量机是寻求最好的分类面，即找到一个平面使得 $dis=\frac{2}{\|w\|}$ 最大。使训练样本集$(x_1, y_1), (x_2, y_2)\ldots(x_n, y_n)$分到这个超平面两侧。这样就将求最佳超平面问题转化成为二次规划问题，采用拉格朗日乘子法归结为如公式（3.1）所示。

$$Q(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_i y_i y_j x_i^T x_j \tag{3.1}$$

其中：

$$Q(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_i y_i y_j x_i^T x_j \tag{3.2}$$

$$\sum_{i=1}^{l}y_i\alpha_i=0,\alpha_i\geq 0 \tag{3.3}$$

对于线性不可分问题。SVM 利用核函数将线性不可分问题转化成线性可分问题。核函数包括四种，线性核函数，多项式核函数，径向基核函数和 S 型核函数。

3.1.2 特征选择

本文通过分析术语特点，选择了以下三类特征：基本特征，互信息特征，词语在知网中的第一义原特征；其中基本特征在文献[8-9]中已经被应用，并取得了很好的效果。针对术语自身的特点，本文加入了互信息特征，针对数据稀疏问题加入了词语在知网中的第一义原特征。

1 基本特征

基本特征如表 1 所示，其中词，词性，上下文特征权重都是二值函数；距离信息权重共考虑四种情况，W_{dis} 表示为：

$$W_{dis}=\begin{cases}1 & \text{if (dis=1)}\\ 2 & \text{if(dis=2)}\\ 3 & \text{if(dis=3)}\\ 4 & \text{if(dis>3)}\end{cases} \tag{3.4}$$

3.1.1 SVM 模型

支持向量机是寻求最好的分类面，即找到一个平面使得 $dis=\frac{2}{\|w\|}$ 最大。使训练样本集$(x_1, y_1), (x_2, y_2)\ldots(x_n, y_n)$分到这个超平面两侧。这样就将求最佳超平面问题转化成为二次规划问题，采用拉格朗日乘子法归结为如公式（3.1）所示。

$$Q(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_i y_i y_j x_i^T x_j \tag{3.1}$$

其中：

$$Q(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_i y_i y_j x_i^T x_j \tag{3.2}$$

$$\sum_{i=1}^{l}y_i\alpha_i=0,\alpha_i\geq 0 \tag{3.3}$$

对于线性不可分问题。SVM 利用核函数将线性不可分问题转化成线性可分问题。核函数包括四种，线性核函数，多项式核函数，径向基核函数和 S 型核函数。

3.1.2 特征选择

本文通过分析术语特点，选择了以下三类特征：基本特征，互信息特征，词语在知网中的第一义原特征；其中基本特征在文献[8-9]中已经被应用，并取得了很好的效果。针对术语自身的特点，本文加入了互信息特征，针对数据稀疏问题加入了词语在知网中的第一义原特征。

1 基本特征

基本特征如表 1 所示，其中词，词性，上下文特征权重都是二值函数；距离信息权重共考虑四种情况，W_{dis} 表示为：

$$W_{dis}=\begin{cases}1 & \text{if (dis=1)}\\ 2 & \text{if(dis=2)}\\ 3 & \text{if(dis=3)}\\ 4 & \text{if(dis>3)}\end{cases} \tag{3.4}$$

表 3.1 依存分析的基本特征

特征	特征描述	取值范围
Word	两个预判断依存关系的词	（0，1）
Pos	两个预判断依存关系的词性	（0，1）
Distance	两个预判断的词在术语中的距离	（1，2，3，4）
Context	第一（二）个词的前一个词和后一个词的词性和词本身	（0，1）

2 互信息特征

互信息在信息论中是作为衡量两个信号关联程度的一种尺度[49]。用 $I(X,Y)$ 代表随机变量X和Y的互信息，公式如（3.5）所示， 互信息可以理解为Y的值透露了多少X的信息量。

$$I(X,Y)=log_2\frac{P(X,Y)}{P(X)\bullet P(Y)} \tag{3.5}$$

其中 $P(X,Y)$ 代表词语X，Y在训练语料中同现的概率； $P(X)$ ， $P(Y)$ 分别代表X，Y在语料中单独出现的概率。

本文将互信息作为组成术语的两个词之间的关联程度大小的量度，把术语中词的出现作为随机过程，在642908句术语的语料上统计每两个词的互信息。分别用X，Y单独出现的次数代替它们单独出现的概率；用它们同现的次数代替它们同现的概率。利用上述方法计算得到的互信息结果一般在 10^{-6} 的数量级上，为了使互信息特征权重和其他特征权重达到相同的数量级，本文将互信息扩大 10^6 倍，然后将互信息的权重设置为 w_{mi} ，公式如下：

$$w_{mi}=\begin{cases}0 & 0.25>10^6\times I(X,Y)\\ 0.5 & 0.25\le 10^6\times I(X,Y)\le 0.75\\ 1 & 0.75<10^6\times I(X,Y)\end{cases} \tag{3.6}$$

3 知网第一义原特征

（1）知网简介

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。知网中的“概念”是对词

汇语义的一种描述，“义原”是组成概念的基本单位。一个词语可以表达为几个概念；例如：“器”在知网中有多个概念,选择其中的两个。

NO.=089177	NO.=089178
W_C=器	W_C=器
G_C=N [qi4]	G_C=N [qi4]
E_C=	E_C=
W_E=ware	W_E= organ
G_E=N	G_E=N
E_E=	E_E=
DEF={implement|器具}	DEF= {part|部件:PartPosition={visc 脏},whole={AnimalHuman|动物}}

其中NO.表示词在知网中的编号，W_C代表这个词本身，G_C代表中文词性，E_C代表中文例子，W_E代表对应的英文解释，G_E代表英文词性，E_E代表英文例子，DEF代表这个词的定义。

知网中的每个概念都由2219个义原来描述，将这2219个义原分为事件、实体、属性、属性值等10个类别。每个类别都由树状结构表示，如图3.1给出了知网中实体的树状结构，其他几类和该结构类似。本文中提到的第一义原代表了这个词的语义类别，例如上个例子中“器”字属于两个语义类，分别为 “器具”和“部件”。在依存分析的过程中加入第一义原特征能够从语义层面上进行分析，提高了分析的准确率。

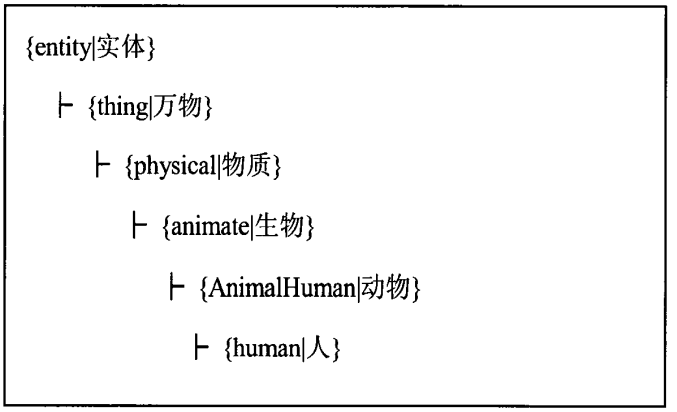


图3.1 知网中实体的树状结构

（2） 概念选择

由（1）节的介绍可知，一个词语在知网中有多个概念，所以在确定词的第一义原时首先要进行概念选择。用tree（t）表示具有n个词的术语t的依存关系树，如下所示。

$$\text{tree(t)}=\left(\langle w_1, \text{kn}(w_1)\rangle, \langle w_2, \text{kn}(w_2)\rangle \cdots < w_i, \text{kn}(w_i)> \cdots \langle w_n, \text{kn}(w_n)\rangle\right) \quad (3.7)$$

其中 w_i 表示t中第i个词，$\text{kn}(w_i)$ 表示 w_i 所依存的词。本文对tree(t)中的每一对依存关系做相关度计算，则依存关系对 $<w_i, \text{kn}(w_i)>$ 中 w_i，$\text{kn}(w_i)$ 的概念 I_{w_i}，$I_{kn(w_i)}$ 表示为：

$$I_{w_i}=\arg \max_{Q_{w_i}} \text{itemSem}(w_i, kn(w_i)) \quad (3.8)$$

$$I_{kn(w_i)}=\arg \max_{Q_{kn(w_i)}} \text{itemSem}(w_i, kn(w_i)) \quad (3.9)$$

其中 Q_{W_i}，$Q_{kn(w_i)}$ 分别表示 w_i，$\text{kn}(w_i)$ 所有的概念集合。因为 w_i 在术语t中可能不止一次出现，在 w_i 出现的词对中，可能每次所选的概念是不同的。设在术语t中 w_i 取得 I_{w_i} 的次数记为 $n(I_{w_i})$，则 w_i 最终所取的概念 $I_i=\arg \max_{Q_i} n(I_{w_i})$，其中 Q_i 表示 w_i 在术语t中求出的所有概念。$\text{itemSem}(w_i, kn(w_i))$ 为 w_i 与 $kn(w_i)$ 的语义相关度，由公式（3.10）求得。

语义相关度是在句法分析中一个短语结构中的两个词能够组成修饰关系、主谓关系、同指关系的程度[s]。本文将相关度计算分成两个层次，义原层次和概念层次。概念 I_1 和 I_2 之间的相关度 $\text{itemSem}(I_1, I_2)$ 表示为：

$$\text{itemSem}(I_1, I_2)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} baseSem(b_i, b_j)}{n \cdot m} \quad (3.10)$$

其中n，m分别为组成 I_1，I_2 义原的个数，$baseSem(b_i, b_j)$ 由刘群[13]提出的公式计算得到。

$$baseSem(b_i, b_j)=\frac{\alpha}{\alpha + dis(b_i, b_j)} \quad (3.11)$$

其中 $dis(b_i, b_j)$ 表示义原 b_i, b_j 的路径长度，α 是可调参数。

（3）知网中未出现词的处理

由于术语属于非通用领域，所以组成术语的词一部分在知网中是不存在的。通过对3000句术语的研究分析得出术语中心词是后置的，组成术语的词的中心字也是后置的。

因此采用逆向最大匹配算法对知网中不存在的词进行切分，直到知网中存在该词为止。具体算法如下：

输入：词语w;

输出：在知网中对应的概念 I_w

```
label:  If(search(w))//在知网中查到该词
            return 概念;//利用(3)中介绍的方法，返回w在知网中的概念
        else
        {
            w=w.substr(2);//截取词语w中除第一个字以外的词
            goto label;
        }
```

