汉语对应英语定语从句结构的一种自动翻译方法[1]

王雷[1,2]　常宝宝[1]　俞士汶[1]

[1]北京大学计算语言学教育部重点实验室；[2]北京大学英语系 北京 100871

Email: {wangleics, chbb, yusw}@pku.edu.cn

摘 要： 受到西方语言语法体系的影响，现代汉语中能够翻译成英语对应的从句的结构越来越多。但是因为汉语传统的语法结构与西方语言语法体系有着很大的不同，利用西方语言的句法分析方法对汉语句子进行语法分析始终无法达到令人满意的结果。句法分析效果不佳，就导致了像从句这类具有嵌套结构的句子翻译效果不会很好。本文尝试利用机器学习中的条件随机场方法先对这类从句结构进行识别，然后利用中心词转录机的方法进行基于依存关系的句法分析并同时进行自动翻译，实验证明这种方法在此类句子的翻译中会获得比统计机器翻译系统更好的效果。

关键词： 汉语，定语从句结构，自动翻译

A Method of Automatic Translation of Attributive Clauses in Chinese Language

Wang Lei[1,2]　Chang Baobao[1]　Yu Shiwen[1]

[1]Key Lab of Computational Linguistics of Ministry of Education; [2] Department of English of Peking University

Beijing 100871

Email: {wangleics, chbb, yusw}@pku.edu.cn

Abstract: Influenced by the grammatical system of western languages, there are more and more syntactic structures in modern Chinese that can be translated into English attributive clauses. But for the obvious differences of the syntactic structures, parsing Chinese by western grammatical rules usually does not lead to satisfactory results, which will result in poor translation performance in embedded syntactic structures. This paper first attempts to identify the attributive clauses by using conditional random field theory and then applied head transducers to parse the clauses and complete the translation. The experiment proves that this method will produce a better result than statistical machine translation models.

Key Words: Chinese language, Attributive Clause, Automatic Translation

1 引言

在一种语言中，句子中包含子句是常见的语法现象。受到西方语言语法体系的影响，在现代汉语中包含子句的复合句子结构越来越多。但是汉语本质上属于“意合型”的语言，词一般没有形态变化，语义因素在句法分析过程中占有举足轻重的地位。众所周知，在句法分析中融入语义因素是非常困难的，这就导致很多情况汉语句法分析的效果并不佳，尤其是对于句子中还嵌套子句这种复杂句法结构。而我们这里所说的子句，并非是指汉语中的小句，而是指把汉语翻译成英语时，需要翻译成英语从句结构的我们所说的汉语词组。

朱德熙[1]指出汉语属于“词组本位”的语法体系，而詹卫东[2]则按照语法功能对汉语词组进行了归类（见表1）。在这些词组类别中，能够翻译成英语定语从句结构的有表1中几类：

[1]本文相关研究得到973课题“文本内容理解的数据基础（课题编号：2004CB318102）”的支持。

表 1 能够翻译为英语定语从句结构的汉语词组

序号	词组类别	实例
1	主谓结构	树叶黄了；小明喜欢看电视；感冒传染
2	述宾结构	喝了三杯酒；企图逃跑；送他香烟；学了三年
3	述补结构	洗干净；做得非常好；吃得完；好得很；拿出来
4	状中结构	快跑；把饭吃完；明天见；屋里坐；认真地学习
5	连谓结构	开着窗户睡觉；安排助理办理；打电话请医生；请他来
6	联合结构（须含动词）	小说和戏剧；又高兴又难过；批评教育
7	附加结构（须含动词）	吃了；努力奋斗过；砍光了
8	的字结构（须含动词）	买菜的；老师表扬了的；冰凉的；慢性的

无论是基于规则还是基于统计的机器翻译方法，正确分析句子结构是进行正确翻译的关键一步。正因为汉语句子结构的特殊性，在自动翻译带有嵌套结构的句子时出现了很多困难，翻译效果不理想。比如下面两个带有可翻译成英语定语从句结构的汉语句子：

例 1　他迫不及待地打开了妈妈寄给他的红色包裹。

例 2　阿强应用概率方法实现的模型被证明不会产生相应的结果。

（注：上面二例中下划线的部分一般翻译成英语的定语从句。）

我们利用 google 的翻译系统来进行翻译会得到如下的结果：

例 3　He could not wait to open his mother sent him a red package.

例 4　Ah-chiang of applied probability methods to achieve the model proved to be not produce the desired results.

