融合语义角色特征的纳西汉语机器翻译方法[1]

丁砣，余正涛，高盛祥，苏萌，周枫
（昆明理工大学 信息工程与自动化学院，云南 昆明 650500）

摘要：为了有效利用纳西语句法特点及语义角色信息，提出一种融合语义角色特征的纳西-汉语机器翻译方法，该方法以依存树到串翻译模型为基础，构建了一个语义角色特征模型获得语义角色的重排序概率，将常用的特征和语义角色特征模型融合到对数线性模型，通过最小错误率来训练模型的参数。在解码过程中根据语义角色特征模型调整目标串的相对顺序。实验结果表明，融合语义角色特征的纳西汉语机器翻译方法有效提高了翻译的准确率。

关键字：语义角色标注；最大熵；依存树到串；机器翻译

A Method of Fusing Semantic Role Features for Naxi Chinese Machine Translation

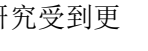
Wei Ding,Zhengtao Yu, Shengxiang Gao,Meng Su ,Feng Zhou
(Faculty of Information Engineering and Automation,Kunming University of Science and Technology, Kunming,Yunnan 650500, China)

Abstract: In order to utilize the Naxi syntactic features and semantic role information effectively, a method of fusing semantic role features is proposed for Naxi-Chinese machine translation. This method is based on the dependency tree to string translation model, constructing a semantic role feature model to obtain semantic roles reordering probability. Further more it integrates common features and semantic role models into the log-linear model, and trains parameter of the model by minimum error rate. In the decoding process, the relative order of the target string is adjusted according to the semantic role features model. The experimental results denominate that the fusion of the semantic role features of the Naxi Chinese machine translation method can effectively improve the accuracy of translation.

Key words: Semantic role labeling; Maximum entropy;Dependency tree to string; Machine Translation

1 引言

纳西语-汉语机器翻译对保护纳西文化遗产和促进纳西语和汉语之间的交流具有积极的作用。在机器翻译方面，经历了基于词和短语的机器翻译方法后[1-2]，基于句法的统计机器翻译成为主流，目前取得了很好的效果，如刘洋等提出的基于对齐模板的树到串翻译模型[3]，熊德意等提出的依存树到串翻译模型[4]，近几年，语义角色分析取得很多成果，如刘挺等提出了基于最大熵分类器的语义角色标注[5]，在此基础上，基于语义的机器翻译研究受到更多学者的关注，如张民等提出了基于双语合成语义的翻译相似度模型[6]。在纳西语机器翻译方面，目前已经取得了一些研究成果，如赵芳婷构建了纳西-汉语双语语料库及双语语料对齐[7]，程立提出了基于统计机器翻译的汉语-纳西语的谓词参数关联模型[8]，李磊提出了一种基于改进的依存树到串的汉语纳西语翻译模板抽取方法[9]。

尽管机器翻译领域取得了成果，但是如果直接运用到纳西-汉语机器翻译是不行的。因为纳西语与汉语在句法结构上有很大差异性，且其属于典型的“动居句尾”型语言，如：对纳西句子  （早

[1] **基金项目**：国家自然科学基金(No.61163022)，科技部科技创新人才基金项目(No.2014HE001)

晨我喝水），其中“  （早晨）”，“  （我）”，“  （水）”，“  （喝）”，将纳西的句法特点融合到纳西-汉语的机器翻译能够提高翻译效果[9]。另外，纳西语与汉语很类似，句子中存在系事、领事、受事等丰富的语义信息，这些语义信息在一定程度上对双语翻译具有一定的约束作用。如：上例中“  （我）”是施事 S，“  （水）”是受事 O，“  （喝）”是谓词 V，在纳西语端的语义角色的顺序是 S O V，而在汉语端却变成了 S“  （我）”，V“  （喝）”和 O“  （水）”，语义角色的顺序置换成了 S V O，这些语义角色信息对于目标语言的调序问题起着很重要的作用。

因此，基于纳西语语义的特点，探讨纳西语语义角色标注，并将语义角色融合到纳西语-汉语的机器翻译模型中以此来提高机器翻译的效果。

2 纳西语语义角色标注

根据纳西语的语义角色特点，借鉴中文、英文语义角色标注规范，并考虑到语义
角色划分的粗细粒度对分类效果的影响，将纳西语语义成分划分为十三类。具体类别如表 1 所示。

