第33卷 第3期 Vol.33 No.3 March 2022

ISSN 1000-9825
CODEN RUXUEW

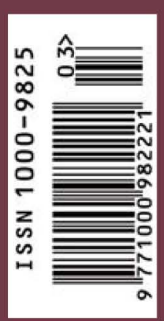
软件学报 Journal of Software

软件学报 Journal of Software

2022 年第 33 卷第 3 期

数据库系统新型技术专题

ISSN 1000-9825



主办

中国科学院软件研究所

中国计算机学会

出版

科学出版社

软 件 学 报
(Ruanjian Xuebao)

第 33 卷第 3 期
2022 年 3 月

目 次

数据库系统新型技术专题

模式识别与人工智能

期刊基本参数: CN11-2560/TP*1990*m*16*384*zh+en*P*¥70*2022*21*2022-03

软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(3):1141−1152 [doi: 10.13328/j.cnki.jos.006201]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多覆盖模型的神经机器翻译*

刘俊鹏, 黄锴宇, 李玖一, 宋鼎新, 黄德根

(大连理工大学 计算机科学与技术学院, 辽宁 大连 116024)

通信作者: 黄德根, E-mail: huangdg@dlut.edu.cn

摘 要: 覆盖模型可以缓解神经机器翻译中的过度翻译和漏翻译问题. 现有方法通常依靠覆盖向量或覆盖分数等单一方式存储覆盖信息, 而未考虑不同覆盖信息之间的关联性, 因此对信息的利用并不完善. 针对该问题, 基于翻译历史信息的一致性和模型之间的互补性, 提出了多覆盖融合模型. 首先定义词级覆盖分数概念; 然后利用覆盖向量和覆盖分数存储的信息同时指导注意力机制, 降低信息存储损失对注意力权重计算的影响. 根据两种覆盖信息融合方式的不同, 提出了两种多覆盖融合方法. 利用序列到序列模型在中英翻译任务上进行了实验, 结果表明, 所提方法能够显著提升翻译性能, 并改善源语言和目标语言的对齐质量. 与只使用覆盖向量的模型相比, 过度翻译和漏翻译问题的数量得到进一步减少.

关键词: 神经机器翻译; 注意力机制; 序列到序列模型; 多覆盖模型; 过度翻译; 漏翻译

中图法分类号: TP183

中文引用格式: 刘俊鹏, 黄锴宇, 李玖一, 宋鼎新, 黄德根. 基于多覆盖模型的神经机器翻译. 软件学报, 2022, 33(3): 1141−1152. http://www.jos.org.cn/1000-9825/6201.htm

英文引用格式: Liu JP, Huang KY, Li JY, Song DX, Huang DG. Multi-coverage Model for Neural Machine Translation. Ruan Jian Xue Bao/Journal of Software, 2022, 33(3): 1141−1152 (in Chinese). http://www.jos.org.cn/1000-9825/6201.htm

Multi-coverage Model for Neural Machine Translation

LIU Jun-Peng, HUANG Kai-Yu, LI Jiu-Yi, SONG Ding-Xin, HUANG De-Gen

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: The over-translation and under-translation problem in neural machine translation could be alleviated by coverage model. The existing methods usually store the coverage information in a single way, such as coverage vector or coverage score, but do not take the relationship among different coverage methods into consideration, leading to insufficient use of the information. This study proposes a multi-coverage mechanism based on the consistency of translation history and complementarity between different models. The concept of word-level coverage score is defined first, and then the coverage information stored in both coverage vector and coverage score are incorporated into the attention mechanism simultaneously, aiming to reduce the influence of information loss. According to different fusion methods, two models are introduced. Experiments are carried out on Chinese-to-English translation task based on sequence-to-sequence model. Results show that the proposed method could significantly enhance translation performance and improve alignment quality between source and target words. Compared with the model with coverage vector only, the number of over-translation and under-translation problem is further reduced.

