北京大学学报(自然科学版) 第 53 卷 第 2 期 2017 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 53, No. 2 (Mar. 2017)
doi: 10.13209/j.0479-8023.2017.036

融合语态特征的日英层次短语翻译模型

王楠 徐金安† 明芳 陈钰枫 张玉洁

北京交通大学计算机与信息技术学院, 北京 100044; † 通信作者, E-mail: jaxu@bjtu.edu.cn

摘要 针对不同语种的被动和可能语态的句法结构差异影响机器翻译质量的问题, 提出融合语态特征的最大熵翻译模型。首先从日语端分出被动语态、可能语态和其他语态, 然后从英语端对被动和可能语态进一步分类, 抽取双语特征训练最大熵规则分类模型, 将语态特征融合到对数线性模型中以改善翻译模型。提高解码器在翻译被动语态和可能语态时规则选择的准确性。实验结果表明, 该方法可以有效地改善日英统计机器翻译的句法结构调序和词汇翻译, 提升被动语态和可能语态句子的翻译质量。

关键词 被动语态; 可能语态; 统计机器翻译; 最大熵模型

中图分类号 TP391

Integrating Voice Features into Japanese-English Hierarchical Phrase Based Model

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

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

Abstract The voice of each language usually keeps different syntactic structure. In machine translation, it causes relatively low translation quality. To resolve this problem, an approach is proposed by integrating voice features into hierarchical phrase based (HPB) models. In the proposed method, corpus is firstly classified into three categories from Japanese side: passive voice, potential voice and others. Secondly, passive and potential sentences are classified into several groups according to the characteristics of English to build maximum entropy models for rules. Finally, bilingual voice features are integrated into log linear model for improving translation results and the accuracy of rule selection during the translation of passive and potential sentences. In Japanese to English translation task, large scale experiment shows that the proposed method can not only improve the problem of long distance reordering but also improve translation quality of both passive and potential voice test sets.

Key words passive voice; potential voice; statistical machine translation; maximum entropy models

日语通过谓词的词尾形式变化表示相应语态, 由于其被动语态和可能语态的部分词尾形式相同, 因而在机器翻译过程中难以正确识别及翻译。日语与英语在语言结构上有显著差异, 日语为 SOV (主宾谓)结构, 英语为 SVO (主谓宾)结构, 句法结构的差异会影响日英机器翻译的质量。其中, 语态不同导致的词汇翻译不准确和结构不当的问题尤为突出。如何正确翻译被动语态与可能语态句子是日英翻译中的重要任务。

现有研究大部分从语义及结构上区分日语的可能语态与被动语态[1], 通过制定翻译规则对不同语态进行处理[2−3], 但基于规则的翻译方法无法直接应用于统计机器翻译系统。统计翻译模型按照概率进行规则选择, 训练语料中可能语态和被动语态的数据稀疏, 统计方法处理远距离调序困难, 难以有效地利用句子全局结构, 这些特征导致翻译精度低

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(2014RC040)和科学技术部国际科技合作计划(K11F100010)资助
收稿日期: 2016−07−29; 修回日期: 2016−10−07; 网络出版日期: 2016−11−28

下。针对日语中的可能语态和被动语态，抽取相应的句法结构特征，分别构建可能语态和被动语态的翻译模型来提高这两种语态的翻译精度，是解决上述问题的一种可行的思路和方法。

近年来，很多研究者通过构建分类模型来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，并用于短语排序。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。Van 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，可以有效地提升翻译质量，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。

虽然关于引入语态信息的统计机器翻译研究不多，但很多研究者通过引入语言学分析，改进了层次短语模型。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。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，提出一种将日语的语态特征融合到统计翻译模型中的方法。首先将语料分为被动语态、可能语态和其他语态三类，针对英语句法特点，将被动语态和可能语态继续细分，抽取双语句法特征训练最大熵模型。翻译过程中，在抽取规则的同时，抽取语态特征，通过最大熵模型将语态特征融合到对数线性模型中来改进翻译模型，帮助解码器更好地进行规则选择。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，并且引入语态特征这一语言学约束来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 0.07~0.82 的 BLEU 值提升，在人工评测中翻译结果的整体可理解度也得到 0.89%~2.03%的提升，并且在日汉语料对比实验得到验证。