表 1 纳西语语义角色分类体系

名称	标记	名称	标记
施事	S	系事	X
受事	O	领事	L
客事	K	工具	I
对象	J	当事	D
时间	H	谓语动词	V
结果	R	标点符号	#
其他	QT		

在定义完语义角色的类型之后，通过汉纳双语词对齐句子语料，借助于词对齐特性，对纳西句子采用了基于条件随机场对纳西句子进行分词及词性标注[10]，然后利用纳西语依存句法分析器对每个节点进行谓词识别[11]，抽取依存角色到语义角色规则，通过规则实现依存关系到语义角色关系转换，并通过人工校对获得语义角色标注语料，标记了 10000 个纳西句子语义角色语料，利用最大熵分类器进行语义角色分类器构建，实现纳西语语义成分标注识别，该模型取得了很好的识别效果，其中在系事、领事、结果、客事、工具的识别准确率在 90%以上[11]。

3 纳西汉语机器翻译模型

3.1 依存树到串模型

使用基于依存树到串模型，在使用模型过程中，定义一个三元组（NDT,CS,A）作为纳西-汉语翻译模板。其中：NDT 表示纳西依存树片段；CS 表示对应的汉语语言串；A 表示 CDT 与 NS 的对齐关系。

三元组中的 A 元素，基于依存树到串的纳西-汉语翻译模板规定翻译对<D,S>要与整个句子对的对齐矩阵保持一致[12]。即：

$$\forall(i,j)\in M, i\in D \leftrightarrow j\in S \qquad (1)$$

一般来说，翻译对两边要全部互为对应，若没有对应关系的词则不生成翻译模板。

对于三元组中的 NDT 元素，则采用 Treelet 来代替子树[9]，当一个节点被抽取出来的时候，这个节点所有的与之相连的直接的兄弟节点也都会被抽取出来[3]。

3.2 语义角色特征模型

为了有效利用源语言端语义角色信息，并解决目标语言的调序问题，利用语义角色模型标注系事（X），领事（L），受事（O），谓词（V）等语义角色,并将源语言端语义角色所构成的序列定义为 SRQ-S，目标语言端的语义角色的序列定义为 SRQ-T。语义角色特征模型描述了 SRQ-S 映射到 SRQ-T 的概率，给定一个谓词 V，它可以表示成如下所示：

$$S(SRQ-S \mid SRQ-T, PRE=V) \qquad (2)$$

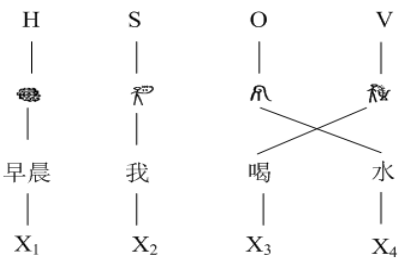


图 1 纳西语汉语语义角色对照图

语义角色特征模型是句中的语义角色重排序的概率。如图 1 所示：以谓词 V“[illegible]（喝）”，受事 O“[illegible]（水）”作为例子，在纳西语端的语义角色的顺序是 OV，O“[illegible]（水）”和 V“[illegible]（喝）”在汉语端置换了顺序，汉语端（SRS-T）的语义角色排序为 X_3X_4，重排序的概率为：

$$S_{S\text{-}A}(\text{S-T}=X_3X_4,\ \text{S-S=OV},\ \text{PRE=}(\text{[illegible]})) = \frac{Count(\text{S-T}=X_3X_4,\text{S-S=OV},\ \text{PRE=}(\text{[illegible]}))}{\sum_{\tau\in\Phi(\text{S-S})} Count(\text{S-T}=\tau,\text{S-S=OV},\ \text{PRE=}(\text{[illegible]}))} \quad (3)$$

