一种汉英翻译模板提取方法*

杨二宝[1] 吕学强 朱靖波 姚天顺[2]
东北大学信息学院 软件与理论研究所 沈阳 110004
Email: paulneverdie@etang.com

摘 要：本文定义了一种汉英翻译模板，并在此基础上提出了一种从语料库中自动学习模板的方法。该方法用语义分类体系约束模板变量，并引入了模板抽象度的概念，以保证模板的正确性。加入基本名词短语的捆绑，和单语语料中的覆盖度统计信息，从而增大模板的覆盖度。在法律领域的试验结果表明，这种方法生成的翻译模板质量具有很高的实用价值。

关键词：机器翻译，翻译模板，语料库，对齐

A Method of Chinese-English Tranlsation Template Extraction

Erbao Yang, Xueqiang Lv, Jingbo Zhu, Tianshun Yao
Institute of software & theory, Northeastern University, Shenyang 110004
Email: paulneverdie@etang.com

Abstract: A type of Chinese to English tranlsation template is defined in this paper, then based on it, a method of template extraction is discussed. In order to increase the accuacy, semantic classification is used as the variable constrains, and the concept of template abstract rato is introduced. In order to increase the coverage, base-NP binding and statistical information is adopted. The experiment in the domain of law text shows that translation template extracted from this method is practically useful.

Keywords: machine translation, translation template, corpus, alignment

1. 前言

翻译模板是基于实例的机器翻译（EBMT）中常用的技术。翻译模板在应用 EBMT 之中，主要起两个作用，一是以解决平行语料库的数据稀疏问题。二是减小实例库的规模。90 年代以后翻译模板成为 EBMT 的发展方向. 模板抽取的方法大致上可以分为三种：一、规则的方法，根据单个实例的某种语言学特性进行抽象；二、统计的方法，根据多个实例的共性归纳出模板；三、规则与统计结合的方法。本文提出了一种规则与统计相结合的模板抽取方法。

*本文得到国家自然科学基金和微软联合资助项目（60203019）资助。
[1]杨二宝：男，硕士生，主要研究方向为多国语机器翻译。
[2]姚天顺：男，教授，博士生导师，主要研究方向为计算语言学理论。

2. 翻译模板的定义

翻译模板是一种翻译规范，是所有符合某种规则的翻译实例的集合，对于集合中每一个元素，它的源语部分必须按照特定的约束翻译成目标语。这是广义上翻译模板的定义。为了方便起见，本文将翻译模板简称为模板。模板的组成成分可以分为固定部分和可变部分（变量）。不同形态的模板主要体现在变量形式的不同，变量可以为词，短语，或者是语法分析之后的某个句子成分。

本文中定义的模板固定部分为单词本身，可变部分为一组符合特定约束的名词或基本名词短语（基本名词短语的集合。用继承结构的语义分类来约束变量，用类名命名变量名。*例如：【人】/明天/要/去/【处所】 // 【人】will go to 【处所】 tomorrow。*本文使用的语义分类取自《同义词词林》的前 4 层。

本文中的模板变量只用来约束名词或基本名词短语，这是因为：①语法功能假设。不同词性的词在句子中的作用是不同。名词和名词短语在运载实体信息，而其他词性的词主要起着维护句子结构的作用。同一种结构可以通过变换名词部分来运载不同的实体信息。②词对齐的因素。通常名词相对其他词性的词在词对齐方面的准确率和召回率较高。③语义消歧的因素。单词在词典中的译项（特别是受限领域）也相对较少。