通过观察发现，翻译的句子无法区分从句中的宾语，即英语定语从句中的先行词（在上两句中分别为“包裹”和“模型”。而在正确的翻译中，这两个词应该分别处于宾语和主语的位置）。这样的翻译结果暴露了统计机器翻译模型固有的问题，即对句子结构分析能力差，这一点尤其体现在定语从句这类需要调序的句法结构上。如果我们能够初步解决从句的识别问题，分析并找出相应的中心词和依存词，将之转化为简单的句法分析，然后实现两步翻译，就有可能解决上述的问题。本文尝试利用条件随机场方法先对这类从句结构进行识别，然后利用中心词转录机的方法进行基于依存关系的句法成分调序同时进行自动翻译。

2 条件随机场和中心词转录机模型

条件随机场（Conditional Random Fields, CRF）是由 John Lafferty 等在文献[3]中提出的一种基于统计的序列标记识别和分割的无向图模型。模型的主要思想来源于最大熵[4]模型，是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率分布。

对于观察值序列 $O=(O_1, O_2, ..., O_i)$ 和状态序列 $S=(S_1, S_2, ... S_i)$，状态序列中的随机变量之间通过指示依赖关系的无向边连接，令 $C(S, O)$ 表示这个图中团的集合，CRF 将输出随机变量的条件概率定义为与无向图中的各个团的势函数（potential function）的乘积成正比：

$$P_A(S|O) = \frac{1}{Z_0}\prod\nolimits_{c\in C(s,o)} \phi_C(S_C, O_C) \qquad (1)$$

在给定一个输入序列的条件下，可以定义一个线性的 CRF 模型，形式如下：

$$P_A(S|O)=\frac{1}{Z_0}exp\left\{\sum_{i=1}^{i}\sum_{j=1}^{j}\lambda_k f_k\left(S_{i-1},S_i,O,i\right)\right\} \quad (2)$$

其中每个 f_k 是观察序列中输出节点的特征，而$Z_0=\sum_s exp\left(\sum_{i=1}^{i}\sum_k\lambda_k f_k\left(S_{i-1},S_i,O,i\right)\right.$是归一化因子。虽然常用的统计模型中还有隐马尔科夫模型（Hidden Markov Model, HMM）[5]和最大熵模型（Maximum Entropy, ME）。但是 HMM 是一种生成模型，需要做出严格的独立性假设。而事实上，大多数构成序列的随机变量都不能被表示成一系列独立的变量，所以 HMM 存在着一些固有的不足，而 ME 模型也有标记偏置的问题[3]。CRF 则综合了 HMM 和 ME 的特点，具有表达元素长距离依赖和交叠性的能力，能方便地融入上下文信息和领域知识，在命名实体识别方面已经取得了较好的效果[6][7][8]。

用于自动翻译的是文献[9]中描述的加权中心词转录机模型（Weighted Head Transducer）。一个加权中心词转录机是一个 5 元组：输入字符集 W；输出字符集 V；一个有限状态集合 Q 包含状态 q_0 ...q_s；一个最终状态集合 $F\subseteq Q$ 以及一个转移状态集合 T。从状态 q 转移至状态 q' 有下列形式：

$$<q,q',w,v,\alpha,\beta,c> \quad (3)$$

这里 w 是 W 的一个成员或者为空字符 ε；v 是 V 的一个成员或者为空字符 ε；整数α是输入位置；整数β是输出位置；实数 c 是转移的权重（或者称为代价）。当α=0 且β =0 时称为中心词转移（图 1）。

中心词转录机与标准转录机的区别在于其引入了读和写的位置α和β。为了解释读写位置的关系，Alshawi[9]引入了转录带的概念。一条转录带被分成了若干方格，其中一个方格标记为 0，其左右的各个方格分别向左标记为-1，-2...，向右标记为+1，+2...（图 2）。