Key words: neural machine translation; attention mechanism; sequence-to-sequence model; multi-coverage model; over-translation; under-translation

机器翻译是使用计算机将一种语言(源语言)自动翻译成另一种语言(目标语言)的技术. 近年来, 神经机器翻译(neural machine translation, NMT)经过不断发展和完善, 显示出强大的翻译性能. 现有的神经机器翻译模型通常采用“编码器-解码器”架构, 其中, 编码器和解码器可以采用不同的网络结构, 依据神经网络拓扑结构

* 基金项目: 国家重点研发计划(2020AAA0108004); 国家自然科学基金(61672127, U1936109)
收稿时间: 2020-04-01; 修改时间: 2020-06-17, 2020-09-30; 采用时间: 2020-10-28

的特点, 分为循环神经网络(recurrent neural network)[1−3]、卷积神经网络(convolutional neural network)[4]和自注意力网络(self-attention network)[5]等. 目前, 神经机器翻译的性能已经显著超越传统的统计机器翻译, 成为当前机器翻译领域的前沿热点[6].

虽然大规模训练数据和强大的计算能力使得神经机器翻译的性能大幅提升, 但是神经机器翻译还存在着过度翻译(over-translation)和漏翻译(under-translation)问题. 在基于短语的统计机器翻译模型中, 文献[7]通过将加入输出序列的翻译结果所对应的源语言短语标记为“已翻译”的方式, 确保所有的源语言短语均被翻译覆盖且只被翻译一次. 但是在神经机器翻译模型中没有显式存储历史翻译信息的结构, 并且每一次解码过程都需要所有源语言词语的参与, 因此极易出现过度翻译和漏翻译问题.

为了向神经机器翻译模型中增加覆盖信息, 文献[8]借鉴统计机器翻译中的覆盖思想提出了覆盖模型(coverage model). 设计一个覆盖向量(coverage vector, CV)记录解码过程中的翻译历史信息, 并指导注意力权重计算, 降低已翻译词语的权重, 使注意力机制更多地关注未翻译词语. 类似地, 文献[9]为源语言词语设置全覆盖编码向量(full coverage embedding vector), 在解码过程中不断削减已翻译词语的编码向量, 以此降低其在未来解码中的作用. 文献[10]利用循环注意力机制(recurrent attention mechanism)给注意力单元提供更多的重调序信息, 并通过条件解码器(conditioned decoder)减少重复翻译. 文献[11]增加了两个额外的循环神经网络分别记录翻译过程中的历史(past)和未来(future)信息, 并利用该信息指导注意力机制和解码状态.

覆盖度还可以与解码算法相结合, 用于在已生成的译文中筛选对原文忠实度最高的结果. 文献[12]通过覆盖惩罚(coverage penalty)和长度归一化(length normalization)改进束搜索(beam search)算法, 使模型在选择翻译结果时考虑该翻译结果对源语言信息的覆盖程度, 避免偏向句子长度更短的翻译结果. 文献[13]提出将覆盖分数(coverage score, CS)引入每一次束搜索过程, 以此减少搜索错误, 并且使覆盖分数的计算适用于源语言词汇和目标语言词汇的多种映射关系.

在翻译过程中, 源语言上下文通常影响翻译的忠实度, 而目标语言上下文则与译文的流利度有关. 基于这种思路, 文献[14]提出了上下文门(context gate)方法, 根据源语言和目标语言上下文的重要程度, 动态控制两种上下文信息在生成目标词语时的影响比重. 该方法可以与覆盖机制相结合, 能够在改善翻译结果对源语言的覆盖度的同时, 提升译文的流利度. 文献[15]通过解码历史增强注意力机制(decoding-history enhanced attention mechanism)建立所有源语言词和目标语言词的结构关系, 使神经机器翻译模型更好地选择源语言端和目标语言端的信息.

此外, 文献[16]针对源语言中被漏翻译的词语特点展开研究, 发现在源语言中, 翻译熵(translation entropy)越大的词语越容易被漏翻译, 并据此设计了一种粗粒度到细粒度(coarse-to-fine)框架, 分别解决句子级和词级的训练和翻译问题, 从而减少熵高词语的漏翻译情况.

虽然上述方法均能在一定程度上缓解神经机器翻译中的过度翻译和漏翻译问题, 但该问题依然不能被完全避免. 如表 1 所示, 在基线系统的翻译结果中, 源语言句子中的“事关全局(related to the overall situation)”被遗漏翻译, 且“体制(system)”被重复翻译一次. 而引入覆盖模型后, 上述问题并未得到明显改正, “事关全局”虽然被未被遗漏但却被错误地翻译为“in the world”, 而“体制”则被漏翻译.