3. 模板的提取过程

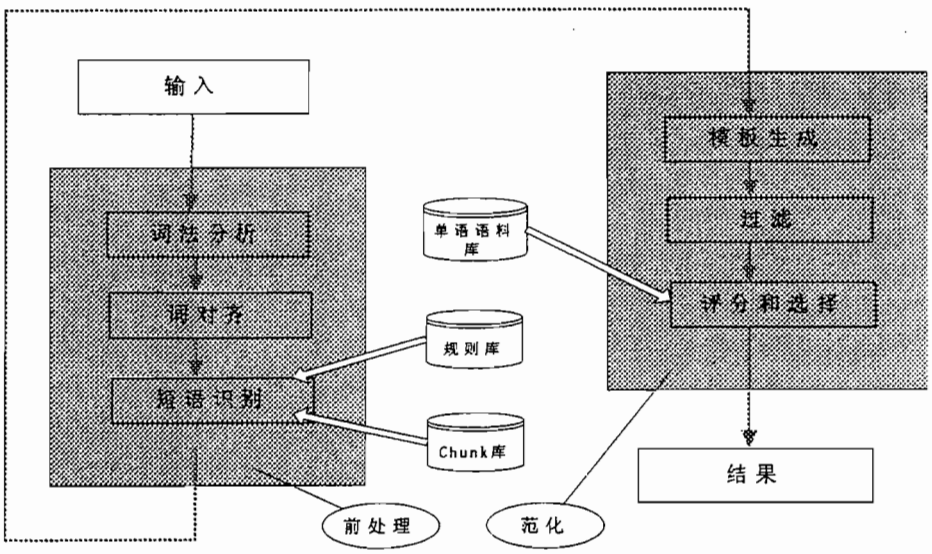

图 1 模板提取流程

从上图 1 中可以看出，模板提取所需要的资源有：语义分类词典，子句级对齐的

双语平行语料，按子句切分的单语语料以及短语捆绑所需的规则库。模板的提取流程大致可以分为两个部分：前处理和提取变量。前处理采用了一些常规的手段，只作简单的介绍。范化是模板自动抽取的核心步骤，也是本文的创新点所在，所以将用较长的篇幅进行介绍。

3.1 前处理

词法分析包括对实例中文部分的分词，词性标注，对英文部分的词型还原。

词对齐采用了东北大学吕学强[8]中的多层次对齐方法。该方法在传统的词对齐方法的基础上，引入了最小求交模型、最大求差模型等基于统计的对齐方法。对于本文实验所用的法律语料，词对齐的正确率达到89%，召回率达到 83%。

基本名词短语捆绑分为源语 Base-NP 识别和目标语对译单元捆绑两个步骤。Base-Np 的识别参考了[9]中的方法，并结合了使用 chunk 库覆盖的方法。第二步采用了王伟在[7]中提出的组块儿对齐的方法。

3.2 范化

因为该步骤是在词对齐的基础上进行的，对源语部分每个词的操作，都能很容易地映射到目标语部分。所以我们只对源语部分进行论述。这一步要解决的问题是：最终的模板应该将原来实例中哪些名词变量化？每一个变量的限制采用类名还是上位类名？

3.2.1 模板生成

生成所有可能的模板。所有的名词都是提取变量的候选单元，名词可以被它所在的语义类或者上位语义类所代替。 C= class (N); F= super(C)。 其中，N 代表名词，C 代表 N 所在的语义类。F 代表 C 的上位类。提取变量的过程中，N 可能被 C 或 F 所取代。

因为一词多义现象的存在，C = class (N) 必须考虑到一个类别消岐的问题。例如“翻译”作为名词时，可以是一个人，可以是一种职业，还可以是一门学科。传统的规则系统中，语义消岐被认为是非常困难的工作。而本文中的类别消岐具有以下优势：1. 受限领域内，名词的义项通常不会太多。2. 利用词对齐的结果可以用目标语对源语进行消岐。3. 模板的固定部分为变量部分提供了大量、准确的上下文信息。本文中的类别消岐主要应用了三个层次的方法。一、借助译文对源语单词进行消岐。二、根据上下文关系进行类别消岐。其中上下文的关系用规则的方法来描述，规则在加标语料中通过统计来获取。三、根据词性进行消岐。

一个翻译实例的源语部分形式化描述为：$s_0\ n_1\ s_1\ n_2\ s_2 \cdots \cdots n_i\ s_i \cdots \cdots n_m\ s_m$ 其中 n_i 为名词，s_i 为其它词或词串，因为经过了基本名次短语的捆绑，不会有两个名词相邻。s_0 和 s_m 可以为空。规定 n_i^0 = class (n_i) ； n_i^1 = super (n_i^0) 。作表格如下：

