文章编号：1003-0077（2011）00-0000-00

融合被动和可能态模型的日汉统计机器翻译*

王楠，徐金安†，明芳，陈钰枫，张玉洁

（北京交通大学 计算与信息技术学院，北京 100044；† 通讯作者，E-mail: jaxu@bjtu.edu.cn）

摘要：日语中谓词语态有不同的词尾变形，其中被动态和可能态具有相同的词尾变化，在统计机器翻译中难以对其正确区分及翻译。因此，本文提出一种利用最大熵模型有效地对日语可能态和被动态进行分类，然后把日语的可能态和被动态特征有效地融合到对数线性模型中改进翻译模 型的方法，以提高可能态和被动态翻译规则选择的准确性。实验结果表明，该方法可以有效提升日语可能态和被动态句子的翻译质量，在大规模日汉语料上，最高翻译 BLEU 值能够由 41.50 提高到 42.01，并且在人工评测中，翻译结果的整体可理解度得到了 2.71%的提升。

关键词：被动态；可能态；统计机器翻译；最大熵模型

中图分类号：TP391　　**文献标识码：**A

Integration of Passive and Active Voice Model into Japanese-Chinese Statistical Machine Translation

WANG Nan，XU Jin'an†，MING Fang，CHEN Yufeng ，ZHANG Yujie

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China;
†Corresponding Author, E-mail: jaxu@bjtu.edu.cn)

Abstract: The suffixes of Japanese predicates have complex formation of different voice. Both passive and potential predicates are formed with the same suffix which originated from the same stem. This phenomenon caused mistranslation in statistical machine translation. In this paper, a new method has been proposed for rule selection among different voice. Maximum entropy models were built which can effectively classify passive and potential voice, and integrate voice features into log linear model to improve translation model. In Japanese to Chinese translation task, large scale experiment shows that our approach improve the translation performance from 41.50 to 42.01 BLEU, and the intelligibility is 2.71 higher via human evaluation methods.

Key words: Passive Voice; Active Voice; Statistical Machine Translation; Maximum Entropy Models

1 引言

日语的“态”是一种语言现象，指在谓语后加接尾辞，且补充语的格规则也相应发生变化的情况。被动态与可能态是日语语态中的两种典型形式。表示主体受到另一事物的动作时使用动词的被动态，而在表示主体具有某种能力、有条件进行某种行为时，使用动词的可能态。这两种语态多数情况是由动词的未然形后加词尾“れる”“られる”构成的。如下例“吃”在日语中对应的动词未然型是“食べる”，其语态变化如表 1 所示。

现有的研究大部分从语义及结构上进行日语被动态与可能态的区分[1]。但在机器翻译任务中，很难有效区分这两种语态，研究人员通过制定翻译规则对不同语态进行处理[2,3]。但基于规则的翻译系统存在规则主观性较强、对语种具有依存性且领域适应能力差等问题，故研究逐渐转向更易于推广使用的统计机器翻译方法。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61370130，61473294）；中央高校基本科研业务费专项资金资助（2015JBM033）；国家国际科技合作专项资助（No. 2014DFA11350）

表 1　动词“食べる”的不同语态

Table 1　The different voice of “食べる”

类别	动词变化	源语言句子	参考译文
未然型	食べる	私はリンゴを食べました	我吃了苹果
可能态	食べられる	私はリンゴが食べられますか	我能吃苹果吗
被动态	食べられる	リンゴは私に食べられた	苹果被我吃掉了

传统统计机器翻译方法利用统计翻译模型和语言模型进行翻译。层次短语翻译模型是当前应用最广泛的翻译模型之一，其中源语言和目标语言所构成的翻译规则通常具有一对多的关系，当解码器按照概率进行规则选择时，由于训练语料中可能态和被动态的数据稀疏问题严重、远距离调序相对困难、难以有效利用句子的全局结构特征实现全局优化，导致其翻译精度低下等问题。因此在翻译过程中，如何正确选择翻译规则，实现语义消歧和调序的优化，进而实现句子的全局优化问题，是系统优化的关键。