1 语态分类

日语中表示主体受到另一事物的动作时使用动词的被动语态，表示具有某种能力或某种可能性时使用动词的可能语态，这两种语态多数情况由动词未然形后加词尾 “れる”和“られる”构成。如“eat”对应日语动词未然形“食べる”，由于其可能语态和被动语态形式都是“食べられる”，导致统计机器翻译难以正确翻译。

本文提出的方法需要对规则进行分类，规则两侧包含源语言与目标语言信息。为了提高翻译精度，不仅需要从日语端分出被动语态、可能语态和其他语态三类，还需从英语端对被动语态和可能语态进一步分类，便于抽取双语特征来训练最大熵规则分类模型。

在英语被动语态的相关研究中，被动语态因时态不同，其表现形式不同，通常按时态分为 8 类(一般现在、一般将来、一般过去、现在进行、过去进行、现在完成、情态动词和其他)。在实验中用到的 50 万日常会话对齐语料中，本文将句数较少、不具代表性的类别暂归至其他类，最终将被动语态分为一般现在时、一般过去时、现在完成时、情态动词和其他，共 5 类。

英语语态中没有可能语态，故将日语可能语态句子对应的英语译文进行如下划分：不同的情态动词(can, could, may, would, will)分为 5 类，动词短语归为一类，形容词和副词(possible, probable etc.)归为一类。将语料中出现次数少且不具有代表性的可能语态句子划分至其他类。最终，将可能语态分为 8 类。

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，规则形式如下：

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

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用~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。然后，利用子短语替换短语规则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，但在泛化过程中没有保留上下文信息，

子短语可以匹配任何的句法成分，因此翻译时容易产生错误。

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。式(2)为对数线性模型中的转换得分：

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

通常使用对数的形式表示：

$$\begin{aligned}\text{score}(d)&=\log P(d)\\&=\log\prod_{i}^{M}\phi_i(d)^{\lambda_i}=\sum_{i}^{M}\lambda_i\log\phi_i(d),\end{aligned} \tag{3}$$

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 $P(e\mid f)$和 $P(f\mid e)$，正反向词汇化权重 $P_{\text{w}}(e\mid f)$和 $P_{\text{w}}(f\mid e)$，N 元语言模型 $p_{\text{lm}}(f)$，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

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语态与可能语态句子，剩余句子归为其他语态，抽取日语句法特征训练测试集分类模型；针对英语句法特点设定规则，将被动语态与可能语态进一步分类，对双语语料进行句法分析，抽取双语语态特征，分别训练两种语态的最大熵规则选择模型；在规则抽取过程中抽取出相应的语态特征，通过最大熵模型融合至规则表中，生成可能语态和被动语态的翻译模型；最后在翻译过程中，通过判断输入句子的语态来选择相应的翻译模型，实现在解码过程中的规则自动过滤。