其中，$\Phi(S-S)$ 指的是目标端所有可能的语义角色排序。

3.3 对数线性模型

对数线性模型不仅可以融合翻译模型，而且可以融合语言模型。在本文中还融合了语义角色特征模型。

$h_m(e,f)$ 是翻译模型的特征函数，λ_m 是 $h_m(e,f)$ 的权重，则机器翻译的目标便是求解特征函数参数以使翻译的概率最大，即使 $P(e|f)$ 最大。则对数线性模型可以表示为公式 4：

$$P(e|f)=\frac{\exp\left[\sum_{1}^{M}\lambda_m h_m(e,f)\right]}{\sum_{e'}\exp\left[\sum_{1}^{M}\lambda_m h_m(e',f)\right]} \quad (4)$$

在此框架下，设计一组特征函数 $h_m(e,f)$，其中 m=1,2,…,M。对于每个特征函数存在相应的模型参数 λ_m，其中 m=1,2,…,M。特征函数 $h_m(e,f)$ 包括常用的特征及语义角色特征模型。

由于（4）式的分母部分对最终的搜索结果没有影响，则机器翻译的搜索模型[4]为：

$$e_{best}=\arg\max_{\alpha}\left\{\sum_{m=1}^{M}\lambda_m h_m(e^*,f)\right\} \quad (5)$$

4 纳西汉语机器翻译

4.1 模板抽取

抽取模板的过程是在纳西语依存树上根据上下文依存关系获取目标语言中对齐词及关联的未对齐词，标记每个节点的属性。其中包括各节点语义角色的属性，使用翻译模板抽取器递归地在已经标注好的节点上抽取所有可能的模板。图 2 给出了依存树各节点对齐关系。

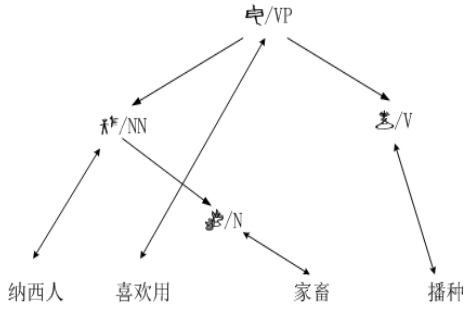


图 2 依存树各节点对齐关系

在抽取模板之前我们为依存树的每个节点标注了属性，利用依存树到串的纳西汉语翻译模板抽取算法即可递归地抽取模板，抽取到部分模板见表 2，图 3 给出了模板抽取算法，其中*代表泛化后的变量。

表 2 抽取出的部分翻译模板

Treelet	String
([illegible]/NN/0)	纳西人
([illegible]/N/0)	家畜
([illegible]/V/0)	播种
([illegible]/VP/0)	喜欢用
([illegible]/VP/0 (*/-1)) ([illegible]/N/0)	喜欢用 * 家畜
([illegible]/VP/0 ($*_1$/-1) ($*_2$/1))	喜欢用 $*_1$ $*_2$

/*算法说明
* 输入：标注好各种属性的依存树节点
* 输出：抽取出的翻译模板
* n 为某一个节点，T 为节点个数-1
*/

定义为 R,L 两个空栈
for i=0 to T do
　　if n.include != 1 then
　　　　标记其父亲节点位置，抽取翻译模　板加入 R 和 L
　　else if n.include ==1 then

```
        取栈 L 中父亲节点位置为 i 的元
    素  和 n.value 生成翻译模板将加入 R
        L(i) = "*"
            将 L(i).value 和 n.value 生成
        翻译模板将加入 R
      end if
    end if
  输出 R
end for
```

图 3 模板抽取算法

4.2 参数训练

使用 Och 提出的最小错误率训练算法（minimum error rate training，MERT）[13]来进行参数训练。翻译后的训练数据集是以译文分数的高低作为优化目标并使用 BLEU 值作为翻译的评测指标，在训练集中得到的分数越高则错误越少，由此调整模型参数。本文在训练语义角色特征的纳西汉语机器翻译方法的参数时，所选择的数据集是通过双语专家审核的双语句子。它的具体的训练过程是如下所示。

在训练集上总的错误数量可定义为：$E(s^s,t^s)=\sum_{i=1}^{S}(s_i,t_i)$。在实际训练中 $E(s_i,t_i)$ 定义为系统输出纳西句子与候选翻译的汉语句子相比的错误的个数。我们的目标是使得最终的机器翻译系统翻译出的汉语句子与训练数据集中的句子相比较起来，出现的错误的数量最少，即通过在这个训练数据集上，迭代模型参数 λ 来达到这个目的。

$$\lambda_M^M = \arg\min_{\lambda^M}\{\sum_{i=1}^{S} E(\hat{t}_i(\hat{s}_i;\lambda_M^M),s_i)\} \quad (6)$$