表 1

	S_0	n_1	S_1	…	n_i	S_i	…	n_m	S_m
C		n_1^0			n_i^0			n_m^0	
F		n_1^1			n_i^1			n_m^1	

在表 1 中我们从 s_0 出发，到 s_m 终止，作迪卡尔积。得到结果如下：

$s_0\ n_1\ s_1 \cdots\cdots n_i\ s_i \cdots\cdots n_m\ s_m$

$s_0\ n_1^0\ s_1 \cdots\cdots n_i\ s_i \cdots\cdots n_m\ s_m$

$s_0\ n_1^1\ s_1 \cdots\cdots n_i\ s_i \cdots\cdots n_m\ s_m$

$\cdots\ \cdots \quad \cdots\ \cdots \quad \cdots\ \cdots$

$s_0\ n_1^0\ s_1 \cdots\cdots n_i^0\ s_i \cdots\cdots n_m\ s_m$

$s_0\ n_1\ s_1 \cdots\cdots n_i^0\ s_i \cdots\cdots n_m\ s_m$

$s_0\ n_1^0\ s_1 \cdots\cdots n_i^0\ s_i \cdots\cdots n_m\ s_m$

$\cdots\ \cdots \quad \cdots\ \cdots \quad \cdots\ \cdots$

$s_0\ n_1^1\ s_1 \cdots\cdots n_i^1\ s_i \cdots\cdots n_m^1\ s_m$

3.2.2 过滤

上面产生的所有可能的模板集合中，有一些太抽象的模板价值不大，应改被过滤掉。

我们规定，用以下三条原则来衡量模板的抽象度。1. 模板长度越小，模板越抽象。模板长度就是模板中含有多少个词，（一个变量算作一个词）2. 模板含有的变量越多越抽象。3. 变量的限制越小，模板越抽象。本文中，变量是用语义分类来限制的。类别在《同义词词林》中的层次越低，限制越大；层次越高限制越小。为了给模板抽象度一个定量的描述，我们定义：$abs = r \times \frac{n \times \sum v_i}{m}$

abs 模板的抽象程度，v_i 为用来限制模板变量的每个语义类的抽象系数。我们根据语义类在《同义词词林》中的层次，赋予它们不同的抽象系数，第一层为 7；第二层为 5；第三层 3；第四层 2；名词本身的抽象系数为 0； n 为变量的个数；m 是模板的长度；r 为一个经验常数，本文试验中取 r = 0.5。在上一步生成的所有可能的模板中，抽象度大于给定阈值的模板被过滤掉。

3.2.3 评分和选取

将过滤剩余的每一个模板评分，然后选取分数最高的模板作为这个实例最终生成的模板。本文的评分标准如下：$score = r_1 \times abs + r_2 \times \mathrm{cov}$

其中 abs 为模板的抽象度，cov 是模板的覆盖度，r1, r2 分别是两个因素的权重。本文的试验中，我们取 r1 = 0.1, r2 = 0.9 。这里引入了一个覆盖度的概念。如果一个模板和某个子句的相似度大于一个给定的阈值，我们就认为该模板可以覆盖这个子句。一个

模板在一个给定的语料库中，可以覆盖的子句的数目，就叫做该模板对于那个语料的覆盖度。因为覆盖度不涉及目标语，所以我们用同质、海量的单语子句库作模板覆盖度的统计。模板的相似算法采用[8]中的方法

算法 1 统计模板 m 在预料库中的覆盖度 cov

输入：一个模板 m，子句库 M

输出：m 在 M 中的覆盖度 cov

算法：

1. 建立相关子句集合 $S = \{s_i \mid sim(m, s_i) > K1\}$
2. $cov = 0$;
3. For each s_i in S
   a) 生成 s_i 所有可能的模板集合 T
   b) 过滤 W
   c) For each t_i in T
      i. 计算 $v_i = Sim(m, t_i)$;
   d) $v = max(v_i)$;
   e) 如果 $v > K2$ 则 $cov = cov + 1$
3. Return cov