近年来关于构建语态模型的统计机器翻译研究不多，但很多研究者通过构建分类模型提高规则选择准确率。Xiong 等[4]选取短语边界词信息构建最大熵模型运用于短语排序，He 等[5]利用非终结符的边界词信息建立最大熵规则选择模型，Nguyen 等[6]利用最大熵模型将位置信息等词汇化特征融入层次短语模型。Iglesias 等[7]使用非终结符的数目和类型对层次短语规则进行分类解决规则选择问题。利用分类模型融合层次短语模型中缺失的上下文信息，可以有效提升翻译质量，但这种方法的不足在于词汇化信息仍然缺少语言学句法的约束。

同时很多研究者引入语言学分析改进层次短语模型。Shen 等[8]使用目标端的句法依存树信息拓展层次短语模型，过滤了大量规则。Čmejrek 等[9]对双语语料进行解析后直接抽取层次短语规则，但没有对冗余规则进行处理。Gao 等[10]使用源端句子的依存结构限制调序提升了翻译性能，这些研究成果表明，将语言学分析融合入翻译系统中可以有效地辅助翻译过程。

本文在总结以上方法的基础上，提出一种把日语的可能态和被动态特征融合入翻译模型的方法。首先把语料分为被动态、可能态和其他语态三类，抽取相应的句法特征构建最大熵分类模型，并对其进行有效的分类。然后在抽取层次短语规则时同步抽取语态特征，使用最大熵模型将语态特征融合入翻译模型。最终构建出可能态和被动态的翻译模型以提高这两种语态的翻译精度。该方法不仅使用最大熵模型融合了丰富的上下文信息，克服了层次短语模型中无法利用上下文信息的缺点，而且引入语态特征这一语言学约束指导解码器根据不同语态选择合适的规则。实验表明，该方法获得了 0.1~0.5 的 BLEU 值的提升，在人工评测中翻译结果的整体可理解度也得到了 2.71%的提升。

本文组织结构如下：第二章介绍层次短语翻译模型，第三章对本文提出的融合被动和可能态的翻译模型做出具体的阐述，第四章描述实验设置及实验结果，并针对实验结果分析本文提出方法的有效性。最后对本文进行总结和展望。

2 层次短语模型

层次短语（Hierarchical Phrase Based）模型[11,12]可以从双语句对中自动地抽取形式语法，不需要语言学上的标注和假设，是当前性能最好的统计机器翻译系统之一。

2．1 规则抽取

层次短语模型使用上下文无关文法（SCFG）规则进行翻译，其规则形式如式(1)所示：

$$X \to \langle \alpha, \gamma, \sim \rangle , \tag{1}$$

其中，X 是非终结符，α 和 γ 分别为规则的源语言目标语言端，包含终结符和非终结符，非终结符的对应关系由 ~ 表示。

层次短语规则的抽取过程如下：基于双语语料的词对齐信息，按照从左至右的顺序抽取短语规则。之后利用子短语替换短语规则，从而得到形式化的句法关系。虽然这种句法关系

简化了建模和解码，但规则在泛化的过程中没有保留上下文信息，导致了子短语可以匹配任何的句法成分，在翻译时往往会产生错误。

2．2 翻译模型

层次短语翻译系统翻译过程可以描述为对于给定的源语言句子 f，从所有可能的翻译结果 e 中，找到得分最高的翻译结果。层次短语翻译系统在翻译过程中使用对数线性模型，其中组合了多个特征。对数线性模型每进行一次转换，都会计算之前步骤的得分总和。公式(2)为对数线性模型中的转换得分，通常使用对数的形式表示，如公式(3)所示：

$$P(\mathrm{d}) = \prod_{i}^{M} \phi_i(\mathrm{d})^{\lambda_i} \ , \tag{2}$$

$$score(\mathrm{d}) = \log \mathrm{P}(\mathrm{d}) = \log \prod_{i}^{M} \phi_i(\mathrm{d})^{\lambda_i} = \sum_{i}^{M} \lambda_i \log \phi_i(\mathrm{d}) \ , \tag{3}$$

其中 ϕ_i 为特征函数， λ_i 为对应的特征权重，d 表示每一步的翻译过程。在层次短语翻译模型中使用了以下特征：正反向翻译概率， $P(e|f)$ 和 $P(f|e)$ ，正反向词汇化权重， $P_w(e|f)$ 和 $P_w(f|e)$ ，N 元语言模型， $p_{lm}(f)$ ，规则数量惩罚， $\exp(-1)$ ， 长度惩罚， $\exp(|f|)$ 。解码器利用对数线性模型将上述特征组合，使用 CYK 形式的算法，利用抽取出的层次短语规则对测试集句子进行翻译。