（4）特征提取

根据 3.13（2）节介绍的概念选择可知，在对测试语料提取在知网中的第一义原特征时，要先确定术语内部的依存关系。所以本文首先预测术语中的每个词可能与该术语哪个词构成依存关系，用 w_i 表示术语 t 的第 i 个词，$kn(w_i)$ 表示 w_i 可能依存的词，公式如（3.11）所示；$P(w_i,kn(w_i))$ 表示两个词的潜在依存强度，经分析得影响潜在依存强度的因素有：两个词在训练语料中的平均距离，在当前术语中的距离和两个词的互信息。

$$kn(w_i)=\arg\max_{kn\in w(t)} P(w_i,kn(w_i)) \tag{3.11}$$

其中 w(t)表示组成术语 t 的词的集合，$P(w_i,kn(w_i))$ 由公式（3.12）计算得出。

$$P(w_i,w_j)=(\lambda_1 I(w_i,w_j)+\lambda_2 dis^{-1}+\xi)/\sqrt{j-i} \tag{3.12}$$

其中 $\lambda_1+\lambda_2=1$，w_i，w_j 表示术语中第 i 和 j 个词，ξ 为平滑因子，dis 代表 w_i,w_j 在训练语料中的平均距离，λ_1 取 0.9，λ_2 取 0.1。之后根据前三步介绍的方法标注出测试语料术语中每个词的第一义原。

3.1.3 SVM 与规则结合的依存分析

在依存分析过程中，本文首先对术语进行规则处理，之后再对术语中除规则以外的

每两个词进行SVM模型处理。

规则 1：词性为后接词，依存于该词的前一个词。

规则 2：词性为介词且词为“用”，依存于该词的前一个词。

利用 SVM 模型对不满足上述两条规则的词对进行分析，返回一个实数，正值说明这两个词存在依存关系，且值越大说明两个词的依存强度越大。同理负值越大说明两个词不能构成依存关系的可能性越大。利用上述特点，本文对术语中的词从左到右依次进行判断。每次选择和当前词依存强度（即 SVM 返回的数值）最大的作为这个词的依存词。根据依存关系公理，依存关系不能交叉，所以需要进行回溯，直至依存关系中没有交叉为止。

3.2 基于 CRF 和 SVM 相结合的语义分析方法

3.2.1 CRF 模型

CRF是一个在给定输入节点条件下计算输出节点条件概率的无向图模型。它是在给定需要标记的观察序列条件下，计算整个标记序列的联合概率分布[52]。假设S表示整个状态序列，O表示整个观测序列，则P(Y|X)表示为 ：

$$\phi(y_{t,n},X)=\exp(\sum_{k}(\lambda_k f_k(y_{t-1},y_t)+\mu_k g_k(y_t,X))) \tag{3.13}$$

其中：

$$f_k(y_{t-1},y_t)=\begin{cases}1 & if\ y_{t-1}\ and\ y_t\ on\ condition\\ 0 & otherwise\end{cases} \tag{3.14}$$

$$g_k(y_t,X)=\begin{cases}1 & if\ X\ and\ y_t\ on\ condition\\ 0 & otherwise\end{cases} \tag{3.15}$$

CRF具有很强的推理能力，并且能够使用复杂、有重叠性和非独立的特征进行训练和推理，能够充分地利用上下文信息作为特征，还可以任意地添加其他外部特征，使得模型能够获取的信息非常丰富。同时，CRF解决了最大熵模型中的数据偏置问题。

3.2.2 语义关系定义

目前各个知识库都涉及到了语义类和语义关系。其中 MindNet 定义了 24 种语义关系，包括属性，目标，原因，方式，部分等。这些语义关系是句法分析器对词典中的释义文本进行分析得到的，所以可靠性较高。Wordnet 分别对名词，动词，形容词，副词

每两个词进行SVM模型处理。

规则 1：词性为后接词，依存于该词的前一个词。

规则 2：词性为介词且词为“用”，依存于该词的前一个词。

利用 SVM 模型对不满足上述两条规则的词对进行分析，返回一个实数，正值说明这两个词存在依存关系，且值越大说明两个词的依存强度越大。同理负值越大说明两个词不能构成依存关系的可能性越大。利用上述特点，本文对术语中的词从左到右依次进行判断。每次选择和当前词依存强度（即 SVM 返回的数值）最大的作为这个词的依存词。根据依存关系公理，依存关系不能交叉，所以需要进行回溯，直至依存关系中没有交叉为止。

3.2 基于 CRF 和 SVM 相结合的语义分析方法

3.2.1 CRF 模型

CRF是一个在给定输入节点条件下计算输出节点条件概率的无向图模型。它是在给定需要标记的观察序列条件下，计算整个标记序列的联合概率分布[52]。假设S表示整个状态序列，O表示整个观测序列，则P(Y|X)表示为 ：

$$\phi(y_{t,n},X)=\exp(\sum_{k}(\lambda_k f_k(y_{t-1},y_t)+\mu_k g_k(y_t,X))) \tag{3.13}$$

其中：

$$f_k(y_{t-1},y_t)=\begin{cases}1 & if\ y_{t-1}\ and\ y_t\ on\ condition\\ 0 & otherwise\end{cases} \tag{3.14}$$

$$g_k(y_t,X)=\begin{cases}1 & if\ X\ and\ y_t\ on\ condition\\ 0 & otherwise\end{cases} \tag{3.15}$$

CRF具有很强的推理能力，并且能够使用复杂、有重叠性和非独立的特征进行训练和推理，能够充分地利用上下文信息作为特征，还可以任意地添加其他外部特征，使得模型能够获取的信息非常丰富。同时，CRF解决了最大熵模型中的数据偏置问题。

3.2.2 语义关系定义

目前各个知识库都涉及到了语义类和语义关系。其中 MindNet 定义了 24 种语义关系，包括属性，目标，原因，方式，部分等。这些语义关系是句法分析器对词典中的释义文本进行分析得到的，所以可靠性较高。Wordnet 分别对名词，动词，形容词，副词

进行了语义关系定义，但是 Wordnet 并没有对不同词类之间进行关系定义。以上两种是英文知识库，在汉语方面，知网定义了 16 种义原关系和 76 种动态角色。清华大学从知网中定义的动态角色中归纳总结了 59 种语义关系并实现了一个语义分析系统，并将系统应用到了语音识别，使得效果有所提高。

本文根据已有的研究和术语的特点定义了 14 种语义关系和两种句法层次的依存关系，其中一部分语义关系来自于知网对关系的定义；另一部分是结合术语自身的特点定义的。两种句法层次的依存关系分别为：“之”字结构和“与”字结构；14 种语义关系及其解释如表 3.2 所示

为了更好地理解术语的语义关系，本文又将 14 种关系中的“属性-宿主”关系进行了进一步划分，分成了 7 个小的语义关系，包括量度属性、外观属性、状况属性、特性、数量属性、类别属性、功能属性。

表 3.2 语义关系分类及解释

语义类别	解释
施事-事件	表示行动的事件类型中“变关系”，“ 变状态”，“ 变属性”，“ 使之动”四类事件中的充当“变”这一功能的实体与事件的关系
受事-事件	表示行动的事件类型中“变关系”，“ 变状态”，“ 变属性”，“ 使之动”四类事件中的充当“被改变”这一功能的实体与事件的关系
属性-宿主	表示宿主本身所具有的性质
材料-成品	事件发生或进行所依赖的材料
整体-部分	表示“蕴涵关系”或表示状态事件中“残疾”类事件中的实体的部件
后接成分	没有实际含义的后缀
方式	事件发生或进行的方式
用途	实体应用的方面或范围。与功能属性的区别是，用途强调应用的范围，功能强调实体发挥的有利作用
否定	带有否定词的修饰
名称	对抽象名词的描述
关系	术语中涉及的关系主要是，方位关系和方向关系
接续	两个以上的事件接连发生或进行，同时它们是密切相关的
程度	事件或属性值的程度
对象	行动或思考时作为目标的事物

3.2.3 基于规则的分词错误处理

本文采用的分词工具是面向通用领域的，应用到专业术语上存在分词错误。这样给语义分析带来困难。例如“多光谱组合滤光片”被拆分成分成“多 光 谱 组合 滤 光

片”；针对这种错误，本文制定以下 4 条规则对分词结果进行处理。由于组成术语的词大部分都是双字词和三字词，而很少出现单字词，所以本文假设分词结果中的单字都是由分词错误造成的，例如上例中的“多光谱”和“滤光片”。合并后对合并词之间的语义关系不予以考虑。

规则 1：5 个连续的单字词，本文将它分解成 2+3 或者 3+2 结构。如果第三个单字词为名词选择 3+2 结构，其他情况选择 2+3 结构。例如“双/m 极/n 性/n 电/n 极/n 铅酸/n 蓄电池/n”合并为：“双极性/n 电极/n 铅酸/n 蓄电池/n”。而“山/n 楂/v 去/v 核/n 器/n”合并为：“山楂/v 去核器/n”。

规则 2：4 个连续的单字词，合并成 2+2 形式。

规则 3：3 个连续单字词，合并为 2+1 或者将三个词合并为一个词。如果第三个词是名词则将三个词合并，其他情况合并为 2+1 的形式。

规则 4：2 个连续的单字词。将两个词合并成为一个词。

3.2.4 基于 CRF 的语义分析

条件随机场（CRF）是一个在给定输入节点条件下，计算输出节点条件概率的无向图模型。它是在给定需要标记的观察序列条件下，计算整个标记序列的联合概率分布[52]。本文选择 CRF 作为语义分析的统计模型，能避免在其他模型中产生的数据偏置问题，符合语义分析的需求。选择表 3.3 所示的原子特征，其中确定 wd_1_atom 与 wd_2_atom 的方法与前面介绍的相同。

表 3 .3 CRF 模型的原子特征

特征	描述	特征	描述	特征	描述
wd_1	修饰词	pos_2	被修饰词词性	pre_wd	修饰修饰词的词
wd_2	被修饰词	wd_1_atom_	修饰词第一义原	pre_pos	修饰修饰词的词性
pos_1	修饰词词性	wd_2_atom	被修饰词第一义原	kn/unkn	被修饰词是否为中心词

3.2.5 结果后处理

和句子之间的语义关系相比，术语所涉及的语义关系范围比较窄，有些语义关系间界限不明确导致 CRF 的分类能力差，为此本文对 CRF 的输出结果进行后处理。对于易混淆的类别采用 SVM 训练二分类器，对 CRF 模型输出的 2-best 结果中的两个语义关系

进行识别，确定词对最终语义关系。进行后处理的过程要满足以下两个条件：

（1） CRF 语义分析模型的打分值低于所规定的阈值；

（2） CRF 语义分析模型输出的 2-best 结果中的两个语义关系是易混淆的类别；

表 3.4 后处理过程特征

特征	特征描述	取值范围
word	两个预判断语义关系的词	（0，1）
sememe	两个预判断语义关系的词在知网中的第一义原	（0，1）
kn/nokn	被修饰词是否为术语中心词	（0，1）