同时：

$$\hat{t}_i(\hat{s}_i;\lambda_M^M) = \arg\max_s\{\sum_{m=1}^{M}\lambda_m h_m(a,s,x)\} \quad (7)$$

其中 $S\in S^S$。

在迭代求解过程中，依照融合语义角色特征的纳西汉语机器翻译模型，使用特征函数的权重 λ 值去搜索翻译的前 N 个作为机器翻译最好的翻译结果，即 N-BEST 翻译列表。在特征函数权重 λ 的 K 维的搜索空间中，先设定 $K-1$ 维空间不变，选择一维空间对其权值进行迭代，当 N-BEST 不再变化时得到 N-BEST 列表。其中迭代的过程中的具体算法如下所示：

第一步：将特征函数的参数 λ^K 赋给初始值，接着根据这个初始值去搜索目标语言端的 N-BEST 列表。

第二步：如果迭代的特征模型参数是 λ_H^K，则用新的特征模型参数 λ_H^K 生成一个新的 N-BEST 列表，将新生成的 N-BEST 列表和第一步生成的 N-BEST 列表合并起来，这样可生成新的扩展 N-BEST 列表。

第三步：重复前面两个步骤，直到得到的 N-BEST 列表不再发生变化，这个模型的参数也就训练结束了。

4.3 解码

解码的过程就是搜索最适合推导，并产生最佳译文的过程。本文的解码过程采用了 Chart[14]算法，自底向上地翻译依存树上的每个节点，查找使用当前的节点为根节点匹配的 Treelet 树对，然后对没有被 Treelet 树对覆盖的节点，根据语义角色特征模型调整其相对顺序。

通过往对数线性翻译模型中增加一个新的特征来融合语义角色特征模型。我们定义几个函数：

$A(i,j,S)$：纳西语端如果 S 的参数全部处在[i,j]范围内，则返回 S 的参数。

$B(i,j,S)$：如果 S 处在[i,j]范围内，则返回 true，否则返回 false。

$C(S_x,S_y)$：如果 S_x 和 S_y 的语义角色排序还没有计算，则返回 true，否则返回 false。

其中，S 表示一个语义角色,包含 X 系事，L 领事，O 受事，V 谓词；$A(i,j,S)$ 表示一个假设。

$S_{S_x-S_y}(H,S_x,S_y)$ 据公式（2），返回被假设 H 覆盖的 S_x,S_y 的语义角色排序的概率。根据 H 的翻译求导，汉语端 S_x,S_y 的位置关系能被检测出来。

当一个新的假设 H 生成，通过算法将语义角色排序模型的概率计算融合到 Chart 解码器中。给定一个假设 H，首先寻找被 H 覆盖的语义角色。然后分别计算语义角色序列排序的概率。

在解码的过程中，将语义角色特征的概率融合到对数线性模型中，使用模型计算相关的模型的参数，直到整个句子完成推导，选择分数最高的翻译作为最后的译文。

5 实验及结果分析

5.1 实验数据集准备

为了进行纳西-汉语机器翻译方法的评测，我们从小学课本、初中课本收集并整理了 53000 对纳西-汉语平行句对，从中选取了 42000 对作为训练集，选取了 8000 对作为开发集，选取了 3000 对作为测试集。具体的实验数据如表 3 所示：

表 3 实验数据情况

数据集		中文	纳西文
训练集	句数	42000	
	词数	331524	21577
开发集	句数	8000	
	词数	68312	40448
测试集	句数	3000	
	词数	25431	45912