3 融合最大熵特征的翻译模型

3．1 翻译系统结构

本文把语态分为被动态、可能态和其他语态三类，通过最大熵模型把语态信息融合入翻译模型中。融合语态特征的翻译系统流程如图 1 所示。

首先对语料进行分类，人工抽取筛选出语料中的被动态与可能态句子，剩余句子归为其他语态；然后对训练语料进行句法分析，抽取出不同语态的特征以训练最大熵模型；在规则抽取过程中抽取出相应特征，通过最大熵模型将语态特征融合到规则表中生成不同语态的翻译模型。最后在翻译过程中，首先判断输入句子的语态，根据语态选择相应的翻译模型，实现在解码过程中的规则自动过滤。

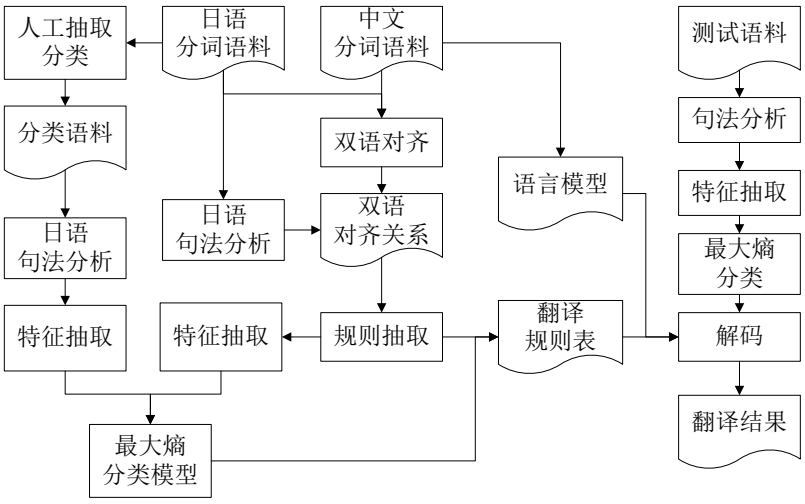

图 1　融合语态特征的翻译系统流程

Fig.1 Flow chart of Japanese-Chinese Translation System with Voice Features

本文主要论述基于最大熵模型的层次短语规则分类、分类特征的选择及最大熵模型与翻译模型的融合，对翻译过程只做简单叙述。

3．2 最大熵规则分类

最大熵（Maximum Entropy）模型能够满足所有已知的约束，对未知信息不做假设，可

以方便地融合多种上下文信息作为语态特征，因此本文选取最大熵模型作为分类模型。假设存在样本集合 $\mathrm{T}=\{(x_1,V_1),(x_2,V_2),...,(x_n,V_n)\}$，其中 $x_i(1\le i\le n)$ 是一个句子的上下文环境，$V_i(1\le i\le n)$ 表示句子的语态类别。最大熵的约束是通过特征函数实现的，对于句子语态分类问题，定义如下特征函数：

$$f(x,V)=\begin{cases}1 & (x=\text{词})\wedge(V=\text{句子语态类别})\\ 0 & \text{其他}\end{cases}, \tag{4}$$

建立语态的最大熵模型如公式(5)所示：

$$\mathrm{P}=\arg\max_{p\subseteq C}H(\mathrm{P}), \tag{5}$$

其中 $H(\mathrm{P})$ 是模型 P 的熵，C 是满足条件约束的模型集合。在给定文本集合和相关约束条件下，存在一个唯一概率模型 P*，其熵值最大，如公式(6)所示：

$$\mathrm{P}(V\mid x)^*=Z(x)\exp(\sum_i\lambda_i f_i(x,V)), \tag{6}$$

$$Z(x)=\frac{1}{\sum_V\exp(\sum_i\lambda_i f_i(x,V))}, \tag{7}$$

其中，$Z(\mathrm{x})$ 是归一化常数，f_i 即为模型特征，λ_i 是模型的参数，即特征函数的权重。通过在训练集上学习可以得出 λ_i 的具体值。

上述公式描述了句子的最大熵概率模型。对于每一条包含核心动词的层次短语规则 $<\alpha,\gamma>$，可以构建以下最大熵语态分类模型：

$$P(V\mid\alpha,\gamma,f(\mathrm{X}_k))=Z(\mathrm{x})\exp[\sum_i\lambda_i f_i(\mathrm{V}(\alpha),f(\mathrm{X}_k))], \tag{8}$$

$$Z(x)=\frac{1}{\sum_{\gamma'}\exp[\sum_i\lambda_i f_i(\mathrm{V}(\alpha),f(\mathrm{X}_k))]}, \tag{9}$$