本文对三对易混淆的语义关系进行训练：（1）“功能属性”与“特性”，（2）“外观属性”与“特性”，（3）“施事-事件”与“受事-事件”，三个分类器均选用表 3.4 所示特征。和 CRF 模型的特征相比，我们删除了词性特征，因为词性特征对这三类分类器的区分度不大，例如“事件施事”与“事件受事”一般都由“动词+名词”结构组成，所以加入词性信息反而使正确率下降。

3.3 实验结果与分析

3.3.1 实验设置

实验采用专利文献中的术语作为训练和测试语料。该语料共有 642908 调术语，本文选取其中的 3000 条作为训练语料，238 条作为测试语料，术语平均长度为 5.07 词/条,内容包括人类生活必需类、作业、运输类、化学、冶金类等。分词采用哈尔滨工业大学的分词工具 IRLAS[50]。

3.3.2 依存分析结果

本文采用两种评价方法，词对准确率（wordPre）和句子准确率(sentPre)。两个评价方法定义如公式（3.16）和（3.17）所示：

$$word\,\mathrm{Pr}e=\frac{\text{标注正确的依存关系个数}}{\text{依存关系总个数}}\times 100\% \tag{3.16}$$

$$sent\,\mathrm{Pr}e=\frac{\text{标注正确的术语个数}}{\text{术语的总数}}\times 100\% \tag{3.17}$$

为了更好地体现系统性能，本文选用两个对比系统：

进行识别，确定词对最终语义关系。进行后处理的过程要满足以下两个条件：

（1） CRF 语义分析模型的打分值低于所规定的阈值；

（2） CRF 语义分析模型输出的 2-best 结果中的两个语义关系是易混淆的类别；

表 3.4 后处理过程特征

特征	特征描述	取值范围
word	两个预判断语义关系的词	（0，1）
sememe	两个预判断语义关系的词在知网中的第一义原	（0，1）
kn/nokn	被修饰词是否为术语中心词	（0，1）

本文对三对易混淆的语义关系进行训练：（1）“功能属性”与“特性”，（2）“外观属性”与“特性”，（3）“施事-事件”与“受事-事件”，三个分类器均选用表 3.4 所示特征。和 CRF 模型的特征相比，我们删除了词性特征，因为词性特征对这三类分类器的区分度不大，例如“事件施事”与“事件受事”一般都由“动词+名词”结构组成，所以加入词性信息反而使正确率下降。

3.3 实验结果与分析

3.3.1 实验设置

实验采用专利文献中的术语作为训练和测试语料。该语料共有 642908 调术语，本文选取其中的 3000 条作为训练语料，238 条作为测试语料，术语平均长度为 5.07 词/条,内容包括人类生活必需类、作业、运输类、化学、冶金类等。分词采用哈尔滨工业大学的分词工具 IRLAS[50]。

3.3.2 依存分析结果

本文采用两种评价方法，词对准确率（wordPre）和句子准确率(sentPre)。两个评价方法定义如公式（3.16）和（3.17）所示：

$$word\,\mathrm{Pre} = \frac{\text{标注正确的依存关系个数}}{\text{依存关系总个数}} \times 100\% \tag{3.16}$$

$$sent\,\mathrm{Pre} = \frac{\text{标注正确的术语个数}}{\text{术语的总数}} \times 100\% \tag{3.17}$$

为了更好地体现系统性能，本文选用两个对比系统：

（1）规定所有的词都依存于术语中心词（baseline1）。

（2）规定所有词只依存于它右侧的紧邻词 (baseline2)。

表 3.5 给出了对比系统和本系统在基本特征（system1）,基本特征+互信息特征（system2）,基本特征+互信息特征+知网第一义原特征（system3）下词对准确率和句子准确率。

表 3.5 术语依存分析实验结果

方法	词对准确率	句子准确率
baseline1	33.67%	17.84%
baseline2	71.78%	50.41%
system1	81.36%	51.51%
system2	82.55%	55.89%
system3	**87.85%**	**65.65%**

实验结果表明，本系统的效果要高于两个对比系统，这说明本文所选用特征是有效的。baseline2 比 baseline1 提高 38.11 个百分点，由此可知依存于紧邻词的概率要远远高于依存于中心词的概率，所以两个词的距离信息对依存关系有很大的影响。

在本系统的三个实验中，加入互信息后词对准确率和句子准确率比 system1 词对准确率和句子准确率分别提高了 1.19 个百分点和 4.38 个百分点，加入词语在知网中第一义原特征后词对准确率和句子准确率分别比 system2 提高了 5.30 和 9.76 个百分点。这是因为互信息虽然可以判定出两个词的关联强度，但是在术语中存在大量未登陆词，所以难免出现数据稀疏问题，导致加入互信息特征提高并不明显。而加入知网第一义原特征后有了明显提高，这说明两个词是否能够构成依存关系很大程度上取决于它们所表达的语义。

训练语料库规模会直接影响到 SVM 的分类效果，为此本文统计了训练语料库规模和依存关系正确率之间的关系。如图 3.2 所示。由图中数据可以看出，当语料规模达到 2500 条时，再增加训练语料，对结果的影响并不是很大，且如果训练语料过多会影响到训练的速度，所以综合考虑本文选用 3000 条术语作为训练语料。

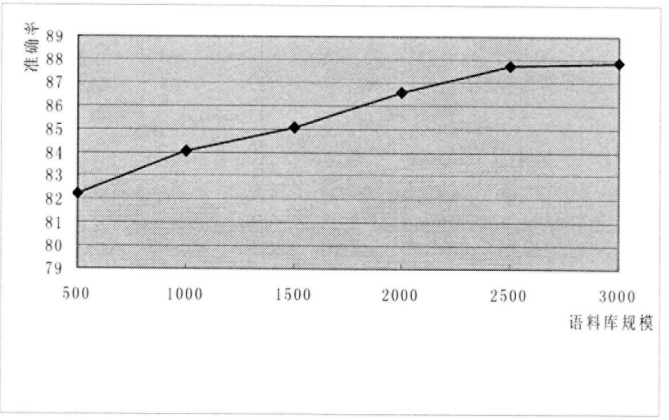


图 3.2 语料库规模与词对准确率关系图

3.3.3 语义分析结果

我们将定义的 14 种语义关系称为“大类语义关系”，将“属性-宿主”细分后的 20 种语义关系为“小类语义关系”。用语义关系正确率对结果进行评价，评价公式如下：

$$wordsem = \frac{\text{正确标出依存关系和语义关系的个数}}{\text{依存关系总个数}} \times 100\% \qquad (3.18)$$

表 3.6 给出语义分析的实验结果。表 3.7 给出在依存分析正确情况下的语义分析实验结果。

表 3.6 术语语义分析实验结果

方法	大类语义关系	小类语义关系
CRF	76.98%	68.31%
CRF+SVM	**77.13%**	**69.05%**

表 3.7 依存分析正确情况下的语义分析实验结果

方法	大类语义关系	小类语义关系
CRF	89.25%	79.20%
CRF+SVM	**89.42%**	**80.06%**

从表 3.7 可以看出如果依存关系完全正确，利用本文的语义分析方法在大类语义关系上可以达到 89.42%的正确率，在小类语义关系上可以达到 80.06%的正确率。而在依存分析的基础上进行语义分析，由于依存分析的效果在 87%左右，所以导致语义分析的效果有所下降。从表 3.6 可以得出，在小类语义关系上效果不好，主要是由于小类语义关系划分比较细致，导致各个类别间差异不明显。从结果可以看出，在对 CRF 模型后处理后在大的语义关系上提高了 0.15 个百分点，在小类语义关系上提高了 0.74 个百分

点。由此可见加入后处理模型对本系统是有效的，但是提高并不明显。

3.3.4 错误实例分析

1 依存分析错误实例分析

为了找出影响依存分析效果的因素，本文做了对比实验 3（system4）：该实验是在训练语料与测试语料分词和词性标注完全正确的情况下进行的，并且训练语料，测试语料，所选特征和 system3 完全相同。实验结果如表 3.9 所示。

表 3.9 分词和词性标注正确时的依存分析结果

方法	词对准确率	句子准确率
system3	87.85%	65.65%
system4	**89.62%**	**69.24%**

从实验结果可以看出，在分词和词性标注完全正确的情况下，词对准确率和句子准确率比 system3 提高了 1.77 和 3.59 个百分点。由 system4 的结果可以看出，分词的好坏直接影响到了依存分析的结果。除此之外，术语中难于理解的词也影响到了依存分析效果。所以依存分析的错误主要由以下几个方面造成。

（1）分词错误

依存分析是在分词的基础上进行的，所以分词的正确与否直接影响到依存分析的好坏。分词错误分为两种。

未登录词问题：例如：“可/v 控/v 气/n 泵/n”与“风力/n 清/a 雪/n 机/n”中的“气泵”和“清雪”；由于分词错误导致最终依存分析错误，图 3.3 给出了两个术语的错误分析结果，图 3.4 给出了正确的依存分析结果。

图 3.3 错误依存分析结果

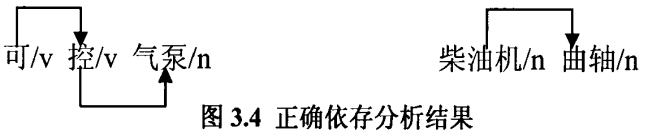


图 3.4 正确依存分析结果

词边界问题：例如“光导诊查器”的分词结果为“光 导诊 查 器”。这种分词错在总的分词错误中占有很高的比重。

（2）术语中存在大量难于理解的词

由于术语属于非通用领域，所以难免会出现一些难于理解的词，要结合专利的上下文才能识别出术语内部之间的修饰关系。例如：“电动机/n 防尘/v 轴承/n 组件/n”，“包装用/b 瓦楞纸/n 板/n 缓冲/v 装置/n”等。

2 语义分析错误实例分析

在语义分析阶段，本文分别统计了 20 种语义关系在测试语料中出现的次数的正确率，如表 3.8 所示。从表 3.8 的实验结果发现：“材料”，“功能属性”，“数量属性”和“后接成分”的正确率在 90%以上，而“外观属性”，“施事-事件”，“对象”的正确率在 40%以下。导致这三类关系正确率低的原因主要有以下两点。首先，术语所涉及的语义关系范围比较窄，有些语义关系间界限不明确，比如“外观属性”和“特性”，虽然我们对这两类错误进行后处理，但是在后处理过程中训练语料不充分，导致数据稀疏问题严重；其次，两个语义类别之间结构相似，在分类过程中容易混淆，比如：“施事-事件”和“受事-事件”，一般都是由“动词+名词”组成。