5.2 实验结果及分析

为了验证融合语义角色特征信息的基于依存树到串的纳西汉语机器翻译方法的效果。我们设计了纳西汉语翻译对比实验。实验系统以短语模型、依存树到串的机器翻译系统为基准进行对比实验。采用基于条件随机场的纳西分词工具对和 ICTCLAS 汉语分词工具对纳西-汉语双语句子进行分词[10]，使用纳西句法分析器获得纳西句法分析树[11]。采用训练获得语义角色标注模型对纳西语言进行语义角色识别。为了对比提出翻译方法的效果，选取基于短语的纳汉翻译系统、基于依存树到串的纳汉句法翻译系统作为基准，与融合语义角色特征的纳汉句法翻译系统进行对比实验，使用 BLEU4 作为实验衡量标准。实验结果见表 4。

表 4 不同纳汉翻译方法对比实验结果

翻译系统（纳-汉）	开发集	测试集
短语模型	21.27%	21.35%
依存树到串模型	22.41%	22.53%
融合语义角色特征的依存树到串模型	23.96%	23.58%

从表 4 实验结果可以看出，基于依存树到串的模型比短语模型效果好，其主要原因是考虑了纳西语的句法的特点。融合语义角色特征的模型比短语模型和依存树到串模型效果好，其中比依存树到串 BLEU 值提高了 1.55，比短语模型 BLEU 值提高了 2.69，分析原因，主要是因为融合了纳西语语义角色特征提高了句法结构树的调序能力，语义角色的约束提高了纳西汉语机器翻译的准确率。

6 总结

语义角色对句子理解有非常重要的作用，本文针对纳西语特点，在树到串句法统计翻译的基础上，融合了纳西语语义角色，实验也验证了提出方法的有效性，能够有效提高翻译的准确率。进一步的研究将主要集中在提取篇章级上下文语义关系帮助提高翻译效果。

参考文献：

[1] Peter F.Brown,John Cocke,Stephen Della Pietra, et al. A Statistical Approach to Machine Translation[J]. Computational Linguistics. 1990, 16 (2):79-85.

[2] Franz Josef Och.Statistical Machine Translation: From Single-Word Models to Alignment Templates[J]. The publications of the Department of Computer Science of RWTH Aachen. 2002.

[3] Yang Liu,Qun Liu,Shouxun Lin. Tree-to-String Alignment Template for Statistical Machine Translation[C].//In proceedings of the ACL 2006. Syendy,2006: 379-386.

[4] Deyi Xiong,Qun Liu,Shouxun Lin.A Dependency Treelet String Correspondence Model for Statistical Machine Translation[C].//In proceedings of the Second Workshop on Statistical Machine Translation.Beijing,China 2007: 40-47.

[5] 刘挺，车万翔，李生.基于最大熵分类器的语义角色标注[J].软件学报，2007, 18(3): 565-573.

[6] 王超超，熊德意，张民.基于双语合成语义的翻译相似度模型[J].北京大学学报(自然科学版),2014,51(2):335-341.

[7] 赵芳婷，余正涛，线岩团，等.纳-汉双语语料库构建及双语语料对齐[J].广西师范大学学报(自然版)，2009, 27(1): 161-164.

[8] Li Cheng,Zhengtao Yu,Jianyi Guo,et al. Predicate-argument Relevance Model for Chinese-to-Naxi SMT [J]. 2015, 11(13): 4857-4861

[9] Lei Li,Zhengtao Yu,Cunli Mao. The Extracting Method of Chinese-Naxi Translation Template Based on Improved Dependency Tree-To-String[C].//Lecture Notes in Computer Science. Zhengzhou,China.2013,8229(3): 350-358

[10] Xiuzhen Yang,Zhengtao Yu,Jianyi Guo,et al.Naxi-Chinese Bilingual Word Alignment Method Based on Entity Constraint[C].//In Proceedings of 14th Workshop on Chinese Lexical Semantics.Zhengzhou,China.2013:378–386

[11] 苏萌. 融合语义角色特征的纳西汉语机器翻译研究 [D].昆明:昆明理工大学.2015:28-44.

[12] Huang L, Knight K, Joshi A. Statistical syntax-directed translation with extended domain of locality[C].//Proceedings of AMTA.2006: 66-73.

[13] Michel Galley,Mark Hopkins,Kevin Knight, et al. What’s in a translation rule[C].//In Proceedings of HLT/NAACL. Japan.2004:273-280.

[14] Dekai Wu.Stochastic inversion transduction grammars and bilingual parsing of parallel corpora[J]. Computational Linguistics.1997,5(2): 377-403.