其中，α 为规则的源语言端，γ 为目标语言端，V 是规则对应的语态类别。X_k 表示其中包含的非终结符。一条规则中可能含有多个非终结符，k 为非终结符对应的编号。非终结符 X_k 中的源语言子短语为 $f(\mathrm{X}_k)$，$V(\alpha)$ 表示源语言短语中上下文的语态信息，$f_i(\mathrm{V}(\alpha),f(\mathrm{X}_k))$ 是一个二值的特征函数，λ_i 为该函数的特征权重。

3．3 特征选择及规则抽取

特征函数的选取直接影响着分类性能。日本学者 Kurohashi[13]利用大量网络资源构建了较为完备的日语格框架库，并运用到句法分析中。Murata 和 Sasano[14,15]从大规模语料中抽取格框架特征，完成将被动态语句转化为主动态的任务。其中对于被动态句子识别的精度很高，说明格框架可以有效区分句子的语态。对于上述最大熵规则分类模型，规定以下特征：

句子的中心结构词特征 F1。日语句中谓语动词及词尾，即句法分析树的根节点信息。

句子主干结构特征 F2，源语言端句法分析树的第一层节点，即中心谓词的格框架信息。被动态与可能态句子的谓语有区别于其他语态的变形，但许多动词的被动态和可能态具有相同的形式，因此需要引入句子的结构特征进行区分。

下面以例句：“地下鉄で私の財布は憎らしい泥棒に盗まれました（在地铁上我的钱包被可恶的小偷偷走了）”说明，如何抽取句子及句子生成的层次规则中的对应特征。

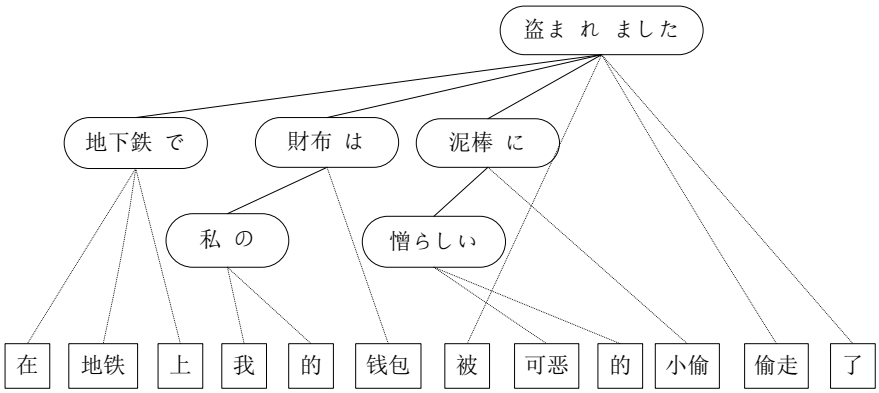

图 2　日语依存句法树示例

Fig.2 An example of Janpanese chunk-based dependency analysis result

首先在抽取前进行句法分析及标注。根据标注结果抽取句子特征如表 2 所示：

表 2　例句中的特征抽取

Table 2　An example of fether extraction

特征类型	特征值
F1	盗ま れ ました
F2	地下鉄 で 財布 を 泥棒 に

以句中短语示例，抽取层次短语规则时，把对应的最大熵特征也抽取出来。首先从词对齐关系中抽取到下面三个短语规则：

X → <れ，被>

X → <泥棒 に，小偷>

X → <泥棒 に 盗ま れ，被 小偷 偷>

由上述规则可以得到含有两个非终结符的层次短语规则：

X → <X1 盗ま X2， X2 X1 偷>

该规则中泛化的部分中含有根节点，即谓语动词或词尾信息，需对该规则进行特征抽取，即抽取其完整的根节点信息（抽取范围包括规则、非终结符和边界词）及非终结符中的句子结构信息。抽取上述规则的最大熵特征如表 3：