表 3.8 各个类别语义分析正确率

语义类别	在测试语料中出现次数	标注正确率
否定	1	1
材料	4	1
方式	16	0.4375
数量属性	15	0.9333
接续	6	0.6666
程度	7	0.8571
施事事件	11	0.3636
用途	8	0.875
功能属性	138	0.9057
对象	6	0.3333
整体部分	38	0.8157
名称	11	0.8181
关系	4	0.75
特性	125	0.8
类别属性	41	0.7561
状况	2	1
量度属性	16	0.6875
受事事件	70	0.8714
外观属性	25	0.32
后接成分	33	0.9697

3.4 本章小结

本章主要介绍了术语语义分析的方法，在依存分析阶段引入基本特征、互信息特征和知网第一义原特征，从语义层次上解决依存分析问题；并利用已定义规则对分词错误处理；在语义分析阶段对语义分析模型输出的结果进行后处理，使得大类语义关系和小类语义关系的准确率都有所提高。

第 4 章 基于语义分析的术语翻译方法

术语翻译作为机器翻译的一部分已经逐渐被人们所重视。本文针对结构复杂的名词术语进行翻译，该过程包含三个部分：短语抽取过程，源语言调序过程和解码过程。在短语抽取阶段，加入了在句法层次上具有依存关系的词串，弥补了传统方法中将连续的词串作为短语的缺陷。在调序阶段，将术语语义分析结果和词对齐结果相结合，从训练语料中提取调序模板，并在解码之前对源语言进行模板匹配，如果存在相应的调序模板，则按照调序模板中的调序方法对源语言调序。最后，在解码阶段，采用基于短语的翻译工具摩西进行解码。

4.1 短语抽取

在短语抽取之前，首先对源语言术语进行依存分析。图 4.1 表示出了“圆形苍蝇捕灭机”的依存树的结构。短语抽取过程分成两个部分：非结构化短语抽取和结构化短语抽取。

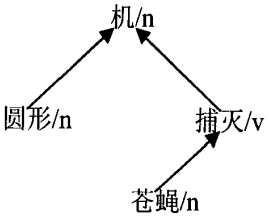


图 4.1 术语依存分析结果

4.1.1 非结构化短语抽取

非结构化短语抽取首先采用词对齐工具 GIZA++进行双向词对齐，生成对齐矩阵的形式，并取得双向对齐的交集和并集。之后采用 GROW-DIAG-FINAL 方法扩展词对齐结果。首先提取双向对齐结果的交集和并集，并以交集作为中心点，检查其上下左右和对角相邻的 8 个点，若在并集中，则作为扩展的对齐点加入到对齐序列中[53]。利用这种扩展的方法抽出的短语是非结构化的，只是形式上连续的词串，因此会漏掉一部分短语。例如：上例中的“圆形苍蝇捕灭机”，如果利用非结构化的短语抽取方法抽取就会漏掉“圆形机”这个短语。基于上述情况本文在短语抽取过程中使用了非结构化和结构化相结合的短语抽取方法。

第 4 章 基于语义分析的术语翻译方法

术语翻译作为机器翻译的一部分已经逐渐被人们所重视。本文针对结构复杂的名词术语进行翻译，该过程包含三个部分：短语抽取过程，源语言调序过程和解码过程。在短语抽取阶段，加入了在句法层次上具有依存关系的词串，弥补了传统方法中将连续的词串作为短语的缺陷。在调序阶段，将术语语义分析结果和词对齐结果相结合，从训练语料中提取调序模板，并在解码之前对源语言进行模板匹配，如果存在相应的调序模板，则按照调序模板中的调序方法对源语言调序。最后，在解码阶段，采用基于短语的翻译工具摩西进行解码。

4.1 短语抽取

在短语抽取之前，首先对源语言术语进行依存分析。图 4.1 表示出了“圆形苍蝇捕灭机”的依存树的结构。短语抽取过程分成两个部分：非结构化短语抽取和结构化短语抽取。

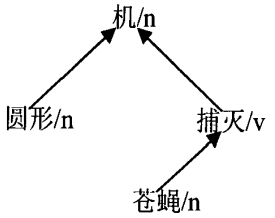


图 4.1 术语依存分析结果

4.1.1 非结构化短语抽取

非结构化短语抽取首先采用词对齐工具 GIZA++进行双向词对齐，生成对齐矩阵的形式，并取得双向对齐的交集和并集。之后采用 GROW-DIAG-FINAL 方法扩展词对齐结果。首先提取双向对齐结果的交集和并集，并以交集作为中心点，检查其上下左右和对角相邻的 8 个点，若在并集中，则作为扩展的对齐点加入到对齐序列中[53]。利用这种扩展的方法抽出的短语是非结构化的，只是形式上连续的词串，因此会漏掉一部分短语。例如：上例中的“圆形苍蝇捕灭机”，如果利用非结构化的短语抽取方法抽取就会漏掉“圆形机”这个短语。基于上述情况本文在短语抽取过程中使用了非结构化和结构化相结合的短语抽取方法。

4.1.2 结构化短语抽取

在结构化短语抽取过程中，将依存分析的结果看成有向图的结构，之后提取出有向图中任意的连通子图，且短语长度限制在 7 个词语之内。最后利用对齐信息提取出短语所对应的目标串。图 4.4 给出了“圆形苍蝇捕灭机”中的所有结构化短语。

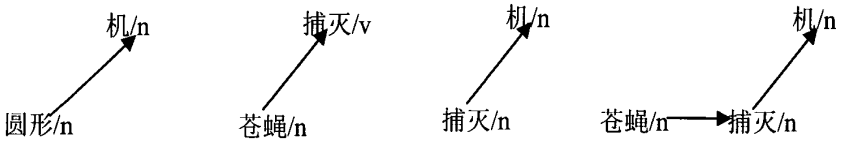


图 4.4 结构化短语实例

1 短语翻译对获取

根据已提取的短语和 GIZA++的对齐信息获得短语翻译对。为了提高翻译的准确率，在翻译对获取的过程中要保证源语言和目标语言之间严格满足对齐一致性，即翻译短语对<Ch,En>满足：

$$\forall\ (a^{'}, b^{'}) \in T: b \le b^{'} \le b + m \leftrightarrow i \le i^{'} \le i + m \tag{4.1}$$

其中 T 表示翻译短语对<Ch,En>的对齐矩阵。公式 4.1 表示，源语言中组成短语的词其对应的翻译必须在翻译短语对的目标语言中，反之，目标语言中的词语其翻译必须在翻译短语对的源语言中。

2 概率计算

在获得短语翻译对后，我们对短语互译的概率进行计算，包括源语言到目标语言的词对齐概率，短语对齐概率和目标语言到源语言的词对齐概率，短语对齐概率。公式 4.2 给出源语言到目标语言词对齐概率计算方法，其中 T 表示目标语言短语，S 表示源语言短语，t,s 分别表示目标语言和源语言的词。公式 4.3 给出源语言到目标语言短语对齐概率的计算方法，其中 $count(S \to T)$ 表示在语料库中源语言 S 翻译成 T 的次数，$count(S)$ 表示源语言 S 在语料库中的总次数。目标语言到源语言的计算方法和上述方法相同。

$$P_{\mathrm{w}}(T/S) = \prod_{t \in T, s \in S} p(t/s) \tag{4.2}$$

$$P_p(T/S)=\frac{count(S\rightarrow T)}{count(S)} \tag{4.3}$$

4.2 源语言调序

汉英术语翻译系统中，语序上的不同一直是翻译效果不好的主要原因。汉语中修饰词一般是在被修饰词的前面出现，而英语中定语通常是后置的。例如“联机 付费 方法 与 系统”对应的英文翻译为“Method and system for online payments”。针对这一问题本文在翻译过程中对源语言进行调序，使其更加符合目标语言的语序。

对于有 N 个词的句子，存在 N！种调序方式，其中有很多种调序方式是不符合语言现象的，所以目前研究者们利用各种方法来对调序进行剪枝，从而能够选择既符合源语言特点又符合目标语言特点的调序方式。机器翻译的调序方法一般都是在句法分析或者依存分析的基础上进行的。这种方法的优点是实现了长距离调序，而且不仅仅是词与词之间，短语和短语之间也可以调序，缺点是只从结构上对句子进行考虑，而没能理解组成句子的词之间的语义关系。

本文对 60 万的训练语料自动抽取了调序模板库。在测试阶段，首先对中文术语进行语义分析得到相应的语义结构，之后在模板库中查找是否存在该语义结构，如果模板库中存在则对术语进行相应的调序。

4.2.1 调序模板

1 模板抽取

调序模板抽取是在 60 万的双语平行语料上进行的，具体分为以下几个步骤：首先根据第三章介绍的方法对汉语术语语义分析，得到每个术语对应的语义分析树；其次，将语义分析树转化成语义结构树；最后，利用 GIZA++对语料进行词对齐，选择双向对齐的结果，并抽取调序模板。图 4.5 给出了“多功能电动节能三轮车”的语义分析树和语义结构树。从图 4.5 可以看出，语义分析树是完全词汇化的，语义结构树是将语义分析树的完全词汇化的节点转化成为该节点在术语中的位置。通过对大量术语的研究发现，术语中很多存在相同语义结构的术语具有相同的调序结果。所以本文将语义分析树转化为语义结构树进行模板抽取。

$$P_p(T/S)=\frac{count(S\rightarrow T)}{count(S)} \tag{4.3}$$

4.2 源语言调序

汉英术语翻译系统中，语序上的不同一直是翻译效果不好的主要原因。汉语中修饰词一般是在被修饰词的前面出现，而英语中定语通常是后置的。例如“联机 付费 方法 与 系统”对应的英文翻译为“Method and system for online payments”。针对这一问题本文在翻译过程中对源语言进行调序，使其更加符合目标语言的语序。

对于有 N 个词的句子，存在 N！种调序方式，其中有很多种调序方式是不符合语言现象的，所以目前研究者们利用各种方法来对调序进行剪枝，从而能够选择既符合源语言特点又符合目标语言特点的调序方式。机器翻译的调序方法一般都是在句法分析或者依存分析的基础上进行的。这种方法的优点是实现了长距离调序，而且不仅仅是词与词之间，短语和短语之间也可以调序，缺点是只从结构上对句子进行考虑，而没能理解组成句子的词之间的语义关系。

本文对 60 万的训练语料自动抽取了调序模板库。在测试阶段，首先对中文术语进行语义分析得到相应的语义结构，之后在模板库中查找是否存在该语义结构，如果模板库中存在则对术语进行相应的调序。