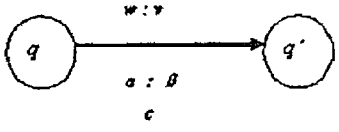


图 1 中心词转录机状态转移

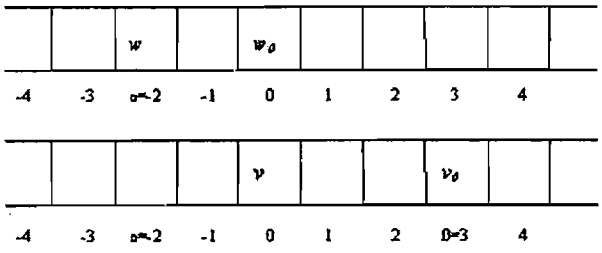


图 2 转录带和读写位置

从输入位置α和输出位置β开始的一个转移动作就是在输入带上的α位置读入字符 w，然后在输出带上的β位置写出字符 v；如果β位置已经被占据，如果β <0，则在左端下一个空格处写 v；如果$\beta\geq 0$，则在右端下一个空格处写 v。同样，如果在α位置处已经读入字符 w，如果α<0，则

在左端下一个空格处读入 w；如果$\alpha \geq 0$，则在右端下一个空格处读入 w。

在用于机器翻译时，转录机可推导出一系列依存树对。树中每一个节点的父节点都被视为中心词，子节点是中心词的依存节点。由依存转录模型推导的源语言和目标语言的依存树各个节点是有序的，经过对目标依存树进行简单的可回溯遍历就可以直接推导出目标句子。我们利用条件概率的形式表示状态转移的权重，对于中心词 w 和 v 及其依存节点 w'和 v'，其概率参数可以这样表示：

$$p(q', w', v', \alpha, \beta | w, v, q) \qquad (4)$$

这里 q 和 q'是一个转移动作的开始状态和结束状态，α和β 分别是源位置和目标位置，我们还需要参数 $p(roots(w_0, v_0))$作为选择 w_0, v_0作为中心词（根节点）的概率[2]。

3 特征选择和语料准备

由于汉语句子中各个语法单位分布的随机性，抽取合适的特征来反映这些复杂的语言现象显得十分重要。因此建立恰当而准确的特征模板是对汉语句子中定语从句结构进行准确识别的关键。此外特征融合的方法也非常重要。在 CRF 的模板中，可以包含一元特征、二元特征以及多元特征，还可以包含复合特征。

在决定汉语句子中的可译为定语从句结构的语法特征中，构成从句中的各个词和这些词的词性无疑是最重要的因素，因此在构建特征模板时首先引入词和词性信息。我们共选取了当前词及其前后两个词、当前词和前后两个词的词性组成一个特征集合，构成的模板如下：

表 2 词和词性的原子特征模板

序号	原子模板	意义
1	CurWord-2, CurWord-2_POS	当前词前 2 个词及其词性
2	CurWord-1, CurWord-1_POS	当前词前 1 个词及其词性
3	CurWord, CurWord_POS	当前词及其词性
4	CurWord+1, CurWord+1_POS	当前词后 1 个词及其词性
5	CurWord+2, CurWord+2_POS	当前词后 2 个词及其词性

在基于词性的基础上，我们又引进了针对汉语定语从句语法特点的一些语法信息。经观察发现，汉语定语从句结构中一定会包含动词、助词“的”和要译成英语定语从句先行词的名词结构（在例 1 中这三个词分别为“寄”、“的”和“包裹”）。所以我们把这三类词作为特征词分别在语法信息模板中进行标注，所形成的特征模板如下：

表 3 语法信息原子特征模板

序号	原子模板	意义
1	CurWord-2_H, CurWord-2_D, CurWord-2_C	当前词前 2 个词是否为动词、“的”、先行词
2	CurWord-1_H, CurWord-1_D, CurWord-1_C	当前词前 1 个词是否为动词、“的”、先行词
3	CurWord_H, CurWord_D, CurWord_C	当前词是否为动词、“的”、先行词
4	CurWord+1_H, CurWord+1_D, CurWord+1_C	当前词后 1 个词是否为动词、“的”、先行词
5	CurWord+2_H, CurWord+2_D, CurWord+2_C	当前词后 2 个词是否为动词、“的”、先行词

[2]本文中各个定语从句结构的中心词均已确定，所以这个概率值为 1。

表 4 词、词性和语法信息组合特征模板

序号	特征组合模板	意义
1	CurWord_H& CurWord_POS	当前词是否为动词和词性
2	CurWord_D& CurWord_POS	当前词是否为“的”和词性
3	CurWord_C&CurWord_POS	当前词是否为先行词和词性