表 1 过度翻译和漏翻译示例

文本类别	例句
源语言句子	深化国家监察体制改革是**事关全局**的重大政治**体制**改革.
参考译文	Deepening the reform of the state oversight system is a major political **structural** reform **related to the overall situation**.
基线系统	Deepening the reform of the national oversight system is a major political **system system** reform.
覆盖模型	Deepening the reform of the state supervisory system is a major political reform **in the world**.
多覆盖融合模型	Deepening the reform of the country’s supervision system is a major political **system** reform **to the overall situation**.

注: 基线系统采用基于循环神经网络的 Seq2Seq 模型, 覆盖模型是根据文献[8]在基线系统上的复现

由此可见, 现有的覆盖模型对覆盖信息的记录和使用并不完善. 可能的原因是, 覆盖向量在使用 GRU 网

络更新时存在信息损失, 导致注意力权重分配不准确, 进而产生了重复翻译和漏翻译现象. 针对上述问题, 本文提出一种多覆盖融合的神经机器翻译模型. 首先定义一种词级覆盖分数, 用于记录源语言词语在解码过程中的注意力累积情况; 而后在解码阶段, 利用覆盖向量和覆盖分数所记录的覆盖信息同时指导注意力权重计算, 使两种覆盖信息互相补充, 从而最大限度减少信息损失.

本文的主要创新点包括两个方面: (1) 定义了词级覆盖分数概念, 并利用覆盖分数指导注意力机制的权重计算, 拓展了覆盖分数在神经机器翻译模型中的使用方式; (2) 提出了多种覆盖信息的融合模型及两种实现方式, 并利用实验验证了方法的可行性和有效性. 在 CWMT2018 中英数据集上的实验结果表明, 多覆盖融合方法能显著提升翻译质量, 并在覆盖模型基础上, 进一步减少过度翻译和漏翻译现象.

1 研究背景

1.1 基于循环神经网络的神经机器翻译模型

基于注意力机制的神经机器翻译模型通常采用“编码器-解码器”结构. 编码器将源语言句子编码成隐层向量表示, 解码器根据编码器的输出逐字预测目标端句子的单词序列. 给定源端输入句子 $X=\{x_1,x_2,\ldots,x_m\}$, 神经机器翻译模型对目标端句子 $Y=\{y_1,y_2,\ldots,y_n\}$的条件概率 $P(Y|X)$进行建模.

$$P(Y\mid X)=\prod_{j=1}^{n}P(y_j\mid y_{<j},X) \tag{1}$$

具体来说, 若当前已生成的目标端输出为$\{y_1,y_2,\ldots,y_{j-1}\}$, 则生成下一个目标词语 y_j 的概率计算公式为

$$P(y_j|y_{<j},X)=g(y_{j-1},\boldsymbol{t}_j,\boldsymbol{s}_j) \tag{2}$$

其中, $g(\cdot)$是非线性函数, $\boldsymbol{t}_j$ 是 j 时刻解码器的隐层状态, $\boldsymbol{s}_j$ 是将编码器的所有隐层状态加权得到的源语言上下文向量. $\boldsymbol{t}_j$ 和 $\boldsymbol{s}_j$ 的计算公式如下.

$$\boldsymbol{t}_j=f(y_{j-1},\boldsymbol{t}_{j-1},\boldsymbol{s}_j) \tag{3}$$

$$\boldsymbol{s}_j=\sum_{i=1}^{m}a_{ij}\cdot\boldsymbol{h}_i \tag{4}$$

$$a_{ij}=\frac{\exp(e_{ij})}{\sum_{k=1}^{m}\exp(e_{kj})} \tag{5}$$

$$e_{ij}=ATT(\boldsymbol{t}_{j-1},\boldsymbol{h}_i)=v_a^T\tanh(W_a\boldsymbol{t}_{j-1}+U_a\boldsymbol{h}_i) \tag{6}$$

其中,

- $\boldsymbol{h}_i$ 是源语言词语 x_i 经编码器编码后得到的隐层向量, 由于在实际中常使用双向循环神经网络作为编码器, 因此 $h_i=[\overrightarrow{\boldsymbol{h}}_i;\overleftarrow{\boldsymbol{h}}_i]$, 其中, $\overrightarrow{\boldsymbol{h}}_i$ 和 $\overleftarrow{\boldsymbol{h}}_i$ 分别是前向和后向循环神经网络的隐层状态, $[\cdot;\cdot]$表示拼接操作;
- a_{ij} 为注意力权重, 表示目标端词汇 y_j 与源语言词语 x_i 之间的关联度;
- $ATT(\cdot)$是计算注意力权重的匹配函数;
- v_a^T , W_a 和 U_a 为权重矩阵.