4.2.1 调序模板

1 模板抽取

调序模板抽取是在 60 万的双语平行语料上进行的，具体分为以下几个步骤：首先根据第三章介绍的方法对汉语术语语义分析，得到每个术语对应的语义分析树；其次，将语义分析树转化成语义结构树；最后，利用 GIZA++对语料进行词对齐，选择双向对齐的结果，并抽取调序模板。图 4.5 给出了“多功能电动节能三轮车”的语义分析树和语义结构树。从图 4.5 可以看出，语义分析树是完全词汇化的，语义结构树是将语义分析树的完全词汇化的节点转化成为该节点在术语中的位置。通过对大量术语的研究发现，术语中很多存在相同语义结构的术语具有相同的调序结果。所以本文将语义分析树转化为语义结构树进行模板抽取。

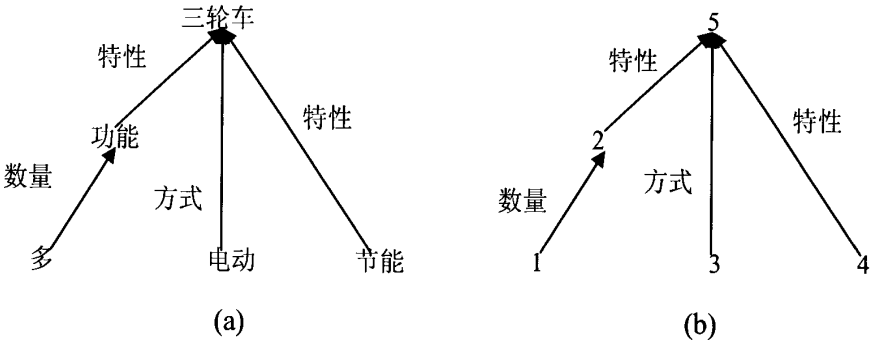


图 4.5 语义分析树与语义结构树

本文利用三元组来定义调序模板<s,i,p>，其中 s 表示中文术语的语义结构树，i 表示中文术语调序信息，p 为语义结构树 s 对应调序信息 i 的概率，公式如（4.4）所示。

$$p=\frac{c(s,i)}{c(s)} \tag{4.4}$$

其中 $c(s,i)$ 表示语料库中语义结构 s 对应调序信息 i 所出现的次数，$c(s)$ 表示语义结构 s 在语料库中出现的总次数。

抽取的调序模板形式为：(1 2)||特性 (2 3)||事件受事 (3 6)||特性 (4 5)||方式 (5 6)||功能 |||4 5 6 1 2 3 ||| 0.685321。第一项中的“(1 2)”表示术语中第一个词是修饰第二个词的，“特性”表示第一个词和第二个词之间的修饰关系为“特性”关系；第二项表示调序信息，第三项表示该语义结构对应这种调序的概率。

2 调序错误处理

由于词对齐工具和语义分析都会产生错误信息，所以导致调序模板中存在大量的调序错误。对于这类错误本文采用互信息和依存分析结合的方法进行排除。定义一条规则：如果一个术语中的两个词相邻词之间依存关系且互信息大于设定的阈值，那么在调序过程中，这两个词之间是不能被插入其他词的。例如：“多功能卷笔刀”的依存分析结果为（1 2）‖（2 5）‖（3 4）‖（4 5），“多”和“功能”之间存在依存关系，且它们的互信息大于规定的阈值。根据上述方法得到调序结果为：2 3 4 5 1，我们认为这种调序结果为错误的，在调序模板抽取的过程中被过滤掉。

4.2.2 术语调序

在调序阶段，首先对中文术语语义分析，得到该术语的语义结构，在根据已有的调

序模板进行调序。表 4.1 给出了一个语义结构对应的不同调序结果，从所给数据可以看出，调序概率最大的结果最符合目标语言语序。

表 4.1 一种语义结构的不同调序方式示例

源语言	目标语言	调序结果	调序概率
备用 电源 自动 切换 器	Automatic switching device for emergency power supply	3 4 5 1 2	0.43023
		1 2 3 4 5	0.32558
		5 4 1 2 3	0.24419
三维 混合 机 辅轴 补偿 机构	Auxiliary shaft compensation mechanism of three-dimensional mixer	4 5 6 7 1 2 3	0.33333
		3 4 5 6 7 1 2	0.66666
		6 7 1 2 3 4 5	0.16666
报纸 杂志 自动 分离 装置	Automatic separator of newspaper and magazine	3 4 5 1 2	0.6246
		1 2 3 4 5	0.2752
列车 自动 灭火 装置	Automatic fire fighting equipment for train	2 3 4 1	0.4872
		3 4 1 2	0.2401
		1 2 3 4	0.2754
		4 1 2 3	0.0031

4.3 解码

本文采用了现有摩西系统实现解码。摩西是不限制语言的基于短语的统计机器翻译系统，只需要双语训练语料即可。

在利用摩西解码时要用到两个软件，词对齐工具 GIZA++和语言模型训练工具 SRILM。首先，要利用 SRILM 训练语言模型，这里考虑到术语长度一般不会超过 15 个词，所以本文采用了三元语言模型。之后，利用 GIZA++抽取出双语词典，最后再根据 4.1.1 小节介绍的方法进行短语扩充，进而得到短语表。

在解码过程中，首先将术语划分成若干个短语的形式，图 4.6 给出了“停车位多重发光警示座”的几种短语划分。从图 4.6 可以看出一个术语会有多个划分结果，而且每一个划分结果中的短语在短语表中可能会有多个解释。这样每个术语的翻译结果可能会

有 N 个，摩西就是对这 N 个结果进行搜索，从而找出最适合的翻译。

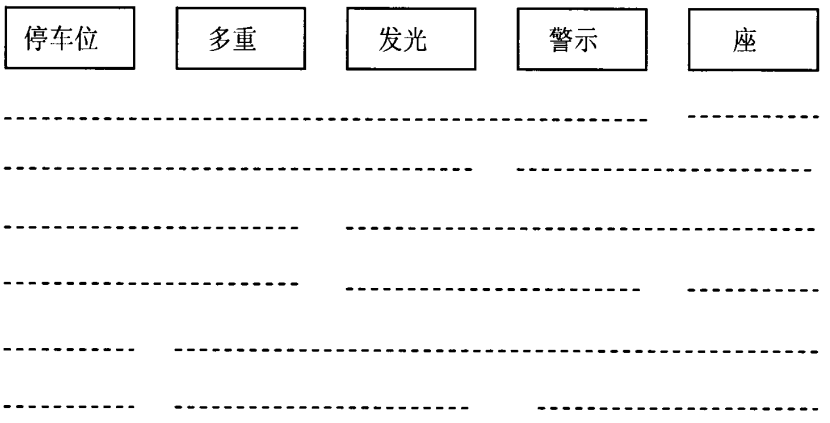


图 4.6 短语划分

4.4 实验结果及分析

4.4.1 评测方法

机器翻译的评测主要分为人工评测和自动评测，前者费时费力，而且存在着一定的主观性；而后者速度快，效率高。所以成为了目前研究的热点，机器翻译的自动评测方法有两种：BLEU 评测和 NIST 评测。

BLEU 评测方法是将机器翻译的结果和参考结果进行相似度计算，并且给出不同的 n 元语法的计算结果，相同的结果越多，BLEU 值越高。一般采用其中 n 的取值 1-4，之后选取 4 个值中的几何平均值。公式如下所示：

$$BLEU = BP \times \exp(\sum_{n=1}^{N} w_n \log p_n) \tag{4.4}$$

大量的实验表明 BLEU 评测方法能够很好的模拟人工评测的过程，所以在机器翻译结果评测过程一般采取 BLEU 值作为评测标准。

NIST 评测方法和 BLEU 方法相似，只不过 NIST 在 BLEU 的基础上，修改了惩罚因子，提高了译文评测的有效性。采用如下公式进行计算：

$$NIST = \sum_{n=1}^{N} \left\{ \sum_{\text{所有同现的}w_1 \ldots w_n} Info(w_1 \ldots w_n) \Big/ \sum_{\text{输出中的所有}w_1 \ldots w_n} (1) \right\} \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{L_{res}}, 1 \right) \right] \right\} \tag{4.5}$$

4.4.2 实验设置

本文的语料来自于专利语料，包括人类生活必需类，作业、运输类，化学、冶金类，

有 N 个，摩西就是对这 N 个结果进行搜索，从而找出最适合的翻译。

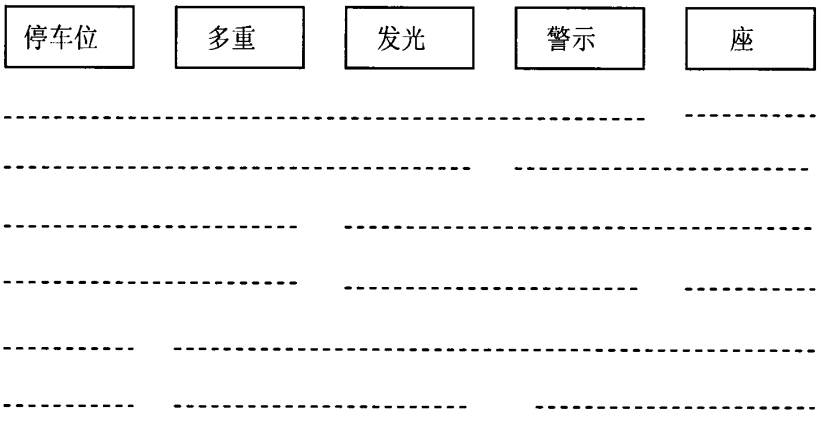


图 4.6 短语划分

4.4 实验结果及分析

4.4.1 评测方法

机器翻译的评测主要分为人工评测和自动评测，前者费时费力，而且存在着一定的主观性；而后者速度快，效率高。所以成为了目前研究的热点，机器翻译的自动评测方法有两种：BLEU 评测和 NIST 评测。

BLEU 评测方法是将机器翻译的结果和参考结果进行相似度计算，并且给出不同的 n 元语法的计算结果，相同的结果越多，BLEU 值越高。一般采用其中 n 的取值 1-4，之后选取 4 个值中的几何平均值。公式如下所示：

$$BLEU = BP \times \exp(\sum_{n=1}^{N} w_n \log p_n) \tag{4.4}$$

大量的实验表明 BLEU 评测方法能够很好的模拟人工评测的过程，所以在机器翻译结果评测过程一般采取 BLEU 值作为评测标准。

NIST 评测方法和 BLEU 方法相似，只不过 NIST 在 BLEU 的基础上，修改了惩罚因子，提高了译文评测的有效性。采用如下公式进行计算：

$$NIST = \sum_{n=1}^{N} \left\{ \sum_{\text{所有同现的} w_1 \ldots w_n} Info(w_1 \ldots w_n) \Big/ \sum_{\text{输出中的所有} w_1 \ldots w_n} (1) \right\} \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{L_{res}}, 1 \right) \right] \right\} \tag{4.5}$$