我们把 1998 年一月份的《人民日报》分词并带词性标记的语料作为原始语料。在研究中我们发现一般来说含有定语从句结构的句子中不会包含标点符号，所以我们首先以标点作为断句标记把句子分成小的单位，每个都作为单独的一个句子进行识别。然后在原始语料中选取了带有可译为英语定语从句的句子 200 个，同时又选了 825 个不带有这种句子结构的句子作为训练语料；然后选择了带有定语从句的句子 60 个、不带有这种结构的句子 200 个作为测试语料。之所以分开选择，是因为我们计划分别对完全是带有定语从句结构的句子做一个识别实验（见表 6 实验一），再把带有这种结构和没有这种结构的两种句子混合起来再做一个识别实验（见表 6 实验二）。

根据上面总结的定语从句必须包含“的”、一个动词和先行词的语法特征，我们对所选语料制定了定语从句判别机制并作了相应的标记如下：

1）如果一个句子不包含“的”、动词或名词结构（语料中标记为：n, an, vn, q, ns, nr），不属于带有定语从句结构的句子；

2）含有定语从句的句子必须至少包含两个动词。如果这两个或多个动词相邻，如果第一个动词不是系动词[3]（“是”、“成为”、“变成”、“成了”、“作为”）或“有”，则这些动词合并视为连动结构，不属于含有定语从句结构的句子。把“的”字左邻的第一个动词标记为“H”；

3）判断句中的“的”是否属于定语从句。如果“的”字前无动词或后无名词结构，该“的”字不属于定语从句。否则把该“的”字标记为“D”；

4）判断句中的先行词。将判断为定语从句中的“的”字后 5 词窗口内的最后一个名词结构前的所有词均标记为“C”。

按照上述原则标记的包含识别结果的语料片段如表 5：

表 5 包含识别结果的语料片段

词	词性	语法信息	标记	识别结果
一	m	O	B*	B
场	q	O	I	I
征服	v	H	I	I
生命	n	O	I	I
禁区	n	O	I	I
的	u	D	I	I
战役	n	C	I	I
开始	v	O	O	O
了	y	O	O	O
,	w	O	O	O

*“B”是定语从句结构的开始标记，最后一个“I”是定语从句的结束标记。

[3] 一般认为，汉语不包含系动词概念，因为本文研究的是翻译，所以这里是指可译成英语系动词的汉语动词。

4 实验和结果分析

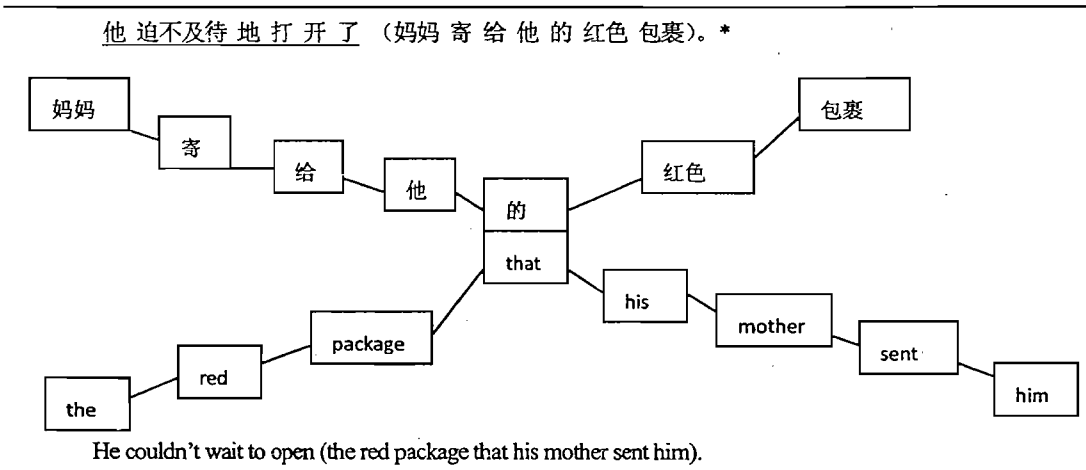
对于从句结构的识别结果我们通过计算准确率、召回率和 F 值作为评分标准。应用 CRF++工具包[10]经过两次实验表明（见表 6），在带有定语从句结构的句子中混入不带有这种结构的句子识别率会下降。定语从句结构的右边界的识别效果比较好，而出现错误的识别基本上都是在左边界。这是因为右边界的特征比较明显，而左边界的特征相对而言不是很明显。定语从句结构在句末（如例 1）的识别效果比较好，而在句首（如例 2）的识别效果就不尽理想，这主要是因为从分布上看大多数的定语从句结构都分布在句尾，训练的数据量比较大。