表 3　规则中的特征抽取

Table 3　An example of feature extraction in rules

特征类型	特征值
F1	盗ま れ ました
F2	泥棒 に

3．4 翻译模型融合

将抽取出规则的最大熵特征使用最大熵模型进行分类，得出三种语态的最大熵概率值 $P(V_1), P(V_2), P(V_3)$，分别对应被动态、可能态和其他语态。规则表中还有一类不包含语态信息的规则，包括没有非终结符的规则（即短语规则）和非终结符中不包含句子中心节点的规则。实验中将这类规则归到其他语态类别，即 $P(V_1)=0,\ P(V_2)=0,\ P(V_3)=1$。

然后将语态特征加入到上一节介绍的翻译模型中，最终生成三个翻译模型。例如将 $P(V_1)$ 加入规则表生成被动态规则表，同理，分别加入 $P(V_2)$、$P(V_3)$ 生成可能态和其他语态的规则表。最终每个单独的翻译模型包含以下特征：正反向翻译概率、正反向词汇化权重、N 元语言模型、规则数量惩罚、长度惩罚，及语态特征 $P(V)$。新增的特征与原有特征地位相同，其权重可以在权重调优的阶段一并进行调节。

通过直接在翻译模型中加入特征的方法，既保留了层次短语模型原有的特征，同时也融

入了新的规则的语态特征，没有增加解码算法的复杂度。在解码阶段，首先对输入的句子进行语态分类，根据分类结果选择不同的翻译模型进行翻译。

4 实验

4．1 实验及工具准备

本文数据来源于从网页端抽取整理的 50 万句日汉日常会话信息，并人工分类抽取出其中的被动态语句及可能态语句。语料相关信息如表 4 所示：

表 4　实验所用语料信息

Table 4 Corpus of experiment

类别	训练集	开发集	测试集
被动态	8292	489	512
可能态	41239	507	509
其他语句	473942	502	500
全部数据集	523473	1498	1521

本文使用 Juman[1]、KNP[2]作为对日语分词及句法分析的工具，使用 stanford-chinese-segmenter[3]工具对中文句子进行分词。使用张乐博士的最大熵工具包[4]作为分类工具。词对齐信息由 GIZA++[5]获得，在目标端句子上使用 SRI 语言模型工具[6]训练出 5 元语言模型。基于东北大学 NiuTrans 统计机器翻译系统[16]进行层次规则的抽取和解码，翻译质量的评价指标为 BLEU-4[17]，最后由 5 名同学对翻译结果进行了人工评测。

4．2 测试集句子分类

首先测试最大熵模型的语态分类效果。因测试集在翻译前需要进行语态识别和分类，最大熵模型的分类准确率直接影响到翻译效果。训练语料使用翻译训练集的 50 万句日汉日常会话语料。从翻译开发集和测试集中抽取 1500 句作为分类测试集，其中 500 句为被动态句子、500 句为可能态句子，剩余 500 句为其他语态。对训练集及测试集进行句法分析后抽取句子的中心结构词 F1 和句子主干结构 F2 训练并测试最大熵模型。使用准确率对语态的识别进行评价。因 F2 为辅助特征，仅仅加入 F2 识别效率非常低，故不作对比。实验结果如表 5 所示，结果表明加入句子的主干结构特征和中心词特征可以有效识别被动态和可能态。

表 5　最大熵分类实验结果

Table 5　The result of maximum entropy classification

测试集	被动态	可能态	其他	全部
+F1	89.4%(447/500)	94.8%(474/500)	97.6%(488/500)	93.93%(1409/1500)
F1+F2	97.8%(489/500)	98.2%(491/500)	99.4%(497/500)	98.47% (1477/1500)

4．3 翻译实验结果

实验中使用已分好类别的测试集进行翻译，这样可以排除分类错误的句子对翻译结果的影响，更好的分析系统性能。BLEU 实验结果如表 6 所示。

仅仅加入 F1 时被动态和其他语态的 BLEU 值较基线系统略有下降，可能态的 BLEU 值上升，但变化幅度很小。对比加入全部特征的翻译结果，仅加入 F1 时部分歧义问题没有得到改善。原因在于仅加入中心词无法有效地区分出被动与可能态谓语动词变形相同的情况。

---

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[3] http://nlp.stanford.edu/software/segmenter.shtml
[4] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html
[5] https://code.google.com/p/giza-pp/
[6] http://www.speech.sri.com/projects/srilm/