K1，K2 是事先设定的两个阈值，满足 0<K1<K2<1。算法的第一步是提高效率的关键所在，子句库 T 是非常庞大的，逐一计算无疑是对计算机的一种摧残。而实际上 T 中大部分子句都是和模板 m 不相关的。因为子句库中不含变量，第一步实际上是利用模板中的不变部分为索引，找到和模板比较相似的字句作为候选集合。

4. 试验

我们的实验是面向法律领域的。双语平行语料取自香港《双语法例资料系统》，经过预处理后取得字句级的翻译实例 134，617 个。单语语料取自网上收集的法律文献，共 1，790，420 子句。翻译的测试集合随机取自上面的单语语料，共 1000 子句。翻译结果的正确性由专家来决定。

评价标准：为了检测生成模板的质量，我们作如下定义：$A = D/L$；$C = L/P$

A 为模板库的平均正确率，C 为模板库的覆盖率，D 为输入中被正确翻译的子句数目，L 为输入中被模板库覆盖子句数目，P 为输入句子的总数

按照本文提出的方法抽取出的翻译模板，和一些传统的模板抽取方法抽取的翻译模板，在翻译效果上进行对比。方法 1 采用[5]中的方法，利用多个实例中相通的部分和不同的部分归纳学习，得出模板。方法 2 采用[3]中的方法，通过两个 parser 找到结构对应关系，

再将某个成分变量化。方法三采用本文提出的方法。结果如下：

表 2 不同模板抽取方法的结果对比

方法	模板库平均正确率%	模板库平均覆盖率%	模板的可读性	模板数
1	74.7	13.4	差	16,769
2	55.7	83	差	5,833
3	89	34	好	84,532

从上面结果中可以看出，方法 1 面临严重的数据稀疏问题，生成模板数量少，所以覆盖度低。因为没有严格的变量约束条件，模板的正确率也不高。这种方法理论上需要非常庞大的平行语料。2 的方法以句子结构来做模板，显然覆盖率是相当高的，但是分析器本身就包含了太多的错误，所以生成模板的正确性不高。而本文中的方法 3 对比前两种方法，取得了较好的平衡点。

5. 结论

本文定义了一种新型的翻译模板。利用语以分类来约束模板变量。并在此基础上推出一种模板抽取方法，该方法引入了模板抽象度的概念，并集成了基本名词短语捆绑，和语料库中的统计信息。实验表明，应用这种方法抽取出来的模板形式简单，意义明确，易于人工校对，准确率和覆盖率也可以达到较好的平衡点。

参考文献

[1] Brown, R.D. 1999. Adding linguistic knowledge to a lexical example-based translation system. Proceedings of the 8th International Conference on Theoretical Methodological Issues in Machine Translation (TMI 99).

[2] Carl, M. (1999) Inducing Translation templates for Example-Based Machine Translation. Machine Translation Summit VII Singapore, 617-624

[3] Kaji, H., Y Kida & Y, Morimoto (1992) Learning Translation Templates form Bilingual Text, Proceeding of the Fourteenth [sic] Linguistics: COLLING-92, Nantes, France, 672-678

[4] Watanabe, H. (1993) A Method For Extracting Translation Patterns for Translation Examples. Proceedings of 5th International Conference on Theoretical and Methodological Issue in Machine Translation (IMI-93) MT in the Next Generation, 292-301, Kyoto Japan.

[5] McTait, K. & A. Trujillo (1999). A language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns. Proceeding of the 8th International conference Theoretical and Methodological Issue in Machine Translation (IMI 99), Chester, England, 98-108

[6] Tiansun Yao, Erbao Yang (2002) A Method of Similar Template-Based Chinese-English MT, CJNLP-2

[7] Wei Wang et al 2001, Finding Target Language Correspondeance for Lexicalized EBMT System Proc. Of NLPRS, pages 455-460, Tokyo, Novermber, 27-29

[8] 吕学强 （2003） 《面向机器翻译的 E-Chunk 获取与应用研究》， 东北大学

[9] 赵军，黄昌宁 1999 《汉语基本名词短语结构分析模型》 计算机学报