4.4.2 实验设置

本文的语料来自于专利语料，包括人类生活必需类，作业、运输类，化学、冶金类，

纺织、造纸类，固定建筑物类，机械工程类，物理类，电学类。本文选择 60 万对汉英术语对作为机器翻译的训练语料，3500 句作为测试语料。其中训练语料主要用来抽取结构化和非结构化短语以及抽取调序模型；测试语料的选择是随机选取的，也包括了上述 8 个大类的术语。

4.4.3 实验结果及分析

1 实验结果

为了证明该方法的有效性，本文采用摩西作为对比实验。实验结果如表 4.2 所示，其中 system1 是只加入了结构化短语的翻译模型，system2 是在 system1 的基础上又加入了调序模型。

表 4.2 术语翻译实验结果

方法	BLUE 评分	NIST 评分
moses	0.1694	6.1306
system1	0.1746	6.2038
system2	0.1757	6.2204

从实验结果可以看出，加入结构化短语模型以后，BLUE 值比摩西系统提高了 0.52 个百分点，这也说明结构化短语在机器翻译中是很重要的，如果只是利用非结构化的短语模型，在翻译过程中会丢掉一部分短语，出现数据稀疏问题。加入调序模型以后实验结果比 system1 的 BLUE 值提高了 0.11 个百分点，虽然有所提高，但是提高并不明显。分析原因有以下几点：（1）在调序模板抽取的过程中，要对术语进行语义分析，而我们的语义分析系统在大的语义类别上只能达到 77.13%的准确率。所以语义分析的错误直接影响到了模板抽取的质量。（2）一个语义结构对应着几个调序方式，如表 4.1 所示，我们选择调序概率最高的作为最后的调序结果，但是在少数情况下，最好的调序结果可能不是调序概率最高的，这样就会导致调序结果错误。（3）调序模板数量有限，所以部分术语在调序模型中找不到相应的调序方法。

2 错误分析

通过对 3600 句测试语料的翻译结果分析，翻译的错误主要来自两个方面。

（1）语序错误。表 4.3 给出了这种错误的几个实例，从表中的实例可以看出这种错误主要表现在英文的定语后置现象。虽然我们在翻译过程中加入了调序模型，把源语言的语序变成符合目标语言的语言习惯，但是由于上述三个原因导致不能对所有的句子进

行调序，所以语序仍然是本文的问题之一。

表 4.3 术语翻译结果中语序错误示例

原文	翻译结果	参考译文
报警 防撬门	Alarm anti-prizing door	Anti-prizing door with alarm
低压 管道 快速 接头	low-presure pipes Fast union	Fast union for low-presure pipes
窗帘 升降 装置	Curtain lifting device	Lifting device for curtain

（2）数据稀疏问题。术语是随着科技的发展不断涌现的，所以术语中总是有大量的新词出现，这也导致了组成术语的词的重复性低，出现了数据稀疏问题。尽管我们选择了 60 万句作为训练语料，但是还是难以覆盖术语中出现的所有词汇。表 4.4 给出了数据稀疏问题的错误实例。

表 4.4 术语翻译中数据稀疏问题错误示例

原文	翻译结果	参考译文
船用 航迹 仪	Marine 航迹 instrument	Marine nautical instrument
鼻腔 舒通 止 血栓	Nasal cavity 舒通 hemostatic embolus	Nasal cavity deroppilation hemostasis suppository
妇 乐保健 带	乐保健 band for woman	"Fule" health band for woman
酒 渣 摊 凉拌 曲 机	wine dregs of 凉拌 curved machine	Mold culture making machine

4.5 本章小结

本章主要介绍将术语语义分析结果应用于术语翻译的方法。首先介绍了术语翻译的过程，包括短语抽取过程和调序过程。短语抽取过程分为结构化短语抽取和非结构化短语抽取；调序过程是对训练语料语义分析之后得到调序模板，再对源语言进行调序。其次给出实验结果。最后对术语翻译的错误进行分析并给出了错误实例。

第 5 章　系统设计与实现

本文工作主要包括三个部分：术语依存分析，术语语义分析和术语翻译。术语依存分析和语义分析都采用了机器学习的方法。之后将术语语义分析结果应用于术语翻译中，使得术语翻译的结果比摩西系统的 BLUE 值有所提高。本章将分别介绍三部分的具体流程。

5.1　系统整体流程

依存分析阶段输入经分词和词性标注的术语 R，找到一棵概率最大的依存分析树 T，此过程表述为：

$$T = \arg\max p(T/R) \tag{5.1}$$

语义分析是对已经分析出依存关系的术语，确定每两个构成依存关系的词之间的语义关系过程。术语翻译分为三个阶段：短语抽取、源语言调序和摩西解码。图 5.1 给出了系统的整体示意图。

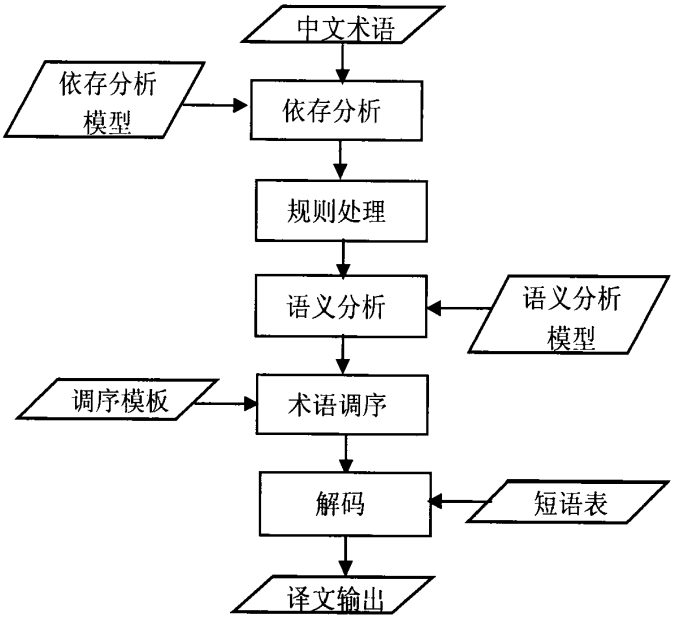


图 5.1 系统示意图

5.2 依存分析模块

依存分析利用了 SVM 的分类能力，训练依存分析模型，从而对术语进行依存分析。在依存分析模型训练过程中包括以下几个过程。

5.2.1 处理语料

我们所用的语料为双语平行语料，语料预处理是把双语平行语料的汉语部分提取出来，之后利用哈尔滨工业大学的分词和词性标注工具 IRLAS 对中文部分进行分词和词性标注。对英文部分的处理是把英文中的大写字母变成小写。

5.2.2 特征提取

在依存分析阶段，我们选取了三类特征：基本特征，互信息特征和知网第一义原特征。其中基本特征包括词，词性，上下文信息和两个词在术语中的距离；这些特征在前一步语料预处理过程可以自动提出。在互信息抽取的过程中，首先利用 3.2.1 节公式(3.5)的互信息计算方法计算语料中出现的每两个词的互信息，结果如表 5.1 所示。从表中的数据可以看出互信息特征的结果在10^{-6}的数量级上，为了和其他特征的数量级保持一致我们将互信息的值扩大了10^{6}倍。提取知网第一义原特征时，要利用到知网的信息，知网中对词语的表示方法如 3.2.1 中的例子所示。一个词可以用多个概念表示，为了规范化我们将一个词的多个概念进行提取，结果如表 5.2 所示。对于有多个概念的词语，我们要对多个概念进行选择，选择适合上下文的概念。

表 5.1 互信息示例

词语	互信息
两用 热水器	1.91624232544949e-006
固态 高铁	0.000753012048192771
多 功能	2.71666453871865e-005
醋 生产	5.30676648104777e-006
应用 有机	5.30740489130435e-006

表 5.2 第一义原结果示例

词语	第一义原
器	能力/器具/部件
机	时间/部件/飞行器/机器
火车	车

5.2 依存分析模块

依存分析利用了 SVM 的分类能力，训练依存分析模型，从而对术语进行依存分析。在依存分析模型训练过程中包括以下几个过程。

5.2.1 处理语料

我们所用的语料为双语平行语料，语料预处理是把双语平行语料的汉语部分提取出来，之后利用哈尔滨工业大学的分词和词性标注工具 IRLAS 对中文部分进行分词和词性标注。对英文部分的处理是把英文中的大写字母变成小写。

5.2.2 特征提取

在依存分析阶段，我们选取了三类特征：基本特征，互信息特征和知网第一义原特征。其中基本特征包括词，词性，上下文信息和两个词在术语中的距离；这些特征在前一步语料预处理过程可以自动提出。在互信息抽取的过程中，首先利用 3.2.1 节公式(3.5)的互信息计算方法计算语料中出现的每两个词的互信息，结果如表 5.1 所示。从表中的数据可以看出互信息特征的结果在10^{-6}的数量级上，为了和其他特征的数量级保持一致我们将互信息的值扩大了10^{6}倍。提取知网第一义原特征时，要利用到知网的信息，知网中对词语的表示方法如 3.2.1 中的例子所示。一个词可以用多个概念表示，为了规范化我们将一个词的多个概念进行提取，结果如表 5.2 所示。对于有多个概念的词语，我们要对多个概念进行选择，选择适合上下文的概念。

表 5.1 互信息示例

词语	互信息
两用 热水器	1.91624232544949e-006
固态 高铁	0.000753012048192771
多 功能	2.71666453871865e-005
醋 生产	5.30676648104777e-006
应用 有机	5.30740489130435e-006

表 5.2 第一义原结果示例

词语	第一义原
器	能力/器具/部件
机	时间/部件/飞行器/机器
火车	车

通过上述步骤的提取，得到如表 5.3 所示的特征模板。其中“类别”有两个值，“1”表示两个词存在依存关系，“-1”表示两个词不存在依存关系。在“上下文”中“/”之前的词表示当前词的前一个词，之后表示当前词的后一个词。

表 5.3 依存分析的特征

类别	词 1	词 2	词性 1	词性 2	上下文 1	上下文 2	第一义原 1	第一义原 2	互信息
1	多	功能	m	n	/功能	多/床垫	多	功用	0.5
-1	多	床垫	m	n	/功能	功能/	多	部件	0
1	保健	镯	n	n	/水晶	水晶/	事物	用具	0.5
-1	保健	水晶	n	n	/水晶	保健/镯	事物	材料	0