表 6　不同测试集的 BLEU 值

Table 6　BLEU-4 scores on the test sets

翻译	被动态测试集	可能态测试集	其他语态测试集
Baseline	42.60	41.50	39.58
+F1	42.13	41.56	39.43
+F1+F2	42.69	42.01	39.71

由于 BLEU 值不能完整地体现语态信息的翻译效果，所以本文以《机器翻译评测大纲》中人工评测规范为标准，对加入全部特征后的测试集翻译结果进行了人工评测。句子评分根据可理解度取 0.0-5.0 分不等，可含一位小数，最后得分是所有打分的算术平均值。最后采用公式(10)使用百分制换算评测结果，

$$\text{总的可理解度} = \text{所有句子得分之和}/\text{总句数}/5\times 100\% \tag{10}$$

本文仅对加入全部特征的最优结果进行了人工评测，评测结果如表 7 所示。

表 7　不同测试集的人工评测

Table 7　The result of human evaluation on test sets

翻译	被动态测试集	可能态测试集	其他语态测试集	全部测试集
Baseline	69.70%	71.02%	67.39%	69.37%
+Maxent Feature	72.44%	74.13%	69.68%	72.08%

分析实验结果可知，本文方法相较于层次短语模型，在被动测试集上 BLEU 值提升 0.09，在可能态测试集上 BLEU 有 0.51 的提高，且没有影响到其他语态的翻译。由于 BLEU 采用 n-gram 的完全匹配，针对语态的处理对 BLEU 值的影响不大。分析人工评测结果可知，本文方法在被动态、可能态及其他语态测试集上相比于基线系统可理解度均有 2.29%~3.11%的提高，在可理解度上优于基线系统。

4．4 实验结果分析

对比实验结果发现，相比传统的层次短语翻译模型，加入语态特征的翻译模型在翻译时消去了部分短语在规则选择时的歧义。如表 8 所示例句，基线系统选择了错误规则，翻译出了可能态的含义。而融合句法特征后，解码器在翻译被动态句子时正确选择了被动态的规则，对“把”字与主语的位置也进行了正确的调序。 表 9 所示例句是可能态，在解码过程中选择了可能态的规则进行翻译，没有丢失词汇信息，相较于基线系统得到了更好的翻译结果。

表 8　被动态句子翻译结果

Table 8　Translation of a passive sentence

原句	[**彼 が**]$_1$ 手紙 を 渡して [**くれ**]$_2$ ました 。
参考译文	**[他]$_1$ [把]$_2$** 信 交 给 我 了 。
Baseline	能 **[把]$_2$** 信 交 给 **[他]$_1$** 了 。
+Maxent Feature	**[他]$_1$ [把]$_2$** 信 交 给 了 。

表 9　可能态句子翻译结果

Table 9　Translation of a potential sentence

原句	右 側 の 曲がり かどに **[見え ます]$_1$** 。
参考译文	在 右侧 拐角处 [**能 看见**]$_1$ 。
Baseline	右边 那个 拐角处 有 转 。
+Maxent Feature	右 转 然后 在 拐角处 有 [**可以 看 到**]$_1$ 。

对实验结果进行分析时发现，日语被动态句子对应的参考译文具有不同的语序。部分日

语被动态句子对应的参考译文是被动句或把字句，另一部分对应的参考译文是主动句。如表10 所示例句，在日语中属于被动态，但是对应的参考译文不是被动句。相比于翻译成“连续记录片在每天七点被直播”，这里使用主动句更符合语言习惯，在没有融合语态信息时BLEU 值更高。在融合语态特征的翻译模型中，对于被动态的不同表达形式无法进行判断分类，是实验中被动态测试集的 BLEU 值没有明显提升的主要原因。

表 10　被动态句子翻译结果

Table 10　Translation of a passive sentence

原句	連続 ドキュメンタリー は 毎晩 七 時 に 放送 さ れた。
参考译文	连续 记录片 在 每 晚 的 七 点 播放 。
Baseline	连续 记录片 在 每天 七点 直播 了 。
+Maxent Feature	连续 记录片 在 每天 七点 都 被 直播 了。