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

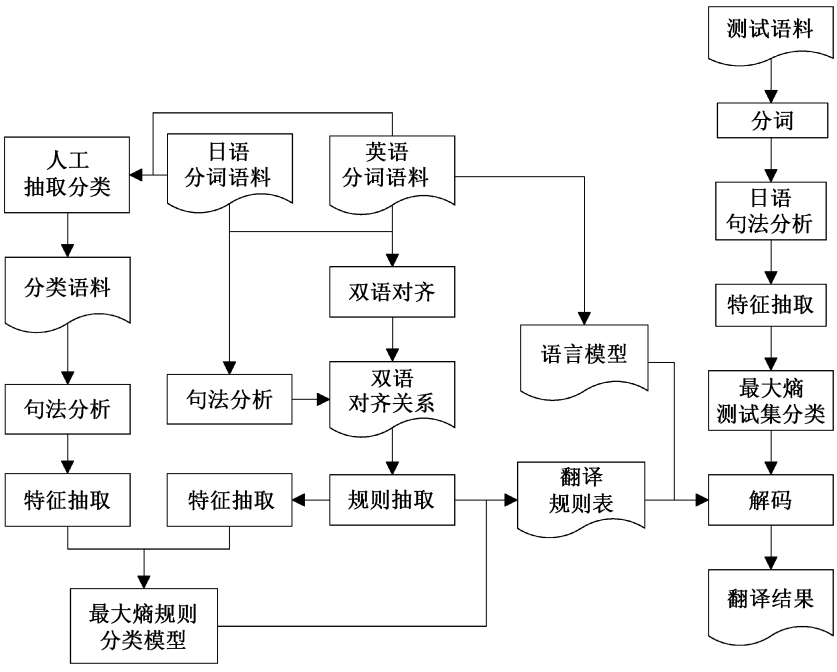

图 1　融合语态特征的日英翻译系统流程

Fig. 1　Flow chart of Japanese-English translation system with voice features

大熵模型作为测试集分类模型和规则分类模型。

假设存在样本集合 $T=\{(x_1, V_1), (x_2, V_2), \ldots, (x_n, V_n)\}$，其中，$x_i(1\leqslant i\leqslant n)$是一个句子的上下文环境，$V_i(1\leqslant i\leqslant n)$表示句子的语态类别。最大熵约束通过特征函数来实现，对于语态分类问题，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下:

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

其中, $H(P)$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，如式(6)所示:

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

$$Z(x)=\frac{1}{\sum\limits_V \exp\left(\sum\limits_i \lambda_i f_i(x,V)\right)}, \tag{7}$$

其中, $Z(x)$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〈α, γ〉，可构建以下最大熵规则分类模型:

$$\begin{aligned}&P(V\mid\alpha,\gamma,f(X_k))\\&=Z(x)\exp\left[\sum_i \lambda_i f_i(V(\alpha),f(X_k))\right],\end{aligned} \tag{8}$$

$$Z(x)=\frac{1}{\sum\limits_{\gamma'}\exp\left[\sum\limits_i \lambda_i f_i(V(\alpha),f(X_k))\right]}, \tag{9}$$

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 $f(X_k)$, $V(\alpha)$表示源语言短语中上下文的语态信息, $f_i(V(\alpha), f(X_k))$是一个二值的特征函数, λ_i 为该函数的特征权重。

3.3 测试集分类特征抽取

从日语端对语料进行分类时，许多动词的被动和可能语态具有相同形式，需要引入句子的结构特征进行区分。Kawahara 等[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 等[14]和 Sasano 等[15]抽取格框架特征，能够有效地识别被动语句。在针对测试集的最大熵分类模型中规定以下特征(F1, F2)。

句子的中心结构词特征(F1): 日语句子中，谓语动词以及词尾多在句尾，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。

以句子“地下鉄で私の財布は憎らしい泥棒に盗まれました (On the subway my wallet was stolen by a hateful thief)”为例，其依存句法树如图 2 所示。在抽取前进行句法分析及标注，抽取特征 F1 为“盗ま れ ました”, F2 为“地下鉄 で 財布 は 泥棒 に”。