由于注意力机制的引入, 源语言上下文向量 $\boldsymbol{s}_j$ 不再是一个单一固定的向量, 而是随着不同时刻 a_{ij} 的不同而变化, 从而使解码器在不同时刻对源语言句子中各个词语的信息考量有所侧重. $f(\cdot)$通常使用长短期记忆网络(long-short term memory, LSTM)或门控循环单元(gated recurrent unit, GRU), 其网络结构可采用单层或多层, 并且多层网络结构的性能比单层网络有更加显著的提升[12]. 多层堆叠长短期记忆网络(stacked long-short term memory)中, 第 i 层和 i+1 层状态的计算公式如下所示.

$$c_t^i,h_t^i=LSTM_i(c_{t-1}^i,h_{t-1}^i,x_t^{i-1}) \tag{7}$$

$$x_t^i=h_t^i \tag{8}$$

$$c_t^{i+1},h_t^{i+1}=LSTM_{i+1}(c_{t-1}^{i+1},h_{t-1}^{i+1},x_t^i) \tag{9}$$

其中, $LSTM_i$ 和 $LSTM_{i+1}$ 分别表示第 i 和 i+1 层 LSTM, x_t^i, c_t^i 和 h_t^i 分别表示 t 时刻 $LSTM_i$ 的输入、记忆单元状态和隐层状态.

最后, 在训练集 $\{(x^p, y^p)\}_{p=1}^J$ 上, 利用极大似然估计对目标函数的参数 θ 进行迭代训练, 目标函数如公式(10)所示.

$$L(\theta) = -\frac{1}{J}\sum_{p=1}^{J}\sum_{t=1}^{n}\log P(y_t^p \mid y_{<t}^p, x^p) \tag{10}$$

1.2 覆盖向量

与统计机器翻译模型中双语词语的硬对齐(hard-alignment)关系不同, 神经机器翻译模型的注意力机制提供的是一种软对齐(soft-alignment)方法, 因此难以对覆盖机制进行建模. 文献[8]提出通过覆盖向量显式存储解码过程中历史信息的覆盖模型, 利用覆盖向量所存储的信息指导注意力评分过程, 从而使注意力机制更多地关注未翻译的词语, 其模型结构如图 1 所示.

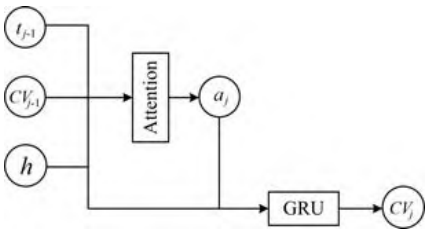


图 1 基于覆盖的注意力模型结构

加入覆盖向量指导后, 注意力权重的计算方法由公式(6)修改为公式(11).

$$e_{ij} = ATT(\boldsymbol{t}_{j-1}, \boldsymbol{h}_i, CV_{i,j-1}) = v_a^T \tanh(W_a \boldsymbol{t}_{j-1} + U_a \boldsymbol{h}_i + V_a CV_{i,j-1}) \tag{11}$$

其中, V_a 为权重矩阵, $CV_{i,j-1}$ 表示 j 时刻前源语言词语 x_i 对应的覆盖向量.