5 结语

本文提出了一种提高日汉统计机器翻译的可能态和被动态的方法，该方法可以提高规则选择的准确性。首先，针对被动态与可能态句子构建分类模型，利用最大熵模型把句子的语态特征融合入层次短语翻译模型中，实现在解码过程中对不同语态规则的自动过滤。最后，实验结果显示，本文提出的方法可以有效提高翻译质量。

今后的工作主要包括：如何有效解决学习数据的不平衡问题，提高分类精度和翻译性能；尝试把可能态和被动态的翻译方法融合到其他统计翻译模型中，如树到串模型；其次，在模型中加入句子的句法结构及句子的全局特征、双语特征，提高翻译精度；再次，尝试融合神经网络语言模型以提高翻译精度。

参考文献

[1] Nakamura H. Two Types of Complex Predicate Formation: Japanese Passive and Potential Verbs[C]//Proceedings of the Pacific Asia Conference on Languages, Information, and Computation. 2007: 340-348.

[2] Alam Y S. A Rule-based Morpho-semantic Analyzer of the Japanese Verb Phrases of Simple Sentences[C]//PACLIC. 2008: 101-112.

[3] 卜,朝暉, 浅井,良信, 王,軼謳, et al. 日中機械翻訳における構文上の対応のずれに関する考察 : 受動態と能動態のずれ、品詞のずれを中心に(翻訳)[J]. 情報処理学会研究報告:自然言語処理研究会報告, 2006, 2006(124):33-40.

[4] Xiong D, Liu Q, Lin S. Maximum entropy based phrase reordering model for statistical machine translation[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 521-528.

[5] He Z, Liu Q, Lin S. Improving statistical machine translation using lexicalized rule selection[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008: 321-328

[6] Van Nguyen V, Shimazu A, Le Nguyen M, et al. Improving a lexicalized hierarchical reordering model using maximum entropy[J]. MT Summit XII, Ottawa, Canada, August, 2009.

[7] Iglesias G, de Gispert A, Banga E R, et al. Rule filtering by pattern for efficient hierarchical translation[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 380-388.

[8] Shen L, Xu J, Weischedel R M. A New String-to-Dependency Machine Translation Algorithm with a Target

Dependency Language Model[C]//ACL. 2008: 577-585.

[9] Čmejrek M, Zhou B, Xiang B. Enriching SCFG rules directly from efficient bilingual chart parsing[C]//Proceeding of the International Workshop on Spoken Language Translation. 2009: 136-143.

[10] Gao Y, Koehn P, Birch A. Soft dependency constraints for reordering in hierarchical phrase-based translation[C]//proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011: 857-868.

[11] Chiang D. A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 263-270.

[12] Chiang D. Hierarchical Phrase-Based Translation[J]. Computational Linguistics, 2007, 33(2):201-228.

[13] Kawahara D, Kurohashi S. Case frame compilation from the web using high-performance computing[C]//Proceedings of the 5th International Conference on Language Resources and Evaluation.2006:1344-1347.

[14] Murata M, Shirado T, Kanamaru T, et al. Machine-learning-based transformation of passive Japanese sentences into active by separating training data into each input particle[C]//Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, 2006: 587-594.

[15] Sasano R, Kawahara D, Kurohashi S, et al. Automatic Knowledge Acquisition for Case Alternation between the Passive and Active Voices in Japanese[C]//EMNLP. 2013: 1213-1223.

[16] Xiao T, Zhu J, Zhang H, et al. NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation[C]// ACL 2012 System Demonstrations. 2012:19-24.

[17] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 311-318.

作者联系方式：
姓名：王楠
地址：北京市海淀区上园村 3 号北京交通大学计算机与信息技术学院
邮编：100044
电话（最好手机）：15201326392
电子邮箱：14120428@bjtu.edu.cn

作者简介：

王楠（1992——），女，硕士研究生，主要研究领域为统计机器翻译。
Email: 14120428@bjtu.edu.cn；

徐金安（1970——），男，副教授，主要研究领域为自然语言处理和机器翻译。
Email: jaxu@bjtu.edu.cn；

明芳（1991——），女，硕士研究生，主要研究领域为统计机器翻译。
Email: 14120416@bjtu.edu.cn；

陈钰枫（1981——），女，副教授，主要研究领域为自然语言处理和机器翻译。Email: chenyf@bjtu.edu.cn；
张玉洁（1961——），女，教授，主要研究领域为自然语言处理和机器翻译。Email: yjzhang@bjtu.edu.cn。