5.2.3 模型训练

本文选择支持向量机（SVM）来训练依存分析模型，第三章对支持向量机做了初步介绍，主要利用到 SVM 的分类能力确定两个词是否存在依存关系。由于 SVM 只识别数字特征，所以将表 5.3 所示的特征模板数字化。图 5.1 给出 SVM 可识别的数字化的特征模板。

```
-1 2938:1 10372:1 12872:1 12877:1 12904:1 12919:1 12944:1 12971:1 12998:1 14058:1
23732:1 30156:1 36369:1 38711:1 39233:1 39939:1
-1 4294:1 10372:1 12859:1 12877:1 12903:1 12930:1 12935:1 12971:1 12998:1 15952:1
23386:1 30156:1 36369:1 38711:1 39341:1 39939:1
1 1102:1 10324:1 12872:1 12899:1 12903:1 12917:1 12957:1 12984:1 13011:1 17715:1
23338:1 26964:1 37735:1 38711:1 39011:1 40076:1
1 3900:1 12640:1 12872:1 12886:1 12904:1 12930:1 12957:1 12984:1 13014:1 14116:1
24887:1 31311:1 38709:1 38711:1 39177:1 39638:1
1 5449:1 12640:1 12872:1 12886:1 12903:1 12930:1 12944:1 12984:1 13014:1 16914:1
25654:1 31311:1 38709:1 38711:1 38878:1 39638:1
-1 4256:1 9932:1 12859:1 12887:1 12903:1 12917:1 12945:1 12971:1 13011:1 18130:1
22946:1 30118:1 34163:1 38711:0 38722:1 40483:1
-1 4256:1 8301:1 12859:1 12899:1 12904:1 12917:1 12945:1 12972:1 12998:1 18130:1
22946:1 29370:1 36128:1 38711:1 38722:1 40289:
```

图 5.1 数字化的特征模板

5.3 语义分析模块

语义分析分包括语义关系定义、基于 CRF 的语义关系识别和基于 SVM 的后处理三部分。

对 3000 句术语依存分析，并将存在依存关系的词语进行语义关系总结，将术语的语义关系总结为 14 个大类，包括“宿主-属性”，“事件-受事”，“用途”，“程度”等。其中又将关系中的“宿主-属性”关系分成 7 个小的语义类别。并给句已定义的语义关系重新对 3000 句术语语义关系标注，将标注的结果作为 CRF 的训练数据。

本文在第三章中已经对 CRF 的语义关系识别过程进行了介绍，主要是特征的选择和特征模板的设置。表 3.3 已经给出 CRF 训练过程的特征集，通过多次实验得到了表 5.4 所示的特征模板，使得效果达到最高。表中 x[0,0]表示当前词的第一个特征，x[0,1]表示当前词的第二个特征，依次类推，x[0,0]/%x[0,2]表示当前词的第一个特征和第三个特征的组合特征。

表 5.4 CRF 特征模板

原子特征	组合特征
U01:%x[0,0]	U11:%x[0,0]/%x[0,1]
U02:%x[0,1]	U12:%x[0,0]/%x[0,2]
U03:%x[0,2]	U13:%x[0,1]/%x[0,3]
U04:%x[0,3]	U14:%x[0,2]/%x[0,3]
U05:%x[0,4]	U15:%x[0,0]/%x[0,1]/%x[0,2]/%x[0,3]
U06:%x[0,5]	U16:%x[0,4]/%x[0,5]
U07:%x[0,6]	U17:%x[0,0]/%x[0,1]/%x[0,4]/%x[0,5]
U08:%x[0,7]	U18:%x[0,0]/%x[0,2]/%x[0,6]/%x[0,7]
U09:%x[0,8]	U19:%x[0,0]/%x[0,1]/%x[0,6]/%x[0,7]
	U20:%x[0,0]/%x[0,1]/%x[0,8]
	U21:%x[0,0]/%x[0,3]
	U22:%x[0,1]/%x[0,2]
	U23:%x[0,0]/%x[0,8]
	U24:%x[0,1]/%x[0,8]

5.4 术语翻译模块

汉语术语语义分析结果可以用于信息检索，机器翻译等各个领域。本文将语义分析结果应用到术语翻译的短语抽取和源语言调序，使术语翻译结果的 BLUE 值提高了 0.63

个百分点。

5.4.1 短语表

本文在短语抽取的过程中，除了非结构化的短语外还加入了结构化的短语。非结构化的短语是连续的词串而不是语言学意义上的短语，这样抽取的短语结果虽然覆盖了很大一部分信息，但是长距离依存所产生的短语却被遗漏，针对这种情况我们加入了结构化的短语。表 5.5 给出了汉语短语和对应的英语短语以及概率。

表 5.5 短语表示例

汉语短语	英文短语	汉到英翻译概率	英到汉翻译概率
胶 软 包 天棚 板	rubber soft-packing ceiling	1	1
效果 检测	testing effect of	0.25	0.5
效果图 像 测量	effective image measaring	1	0.5
效率 非 接触 式 充电 装置	efficient contactless charger	1	1
直 压 式	Direct press type	0.166667	1

5.4.2 调序

机器翻译的调序一般包括对源语言的调序和对目标语言的调序两种，本文采用对源语言进行调序，使源语言的语序更符合目标语言的顺序。首先运用 GIZA++对双语平行语料词对齐，其次对汉语术语进行语义分析，最后利用前两步的结果抽取调序模板，对汉语术语进行调序。

经过调序后我们利用摩西进行解码，第四章详细介绍了摩西解码的过程，这里就不再介绍。翻译结果如表 5.6 所示。

表 5.6 术语翻译示例

汉语术语	英文翻译
多 层 复合型 保温 絮 片	Multi-layer composite heat insulation flocculus
电 致 发光 材料 包膜	Electroluminescent material coating
固定式 网络 电阻器	Fixed network resistor
电流 接收 传感器 装置	Current receiving sensor device
电 极 型 荧光灯	Electrode type fluorescent lamp
多 功能 飞碟 空气 清新 器	Multifunctional flying saucer air freshing device
多头 磨光 机	Multi-head polishing machine

个百分点。

5.4.1 短语表

本文在短语抽取的过程中，除了非结构化的短语外还加入了结构化的短语。非结构化的短语是连续的词串而不是语言学意义上的短语，这样抽取的短语结果虽然覆盖了很大一部分信息，但是长距离依存所产生的短语却被遗漏，针对这种情况我们加入了结构化的短语。表 5.5 给出了汉语短语和对应的英语短语以及概率。

表 5.5 短语表示例

汉语短语	英文短语	汉到英翻译概率	英到汉翻译概率
胶 软 包 天棚 板	rubber soft-packing ceiling	1	1
效果 检测	testing effect of	0.25	0.5
效果图 像 测量	effective image measaring	1	0.5
效率 非 接触 式 充电 装置	efficient contactless charger	1	1
直 压 式	Direct press type	0.166667	1

5.4.2 调序

机器翻译的调序一般包括对源语言的调序和对目标语言的调序两种，本文采用对源语言进行调序，使源语言的语序更符合目标语言的顺序。首先运用 GIZA++对双语平行语料词对齐，其次对汉语术语进行语义分析，最后利用前两步的结果抽取调序模板，对汉语术语进行调序。

经过调序后我们利用摩西进行解码，第四章详细介绍了摩西解码的过程，这里就不再介绍。翻译结果如表 5.6 所示。

表 5.6 术语翻译示例

汉语术语	英文翻译
多 层 复合型 保温 絮 片	Multi-layer composite heat insulation flocculus
电 致 发光 材料 包膜	Electroluminescent material coating
固定式 网络 电阻器	Fixed network resistor
电流 接收 传感器 装置	Current receiving sensor device
电 极 型 荧光灯	Electrode type fluorescent lamp
多 功能 飞碟 空气 清新 器	Multifunctional flying saucer air freshing device
多头 磨光 机	Multi-head polishing machine

5.5 本章小结

本章主要介绍了汉语术语语义分析的各个模块的设计与实现，分别从依存分析和语义分析两个部分介绍。并在语义分析的基础上实现了汉语术语翻译，翻译的过程分别介绍短语抽取与调序。

结　论

随着科技的发展，专利术语也随之增多，理解术语内部的组成结构以及术语所表达的意义无论对信息检索还是机器翻译都有着重要的作用。目前的术语翻译主要有两种方法：基于网络的方法和基于统计的方法，而语言学特征的机器翻译还停留在句法层次上，没能真正理解术语所表达的语义信息。

本文提出了基于统计与规则相结合的术语语义分析方法，并在此基础上对术语进行翻译。翻译结果和摩西系统对比，BLUE 值有所提高，这也说明了本系统方法的有效性。

全文的主要工作及得到的主要结论总结如下：

1．提出了一种基于 SVM 的汉语术语依存分析方法。该方法把依存分析看做分类过程，选用基本特征，互信息特征和知网第一义原特征训练依存分析模型。其中知网第一义原代表了该词的语义类别，加入该特征后系统的效果有了明显提高，这也说明从语义层次上来解决依存分析问题是有效的。为了证明该系统的有效性，本文结合术语的特点做了两个对比实验：实验 1 规定术语中所有词都修饰中心词，实验 2 规定所有词都修饰与它相邻的词。通过实验结果可以看出本文方法的性能远远高于上述两种方法。

2．在依存分析的基础上，提出了一种基于 CRF 与 SVM 相结合的术语语义分析方法。在进行语义分析之前，针对分词错误进行规则处理；之后对组成术语的词之间的语义关系进行总结，归纳出 14 种语义关系。并对其中的“宿主-属性关系”进一步细分，分成了 7 个小的语义类别。并对 3000 句术语进行人工标注，作为 CRF 的训练语料训练语义分析模型。由于术语所涉及的语义关系范围窄导致上述语义分析模型的分类能力差，为此本文加入了基于 SVM 的后处理模型，该模型针对只针对三个易混淆类别进行处理，使系统效果有所提高。

3．最后提出一种基于语义分析的术语翻译方法。该方法包括三个部分：短语抽取，源语言调序和解码。传统的基于短语的机器翻译抽取的短语是连续的词串，而非语言学意义上的短语，针对该问题，本文在短语抽取的过程中加入了结构化短语的抽取，提高了短语表的覆盖度，在一定程度上解决了数据稀疏问题。在源语言调序阶段，本文提取调序模板，根据调序模板对源语言进行调序。在解码阶段，本文利用了开源的统计翻译工具摩西进行解码。

对今后工作的建议

利用本文的方法，虽然取得了很好的术语语义分析效果，并使术语翻译的结果有所改善，但是通过分析错误实例，本文认为该方法还有很大的提升空间。对今后工作的建议如下：

1．在依存分析的错误实例中，大部分是由分词错误引起的，所以下一步工作要开发一个术语分词系统。

2．在语义分析的后处理过程中，本文只针对三类易混淆的类别进行后处理，下一步工作将扩大易混淆类别，并加入到后处理的过程中。

3．在术语翻译的调序过程中，如果一个语义结构对应多个调序结果，本文选择第一个作为最终调序，这也带来了一部分调序错误。下一步准备加入互信息和其他语义信息改善调序结果。

参 考 文 献

[1] 冯志伟. 现代术语学引论[M]. 北京: 语文出版社, 1997:1-5

[2] 宗成庆. 统计自然语言处理.北京：清华大学出版社, 2008.5

[3] Chris Quirk , Arul Menezes and Colin Cherry. Dependency Treelet Translation : Syntactically Informed Phrasal SMT [A] . In Proceedings of the ACL 2005 [C]. 2005

[4] Yuan Ding and Martha Palmer. Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars [A]. In : Proceedings of the ACL 2005 [C] .2005

[5] Lin Dekang. A Path2based Transfer Model for Machine Translation [A].In Proceedings of COLING 2004 [C] .