在每一步解码完成后, 利用 GRU 网络对每个源语言词语覆盖向量中存储的覆盖信息进行更新, 更新方法如公式(12)所示.

$$CV_{i,j} = GRU(CV_{i,j-1}, a_{ij}, \boldsymbol{h}_i, \boldsymbol{t}_{j-1}) \tag{12}$$

1.3 覆盖分数

文献[13]提出了表示源语言句子翻译程度的覆盖分数概念. 给定一个翻译句对(X,Y), 将源语言词语 x_i 的覆盖度定义为所有目标词语对该源语言词语的注意力权重之和, 如公式(13)所示.

$$coverage_{x_i} = \sum_{j=1}^{|Y|} a_{ij} \tag{13}$$

在此基础上, 利用所有源语言词语的覆盖度表示源语言句子的覆盖分数, 计算公式如公式(14)所示.

$$cs(X,Y) = \sum_{i=1}^{|X|} \log \max(coverage_{x_i}, \beta) \tag{14}$$

其中, β为可调参数. 最后, 将模型预测的条件概率与译文的覆盖分数线性组合, 使模型在选择译文时兼顾对源语言句子翻译覆盖程度. 改进后的评价函数如公式(15)所示.

$$score(X,Y) = a \cdot \log P(Y|X) + b \cdot cs(X,Y) \tag{15}$$

其中, $\log P(Y|X)$表示模型预测的条件概率值, a 和 b 是用于平衡条件概率和覆盖分数作用的参数.

2 多覆盖融合模型

虽然覆盖向量和覆盖分数均能显式记录翻译过程中的覆盖信息, 但二者在信息的存储方式和使用方式上都有所不同: 前者以向量的形式对信息进行存储和更新, 并通过指导注意力权重计算的方式传递翻译历史;

而后者以常量的形式进行累加, 并作为评价指标用于翻译结果的选择. 两种方法各有优缺点: 覆盖向量存储的信息抽象程度更高, 但在利用 GRU 网络进行更新时, 由于重置门(reset gate)会自动丢弃一定比例的历史信息, 并且更新门(update gate)也会丢掉一部分新的覆盖信息, 因此可能造成覆盖信息损失; 而覆盖分数虽然表达直观, 但难以确定取值界限, 因而无法直接根据数值大小衡量不同词语的覆盖程度.

由于在任意时刻覆盖向量和覆盖分数中存储的信息具有一致性, 且由上述分析可知二者具备一定的互补性, 因此提出一种将两种方法优点相结合的多覆盖融合模型. 利用覆盖向量和覆盖分数存储的信息同时指导注意力机制, 从而降低信息损失对注意力权重分配的影响. 为了将覆盖分数引入注意力机制, 首先定义了词级覆盖分数概念; 然后, 根据覆盖向量和覆盖分数融合方式的不同提出了层次型和平行型两种多覆盖融合模型. 总体框架如图 2 所示.

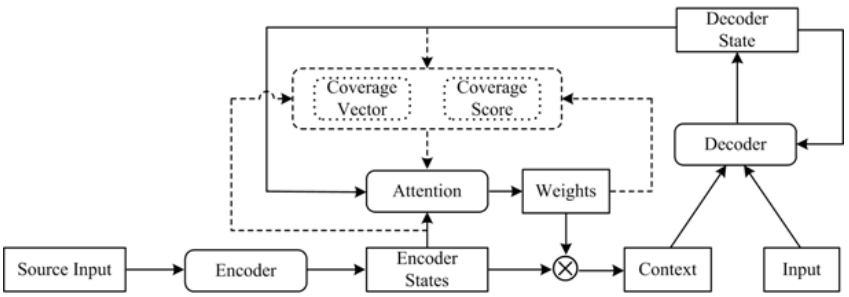


图 2 多覆盖融合的神经机器翻译模型结构

2.1 词级覆盖分数

词级覆盖分数用于表示任意时刻译文结果对每个源语言词语的覆盖程度. 给定源语言句子 X 和 j 时刻目标端的预测序列$\{y_1,y_2,\ldots,y_j\}$, 那么j时刻源语言x_i的覆盖分数定义为当前已生成的所有目标端词语$\{y_1,y_2,\ldots,y_j\}$对源语言 x_i的注意力权重之和的截断取值, 如公式(16)所示.

$$CS_{ij}=\max\left(\sum_{k=1}^{j}a_{ik},1\right) \tag{16}$$

这里, 截断函数 max(·)的作用主要有以下两个方面.

(1) 改善源语言词语覆盖程度的可比较性

在翻译过程中, 存在某些源语言词语对应多个目标词语的情况. 因此, 这些源语言的注意力分数累加和存在大于 1 的可能, 这使得不同源语言词语之间的覆盖程度难以比较. 经过截断取值之后, 覆盖分数 CS_{ij}的取值范围设定在[0,1]之间, 那么对于所有覆盖分数取值为 1 源语言词语, 认为该词已被翻译覆盖; 而覆盖分数小于 1 的词语, 在后续解码过程中, 仍有继续被翻译覆盖的可能.