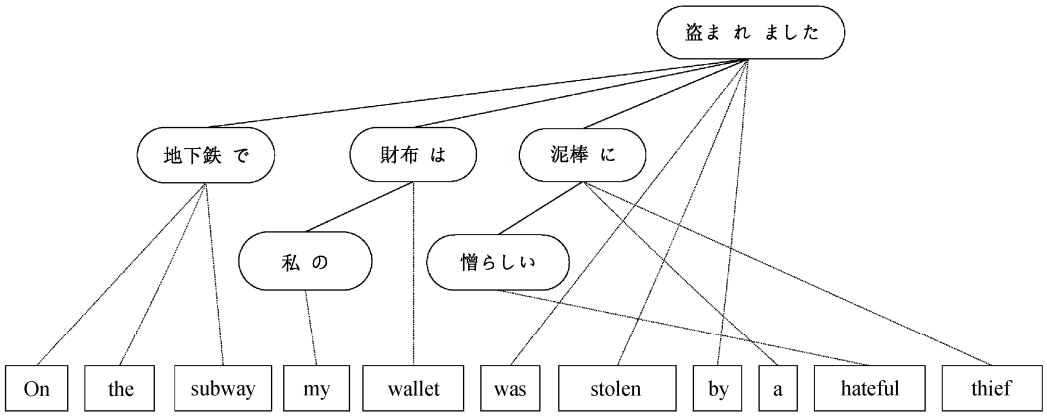

图 2 日语依存句法树示例

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

3.4　规则分类特征抽取

针对英语语法特点，将可能语态与被动语态细分后，按照层次短语规则，需抽取双语特征来训练最大熵模型。翻译过程中，在抽取层次短语规则的同时抽取其语态特征。用上述例句中的短语来示例，首先从词对齐关系中抽取到以下 3 个短语规则：1) $X \rightarrow$〈れ, by〉; 2) $X \rightarrow$〈泥棒 に, a thief〉; 3) $X \rightarrow$〈泥棒 に 盗ま れ, stolen by a thief〉。根据上述规则，可以得到包含两个非终结符的层次短语规则：$X \rightarrow \langle X_1$ 盗ま X_2, X_2 stolen $X_1 \rangle$。

泛化的部分含有根节点信息，需对该规则进行特征抽取。规定词性及目标语言句法特征如表 1 所示。抽取 F1 时，需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)。F2 仅指非终结符中的句子结构信息。表 1 中最后一列为根据上述规则抽取出的对应的最大熵特征。

3.5　翻译模型融合

分别训练被动语态与可能语态的最大熵模型，将每个层次短语规则作为测试集，得出不同类别的最大熵概率值。将被动语态 5 类最大熵概率值加入原规则表中，生成被动语态规则表，用同样方法生成可能语态规则表。在层次短语规则表中，有一类不包含语态信息的规则，包括短语规则(没有非终结符)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，语态特征位置为 0。

最终生成两个翻译模型，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。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，首先对输入的句子根据语态分类，然后照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

由于公开数据集多为科技文献和新闻语料，其中可能语态句子数据稀疏的问题比较严重，针对本文研究的歧义句数量很少，故实验采用语态信息更为丰富的日常会话语料作为数据集。本文数据来源于从网页端抽取整理的 50 万句日英日常会话信息，并人工分类抽取出被动语态及可能语态语句。分类情况及语料相关信息见表 2。

为了验证方法的有效性，本文从 50 万句日汉语料上进一步验证翻译模型的效果。日汉日常会话语料分类情况见表 3。对比实验中暂不针对中文端进一步分类，仅使用源语言特征构建最大熵规则分类模型。

本文使用 Juman[①]和 KNP[②]作为对日语分词及句法分析的工具。使用 Stanford-Chinese-Segmenter 工具[③]对中文句子进行分词。使用最大熵工具包[④]作为分类工具。词对齐信息从 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

表 1　规则特征抽取

Table 1　An example of feature extraction in rules

语言端	特征类型	特征名称	特征描述	特征值
源语言端	句法特征	F1	句子中心结构词	盗ま れ ました
		F2	非终结符中包含的句子主干结构词	泥棒 に
	词性特征	P1	中心结构词词性	動詞 接尾辞 接尾辞
		P2	句子主干结构词词性	名詞 格助詞
目标语言端	句法特征	F3	词对齐关系中 F1 对应的谓语动词	was stolen by
	词性特征	P3	谓语动词词性	BEV V-PP PREP

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 2 日英语料信息

Table 2 Japanese-English corpus of experiment

语态	类别	总句数	训练集	开发集	测试集
被动语态	一般现在时	5728	5328	200	200
	一般过去时	2757	2357	200	200
	现在完成时	1098	698	200	200
	情态动词	653	453	100	100
	其他	757	357	200	200
可能语态	can	14093	13693	200	200
	could	6542	6142	200	200
	may	2304	1904	200	200
	would	2813	2413	200	200
	will	2201	1801	200	200
	动词短语	305	205	50	50
	形容词/副词	764	364	200	200
	其他	8129	7729	200	200
其他语态	其他	452181	451181	500	500

表 3 日汉语料信息

Table 3 Japanese-Chinese corpus of experiment

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

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人工评测。

4.2 分类实验结果

首先验证测试集分类效果，其准确率直接影响到翻译结果。分别抽取 3 种语态各 500 句作为分类测试集，剩余的为训练集。使用准确率进行评价，实验结果见表 4。结果表明，与只加入中心词特征相比，加入句子的主干结构特征后准确率明显提升，可以更有效地识别被动语态和可能语态。