[6] 冯志伟. 特思尼耶尔从属关系语法[J].国外语言学. 1983,1

[7] 周明, 黄昌宁. 面向语料库标注的汉语依存体系的探讨. 中文信息学报, 1994, 1(3)

[8] 罗强, 奚建清, 一种结合 SVM 学习的产生式依存分析方法. 中文信息学报, 2007, 21(4): 21-26

[9] XU Yun, ZHANG Feng. Using SVM to construct a Chinese dependency parser [J] . Journal of Zhejiang University Science A , 2006, 7 (2) : 192-203

[10] Wenliang Chen, Daisuke Kawahara, Kiyotaka Uchimoto et al. Dependency parsing with short dependency relations in unlabeled data. In Proceedings of IJCNLP-2008, Hyderabad, India, January 8-10.

[11] 辛霄, 范士喜, 王轩等. 基于最大熵的依存句法分析. 中文信息学报, 2009, 23(2): 19-22

[12] 计峰，邱锡鹏. 基于序列标注的中文依存句法分析方法. 计算机应用与软件,2009,26(10):133-135

[13] Ryan McDonald, Fernando Pereira. Online Learningof Approximate Dependency Parsing Algorithms.[C]/ / EACL 2006.

[14] 郭铙华, 周昌乐. 一种汉语语句依存关系网分析策略与生成算法研究. 浙江大学学报, 2000,27(6) 637-645

[15] Ryan McDonald, Koby Crammer, and Fernando Pereira. Online large-margin training of dependency parsers. In Proceedings of ACL. 2005a:91–98

[16] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Haji˘c. Non-projective dependency parsing using spanning tree algorithms. In Proceedings of HLT-EMNLP. 2005b

[17] Joakim Nivre, Johan Hall, Jens Nilsson, Gulsen Eryigit, and Svetoslav Marinov. Labeled Pseudo projective dependency parsing with support vector machines. In Proceedings of CoNLL, 2006:221–225

[18] Dan Klein and Christopher D. Manning. Corpusbased induction of syntactic structure: Models of dependency and constituency. In Proceedings of the ACL. 2004

[19] Terry Koo, Xavier Carreras, and Michael Collins. Simple semi-supervised dependency parsing. In Proceedings of the ACL. 2008

[20] Eisner，J.M.Three new Probabilistie models for dependeney Parsing:an exploration.In Proceedings of COLING'96.1996a:340-345

[21] Eisner, J.M.An empirical comparison of probability models for dependeney grammar[R], Technical Report IRCS-96-11, Institute for Research in Cognitive Science ,University of Pennsylvania.1996b

[22] Eisner, J.M. Bilexical grammars and their cubi — time parsing algorithms[R]. Advances in Probabilistic and Other Parsing Teclmologies, 2000:29-62

[23] 刘挺, 马金山, 李生. 基于词汇支配度的汉语依存分析模型. Journal of Software, September 2006 17(9): 1876−1883

[24] MeDonald R, Crammer, KandPereira F. Online large-margin training of dependency parsergA]. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), 2005: 91-98

[25] MeDonald R, Pereira F, Ribaroy K and Haji.J. Non 一 projective dependency parsing usings spanning tree algorithms. In proceedings of HLT 一 EMNLP.2005b

[26] MeDonald R and Pereira F. Online learning of approximate dependeney parsing algorithms. In proceedings of the 11th Conference of the European Chapter of the association for Computational Linguistics (EACL). 2006: 81-88

[27] Nivre, J, Constrains. on non-projcetive dependeney parsing. In proceedings of the 11th Conference of the European Chapter of the association for Computationa lLinguisti (EACL).2006:73-80

[28] Yamada H and Matsumoto Y. Statistical dependency analysis with support veetor macehines. In Proceedings of IWPT'2003. 2003: 19-206

[29] Li MQ, Li JZ, Dong ZD,et al . Building a large Chinese corpus annotated with semantic dependency. In the 2nd SIGHAN Workshop on Chinese Language Processing. 2003:84-91

[30] 李明琴, 李涓子, 王作英, 陆大给. 语义分析和结构化语言模型. 软件学报, 2005, 16(9)

[31] 李钝, 乔保军, 曹元大. 基于语义分析的词汇倾向识别研究. 模式识别与人工智能, 2008, 1(4)

[32] 邹娟, 周经野, 邓成. 一种基于语义分析的中文特征值提取方法

[33] 徐波, 孙茂松, 靳光瑾. 中文信息处理若干重要问题. 科学出版社, 2003

[34] N.Chomsky. Aspects of Theory of Syntax Structure [M]. Revue et carriage, Paris, 1976

[35] N.Chomsky. Lectures on Government and Binding [M]. Dordrecht: Foris, 1981

[36] J.G.Carbonell, M.Tomita. Machine Translation: Theoretical and Methodological Issues [M]. Cambridge University Press, 1987

[37] Philipp Koehn. Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation Models [A]. In Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas[C], Washington, DC, 2004: 115-124

[38] D.Xiong,Q. liu,et al. Maximum EntropyBased Phrase Reordering Model for Statistical Machine Translation. Sydney,Australia In Proceeding of the ACL 2006

[39] Michel Galley Christopher D. Manning A Simple and Effective Hierarchical Phrase Reordering Model proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008: 848–856

[40] 熊德意, 刘群, 林守勋. 基于句法的统计机器翻译综述. 中文信息学报, 2008, 22(2)

[41] Dekai WU. A polynomial-time algorithm for statistical machine translation [A]. In Proceedings of 34th Annual Meeting of the Association for Computational Linguistics[C]. 1996

[42] David Chiang. A hierarchical phrase-based model for statistical machine translation. In Proceedings of 43rd Annual Meeting of the ACL, 2005: 263-270

[43] Liang Huang , Kevin Knight , and Aravind Joshi. Statistical Syntax2Directed Translation with Extended Domain of Locality [A] . In : Proceedings of the 7th AMTA [C] . Boston , MA : 2006

[44] Chris Quirk, Arul Menezes and Colin Cherry.Dependency Treelet Translation: Syntactically Informed Phrasal SMT [A] . In : Proceedings of the ACL 2005 [C]. 2005

[45] Daniel Marcu. Argmax Search in Natural Language Processing [R]. Invited talk in ACL COL IN G

2006

[46] Xianchao Wu, Naoaki Okazaki, Takashi Tsunakawa, Jun'ichi Tsujii. Improving English-to-Chinese Translation for Technical Terms Using Morphological Information. The Eighth Conference of the Association for Machine Translation in the Americas[C],2008

[47] 王金玲, 张桂平, 叶娜等. 基于多特征的日-汉术语翻译技术的研究. 2009 年全国模式识别学术会议暨首届中日韩模式识别学术研讨会, 南京理工大学, 2009: 670-674

[48] Gaolin Fang, Hao Yu, Fumihito Nishino. Chinese-English Term Translation Mining Based on Semantic Prediction. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistic[C], 2006

[49] 黄德根, 马玉霞, 杨元生. 基于互信息的中文姓名识别方法. 大连理工大学学报, 2004, 44(5): 744-748

[50] IRLAS.http://www.ir.hit.edu.cn

[51] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[C]. 第三届汉语词汇语义学研讨会论文集, 北: [s. n.], 2002: 59 -76

[52] 周晶, 吴军华, 陈佳等. 基于条件随机域 CRF 模型的文本信息抽取. 2008, 29(23): 6094-6097

[53] D. Ortiz-Martinez, I. Garcia-Varea, F.Casacuberta. Thot: A Toolkit to Train Phrase-based Statistical Translation Models [A]. Proceedings of the 10th Machine Translation Summit[C], Thailand, 2005

致　谢

时间过得很快，转眼间两年半的研究生生活即将结束，我的学生生涯也接近尾声。在这里我要向所有关心我，帮助我的人表示衷心的感谢。

感谢我的导师张桂平教授，她对科学高屋建瓴，真知灼见使我感受到学者的风采，对事业求真务实，让我体会到了人生的精彩。张老师事务繁忙，但依然挤出时间关心我们的学习与生活，时常请来业界的知名人士为我们做报告，使我们开拓眼界，并能及时了解自然语言处理界的研究动态。也让我们真实生动的体会到每天都在进步，收获了自信。同时，张老师严格要求我们，对我们直言不讳，使我们清楚的看到了自己的不足。张老师教会我们如何做研究，如何做人。在此，向张老师致以最真诚的谢意。

感谢我的导师蔡东风教授，蔡老师对科学的执着，不知疲倦的工作作风使我看到了学者的风范。无论从生活还是学习蔡老师都给予我们极大的帮助，像亲人一样关心着我们每一个人。在我课题研究和论文撰写过程中蔡老师都给予了极大的帮助，每当课题遇到问题时蔡老师能够提出宝贵意见，使我及时改正研究观点上的误区。同时，蔡老师也为我们提供了外出学习和开会的机会，使我们的研究生生活更加充实和丰富。在此，向蔡老师表达我深深的敬意!

感谢实验室的叶娜老师，季铎老师，周俏丽老师，白宇老师和王裴岩老师，感谢你们在我的学习过程中给予的帮助。特别要感谢叶老师在课题研究和论文撰写过程中给我提供的帮助。

感谢陪伴我两年半的同窗好友，是你们的陪伴让我的生活更加多彩！感谢马丽丽师姐和刘新师姐在我写论文的过程中给我的帮助。感谢师弟师妹在我实验过程中为我标注语料。

感谢我的父母和姐姐，感谢你们在我最困难的时候给我的关心和支持，使我克服种种困难，你们是我坚强的后盾。

最后，要感谢答辩委员会的各位老师评审我的论文和出席我的毕业答辩会。

作者：陈小芳

攻读硕士期间发表（含录用）的学术论文

1 **陈小芳**，张桂平,蔡东风,叶娜. 基于统计和规则相结合的汉语语义分析方法. 第六届全国信息检索, 哈尔滨, 哈尔滨工业大学. 2010,8：482-488

2 叶娜，**陈小芳**，蔡东风. 面向专利文献的术语自动处理技术. 沈阳航空工业学院学报. 2010,8：32-35s

汉语术语语义分析技术研究及其应用

作者：陈小芳
学位授予单位：沈阳航空航天大学