(2) 借鉴硬对齐思想并融入注意力机制

将源语言词语的覆盖分数取值控制在[0,1]的区间内, 更容易直观地表示当前翻译结果对源语言的覆盖情况和未来翻译过程中仍需被注意力机制关注的, 因此便于融入注意力机制并指导注意力权重的计算.

2.2 层次多覆盖模型

层次多覆盖模型(hierarchical multi-coverage, HMC)主要基于覆盖向量和覆盖分数存储信息的一致性原理, 模型结构如图 3 所示, 首先, 利用覆盖向量指导的注意力模型计算当前时刻分配给所有源语言词语的注意力权重; 然后, 利用覆盖分数存储的信息对注意力权重进行校验和重新分配, 从而减少由于覆盖信息丢失而造成注意力权重分配错误的现象. 层次多覆盖模型注意力权重的计算方法如公式(17)–公式(19)所示.

$$\tilde{e}_{ij}=ATT(\boldsymbol{t}_{j-1},\boldsymbol{h}_i,CV_{i,j-1}) \tag{17}$$

$$\tilde{a}_{ij}=\frac{\exp(\tilde{e}_{ij})}{\sum_{k=1}^{m}\exp(\tilde{e}_{kj})} \tag{18}$$

2022 年第 3 期	数据库系统新型技术	李国良,于戈,杨俊,范举
2022 年第 4 期	面向开放场景的鲁棒机器学习研究	陈恩红,李宇峰,邹权
2022 年第 5 期	领域软件工程	汤恩义,江贺,陈俊洁,李必信,唐滨
2022 年第 6 期	系统软件安全	杨珉,张超,宋富,张源
2022 年第 6 期	定理证明理论与应用	曹钦翔,詹博华,赵永望
2022 年第 7 期	智能系统的分析和验证	明仲,张立军,秦胜潮
2022 年第 8 期	形式化方法与应用	陈立前,孙猛
2022 年第 9 期	融合媒体环境下的媒体内容分析与信息服务技术	汪萌,张勇东,俞俊,张伟
2022 年第 10 期	智慧信息系统新技术	邢春晓,王鑫,张勇,于戈

登录软件学报网站: http://www.jos.org.cn 免费下载专刊/专题全文.

软 件 学 报

Ruanjian Xuebao

(月刊, 1990 年创刊)

第 33 卷 第 3 期 2022 年 3 月

Journal of Software

(monthly)

(Started in 1990)

Vol.33 No.3 Mar. 2022

主管单位	中国科学院
主办单位	中国科学院软件研究所 中国计算机学会
主 编	赵 琛
编 辑	《软件学报》编辑部 (北京 8718 信箱 邮编 100190) 电话: 010-62562563, E-mail: jos@iscas.ac.cn http://www.jos.org.cn
编辑部主任	方 梅
出 版	科 学 出 版 社 (北京东黄城根北街 16 号 邮编 100717)
印 刷	北京宝昌彩色印刷有限公司
总发行处	中国邮政集团公司北京市报刊发行局
订 购 处	全国各地邮局
国外总发行	中国国际图书贸易总公司 (北京 399 信箱 邮编 100044)

Sponsored by the Chinese Academy of Sciences

Published by Institute of Software, The Chinese Academy of Sciences (ISCAS) and China Computer Federation

Editor-in-Chief: ZHAO Chen

Edited by Editorial Board of Journal of Software
(P.O.Box 8718, Beijing 100190, P.R.China)
Tel: 8610-62562563, E-mail: jos@iscas.ac.cn
http://www.jos.org.cn

Distributed by Science Press (16 Donghuangchenggen North Street, Beijing 100717, China)

Printed by Beijing Baochang Color Printing Co., Ltd

Generally Distributed by Beijing Bureau for Distribution of Newspapers and Journals

Domestically Distributed by All Local Post Offices in China
Overseas Distributed by China International Book Trading Corporation (P.O.Box 399, Beijing 100044, China)

ISSN 1000-9825
CN 11-2560/TP

国内邮发代号: 82-367
国外发行代号: M4628

定价: 70.00 元

公 开 发 行