4.3 翻译实验结果

实验中使用已分定类别的测试集进行评价。由于 BLEU 值不能完整地体现语态信息的翻译效果，所以本研究以《机器翻译评测大纲》[①]中关于人工评测的规范为标准，对翻译结果进行人工评测。根据可理解度，句子评分取 0~5.0 分不等，含一位小数，最后得分为所有打分的算术平均值，使用百分制换算评测结果。

表 4 最大熵测试集分类实验结果

Table 4 The result of maximum entropy classification of test sets

特征	准确率/%
F1	93.93 (1409/1500)
F1+F2	98.47 (1477/1500)

在被动语态与可能语态测试集上 BLEU-4 及人工评测结果如表 5 和 6 所示。日汉对比实验结果如表 7 所示。分析实验结果可知，与层次短语模型相比，本文方法在被动测试集上 BLEU 值有 0.17~0.82 的提升，在可能语态测试集上 BLEU 有 0.07~0.58 的提高。由于 BLEU 采用 N-gram 的完全匹配，因此针对语态的处理对 BLEU 值的影响不大。分析人工评测结果可知，与基线系统相比，本文方法在被动语态及可能语态上的可理解度均有 0.89%~

① http://www.liip.cn/cwmt2013/

表 5　被动语态测试集的 BLEU 值

Table 5　BLEU-4 scores on the passive test sets　%

评测方法	翻译模型	一般现在时	一般过去时	现在完成时	情态动词	其他	全部
BLEU-4	Baseline	32.10	36.62	28.10	31.31	27.42	30.24
	+Maxent Features	32.92	36.79	28.43	31.52	27.85	30.62
人工评测	Baseline	74.08	76.40	66.25	73.87	67.31	71.33
	+Maxent Features	76.11	77.89	67.32	74.76	68.53	72.71

表 6　可能语态测试集的 BLEU 值

Table 6　BLEU-4 scores on the potential test sets　%

评测方法	翻译模型	can	could	may	would	will	动词短语	形容词/副词	其他	全部
BLEU-4	Baseline	38.65	35.44	33.68	36.48	27.42	33.69	34.16	39.38	36.61
	+Maxent Features	39.22	35.78	33.97	36.99	27.85	34.27	34.50	39.45	37.14
人工评测	Baseline	78.61	75.86	70.17	74.55	68.20	72.49	73.17	79.21	74.19
	+Maxent Features	79.73	77.03	71.25	76.32	69.88	74.04	74.30	80.13	75.47

表 7　日汉对比实验 BLEU 值

Table 7　BLEU-4 scores on Japanese-Chinese test sets　%

评测方法	翻译模型	被动语态测试集	可能语态测试集	其他语态测试集
BLEU-4	Baseline	42.60	41.50	39.58
	+Maxent Feature	42.69	42.01	39.71
人工评测	Baseline	69.70	71.02	67.39
	+Maxent Feature	72.44	74.13	69.68

2.03%的提高，可理解度在整体上优于基线系统。日汉对比实验中，与基线系统相比，本文方法在被动测试集上 BLEU 值提升 0.09，在可能语态测试集上 BLEU 有 0.51 的提高，可理解度提高 2.29%~3.11%。对比实验结果可知，日汉翻译 BLEU 值的提升幅度略小于日英翻译，但译文可理解度在整体上比日英翻译提升更多。

4.4　翻译结果分析

首先，分析实验结果可知，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 例句 1 是被动语态语句，表示主语被迫去教堂。基线系统在解码时，显然选用了可能语态的规则进行翻译，误译为“would you like to go to church” (你可以去教堂吗)，出现明显的语义错误。加入语态特征后，解码时更倾向于选择被动语态的规则进行翻译，译出被强迫的含义，语义及结构上都优于基线系统。

其次，本文提出的翻译模型在句法结构调序方面也明显优于基线系统。表 8 中例 2 意为 Jimmy 被 Jennie 抛弃了，基线系统翻译时对主谓宾关系没有进行正确调序。加入语态特征后译出正确的主语和宾语，说明融合最大熵特征的翻译模型提高了解码器规则选择的正确率。