经观察发现，左边界识别错误主要分为两类：一是标点符号、一些属于停用词的词作为左边界；二是诸如助词、量词、连词等词性的词作为了左边界。所以我们制定了两条规则，凡是结果中以第一类符号或词作为左边界的结构，我们把左边界后移一个词；而以第二类错误中的词作为左边界的结构，我们把左边界前移一个词。我们把应用 CRF 进行识别的两个试验和把实验二经过规则修正以后的结果做了对比列表 6：

表 6 识别试验的结果

	准确率（%）	召回率（%）	F 值（%&β=1）
实验一	78.9%	75.0%	76.9%
实验二	75.0%	69.2%	72.0%
实验二修正后的结果	80.0%	73.8%	76.8%

在识别出可译为定语从句的结构后，把这些结构输入到中心词转录机中去，按照图 3 中的翻译模型进行翻译。其中会得到两个确定的句法单位和相应位置：一是子句中的助词“的”，另外一个就是右边界的词一定是英语定语从句中的先行词。我们以固定的“的”为中心词（译成英语句子中的连词“that”为中心词）进行依存语法分析[11]并输出翻译结果，即可得到汉语句子中对应英语定语从句的结构的翻译结果，如图 3：



*划线部分由统计模型翻译，括号内为识别的定语从句结构和翻译结果。

图 3 利用中心词转录模型对定语从句结构进行翻译

按照上述方法对例 1 和例 2 中的句子重新翻译，得到例 5 和例 6。我们发现定语从句先行词位于正确的位置上，句子顺序符合英语的习惯，结果得到了改善。

例 5　He can not wait to open the red package that mother sent him.

例 6　The model Ah-chiang realized with probability methods proved not produce the desired results.

5 未来工作

本文针对统计机器翻译模型在汉英翻译中处理嵌套句子结构时效果不佳的问题，以汉语需要翻译成英语定语从句的结构为例，提出了一种首先利用机器学习中的条件随机场进行子句识别，然后利用中心词转录机模型进行语法分析并翻译，最后基于统计进行全句翻译的方法，改善了句子整体的翻译效果。

但是由于这类语法结构在汉语中的分布的不对称性，从句所处的位置对于实验的结果影响比较大。此外本研究中语料规模还比较小，开放测试也需要在一定条件下进行。在利用中心词进行转录翻译时仍然要依赖统计方法，所以在翻译词的选择上仍然会出现错误，翻译的结果也存在不令人满意的地方，这些问题希望在今后的研究工作中能够得以改善。此外本文只针对汉语中一般要翻译成英语定语从句结构作了研究，希望今后能对其它类似的句法结构，如英语中的主语从句和宾语从句，做进一步深入的研究。

参考文献

[1]朱德熙.《语法答问》. 北京：商务印书馆, 1985.

[2]詹卫东.《面向中文信息处理现代汉语短语结构规则研究》. 北京：清华大学出版社，广西科学技术出版社, 2000.

[3]Lafferty, J., McCallum, A., Pereita, F.. Conditional random fields: probabilistic models for segmenting and labeling sequence data. The Journal of Machine Learning Research, 2001. ICML01: 282-289.

[4] Baum, L. E., Petrie, T.. Statistical inference for probabilistic functions of finite state Markov chains. Annual Mathematics Statistics. 1966, 37(6):1554–1563

[5]Berger, A., Pietra S. D., Pietra V. D.. A maximum entropy approach to natural language processing. Computational Linguistics, 1996, 22(1): 39-71.

[6]宗萍，施水才，王涛，吕学强. 基于条件随机场的英文地理行政实体识别. 现代图书情报技术, 2009(2): 51-55.

[7]周俊生，戴新宇，尹存燕，陈家骏. 基于层叠条件随机场模型的中文机构名自动识别. 电子学报,2006(5): 36-41.

[8]郭家清，蔡东风，王智超，刘浩公. 一种基于条件随机场的人名识别方法. 通讯和计算机, 2007(4): 22-25.

[9]Alshawi, H., Bangalore, S., and Douglas, S.. Learning dependency translation models as collections of finite-state head transducers, Computational Linguistics, 2000, 26(1): 45-60.

[10] http://chasen.org/~taku/index.html.en

[11]Alshawi, H. Head automata and bilingual tiling: Translation with minimal representations. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), 1996: 167–176.