此外，与基线系统相比，在词汇选择方面融合语态特征的翻译模型的性能更优。这可能是语态测试集中融合语态特征系统的翻译质量也有所提升的主要原因。例如表 8 中例句 3，原意为询问“是否可以乘车去观光”，与翻译成“is it possible”相比，翻译成“can”更符合语言习惯。

本文方法在语态翻译上有所改善，但仍存在语态信息的误译和漏译。首先，对测试集分类时仍会出现分类错误的句子；其次，分类正确的语句在翻译过程中也会有漏译的情况。主要原因是数据稀疏问题严重，动词在不同语态的分布不均衡。语料中目标语言与源语言语态不一致也是导致漏译的原因之一。

表 8 被动语态和可能语态句子翻译结果

Table 8 Translation of passive and potential sentences %

项目	结果
例句 1	子供 の 時 よく 教会 に **[**行か さ れ ました**]** 。
参考译文	when i was a child i **[**was forced to go**]** to church a lot .
Baseline	**[**would you like to go**]** to church often when i was a child .
+Maxent Feature	when i was a kid i [was made to go] to church often .
例句 2	[ジミー] は [ジェニー] に 振ら れ た 。
参考译文	[jimmy] was jilted by [jennie] .
Baseline	[jenny] was dumped by 's [jimmy] .
+Maxent Feature	[jimmy] was dumped by [jenny] .
例句 3	タクシー で 観光 する こと は でき ます か 。
参考译文	can i use a taxi for sightseeing ?
Baseline	is it possible to go sightseeing by taxi ?
+Maxent Feature	can i go sightseeing by taxi ?

5 结语

本文针对日英统计机器翻译, 提出一种融合语态特征的层次短语翻译模型, 可以提高规则选择的准确性, 从而提升被动语态与可能语态的翻译精度。首先抽取双语特征构建分类模型, 然后利用最大熵模型将语态特征与层次短语翻译模型相融合, 实现在解码过程中对不同语态规则的自动过滤。实验结果显示, 本文提出的方法可以有效地提高翻译质量。

今后的工作主要包括: 如何有效地解决学习数据的不平衡问题, 进而提高分类精度和翻译性能; 针对公开数据集语料特点进行语态分类, 验证本文方法; 与基于句法的翻译模型进行语态翻译对比, 如树到串模型; 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 // Proceedings of the Pacific Asia Conference on Languages, Information, and Computation. Seoul, 2007: 340–348

[2] Alam Y S . A rule-based morpho-semantic analyzer of the Japanese verb phrases of simple sentences // PACLIC 22. Cebu City, 2008: 101–112

[3] 卜朝晖, 浅井良信, 王軼諞, など. 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). 情報処理学会研究報告: 自然言語処理研究会報告 2006(124). 東京, 2006: 33–40

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

[5] He Z, Liu Q, Lin S. Improving statistical machine translation using lexicalized rule selection // Scott D. Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, 2008: 321–328

[6] Van Nguyen V, Shimazu A, Le Nguyen M, et al. Improving a lexicalized hierarchical reordering model using maximum entropy // MT Summit Ⅻ. Ottawa, 2009: 73–80

[7] Iglesias G, De Gispert A, Banga E R, et al. Rule filtering by pattern for efficient hierarchical translation // Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens: 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 // Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Columbus: ACL, 2008: 577−585

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

[10] Gao Y, Koehn P, Birch A. Soft dependency constraints for reordering in hierarchical phrase-based translation // Proceedings of the EMNLP & ACL. Oregon, 2011: 857−868

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

[12] Chiang D. Hierarchical phrase-based translation. Computational Linguistics, 2007, 33(2): 201−228

[13] Kawahara D, Kurohashi S. Case frame compilation from the web using high-performance computing // Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, 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 // Proceedings of the COLING/ACL on Main Conference Poster Sessions. Sydney: 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 // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: 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 // ACL 2012 System Demonstrations. Jeju Island, 2012: 19−24

[17] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th ACL. Philadelphia, 2002: 311−318