文章编号：1003-0077(2013)04-00016-06

基于多粒度的英汉人名音译

于 恒[1]，涂兆鹏[1]，刘 群[1]，刘 洋[2]

（1. 中国科学院 计算技术研究所 智能信息重点实验室，北京 100190；
2. 清华大学 计算机科学与技术系，北京 100084）

摘 要：音译是解决人名翻译的重要方法。在英汉人名音译问题中，翻译粒度问题一直是研究的重点之一。该文提出一种基于多粒度的英汉人名音译方法。将多种粒度的英文切分通过词图进行融合，并使用层次短语模型进行解码，从而缓解了由于切分错误而导致的音译错误，提高了系统的鲁棒性。实验结果表明基于多粒度的音译方法融合了基于各种粒度音译方法的优点，在准确率上提高了 3.1%，在 BLEU 取得了 2.2 个点的显著提升。

关键词：人名音译；多粒度；词图

中图分类号：TP391 **文献标识码**：A

Lattice-based Multi-granularity Name-Entity Machine Transliteration

YU Heng[1], TU Zhaopeng[1], LIU Qun[1], LIU Yang[2]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Department of Computer Science and Technology, Tsinghua, Beijing 100084, China)

Abstract: Machine Transliteration is an important approach for Name-Entity translation. In English to Chinese transliteration, the translation granularityis of great importance. In this paper we introduce a Multi- granularitymachine transliteration method. We use word lattice to combine multiple syllable segmentation, and decode with hierarchical phrase-based translation model. Experimental results show that our method combines the advantage of different granularityand improve the robustness of the system. We achieve an improvement of 3.1% on precision, and 2.2 points on BLEU over the baseline system.

Key words: name entity machine transliteration; multi-granularity; word-lattice

1 引言

音译作为一种按照文字读音进行近似翻译的方法，在人名翻译中有着广泛的应用[1]。人名音译接受一个源语言的人名作为输入，在保证发音基本不变的原则下，输出与该人名以目标语言表示的翻译。例如，“Julianne”→“朱丽安”。由于音译从读音角度处理翻译问题，在处理未登录词翻译问题上有着良好的效果，因此在很多跨语言任务如机器翻译、跨语言检索以及跨语言问答系统中有着广泛的应用。

由于语言习惯的不同，人名音译过程中，应当适当调整源语言的序列结构（即切分），以使之符合目标语言的语言习惯。因此翻译粒度一直是音译研究的重点之一。Knight 和 Graehl[2] 在日英人名音译中，以英文音素和日文音素为单位，通过发音相似性寻求转换。Al-Onaizan 和 Knight[3]，Sherif[4] 提出以字母为单位，跳过发音过程，直接进行翻译。Wei-Hao Lin 和 Hsin-His Chen[5] 使用音节相似度模型进行人名音译。邹波、赵军[6] 将音节切分问题转换为序列标注问题，采用机器学习的方法进行人名音译。以上方法从不同角度处理音译粒度问题，取得了良好的效果，但每种方法均存在不足之处，主要有以下几个方面。

收稿日期：2012-12-21 **定稿日期**：2013-01-30

基金项目：国家 863 重大项目资助(2011AA01A207)

作者简介：于恒(1988—)，男，博士研究生，主要研究方向为自然语言处理，机器翻译；涂兆鹏(1988—)，男，博士研究生，主要研究方向为自然语言处理，机器翻译；刘群(1966—)，男，教授，主要研究方向为自然语言处理，机器翻译。

(1) 以字母为粒度的方法能够生成较为广泛的音译规则，但规则错误率较高，无法充分利用发音信息辅助切分。

(2) 以音节为粒度的方法利用发音信息进行音节切分，生成准确度较高的音译规则，但模型鲁棒性较差，对一些特例或歧义性音译无法得到正确切分。

(3) 采用机器学习方法的音译策略能够从语料中自主学习音译知识。但对标注语料的依赖性较强，对语料外的切分问题处理能力不佳。

因此，本文提出基于多粒度的英汉人名音译方法。通过词图融合各种粒度的切分，从而缓解了因切分错误而导致的音译错误，在充分利用语言学知识的同时又提高了模型的鲁棒性和音译规则的多样性。实验结果表明，在英汉人名音译中基于多粒度音译方法效果好于单一粒度的音译方法，在准确率上提高 3.1%，在翻译 BLEU 值上提高 2.2 个百分点。

2　统计音译模型

音译问题可以应用 P Brown[7] 提出的噪声信道模型进行建模。当观察到噪声信道的信号为 O 时，我们可以得到一个可能的输入序列集合 $F(O)$，其中的每组输入序列 f 都能得到对应的输出序列 e。我们的目标是找到概率最高的 $\hat{e}$ 作为输出。

$$\begin{aligned}\hat{e} &= \arg\max_{e}\max_{f\in F(O)}\Pr(e, f \mid O)\\ &= \arg\max_{e}\max_{f\in F(O)}\Pr(e)\Pr(f \mid e, O) \qquad (1)\\ &= \arg\max_{e}\max_{f\in F(O)}\Pr(e)\Pr(f \mid e)Pr\ (O \mid f)\end{aligned}$$

在人名音译问题中，O 即为输入英文人名，f 为可能的音节切分序列，e 为人名翻译。模型的目标是从 O 中获取最佳的切分序列 f，然后利用音译规则进行解码，得到正确的音译结果 e。理论上，我们可以简单地通过穷举 $F(O)$ 集合中的所有可能序列 f 来得到最佳翻译，但这样做会带来巨大的计算开销。实际上，许多可能的序列都具有相同的子片段，因此通过词图对这些可能的序列进行表示并在此基础上进行解码会大大提升系统的性能。

3　基于词图的解码

3.1　词图

词图 $G=<V, E>$ 为一个由点和有向边构成的有向无环图，其中 V 为点集合，E 为有向边集合。形式上是一种带权有限自动机。如图 1 所示。

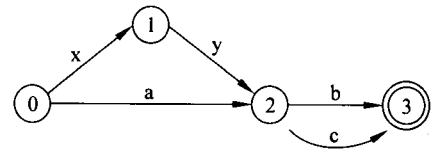

图 1　词图示意图

注：图中节点 3 为终结节点，作为结束状态

词图可以表示各种输入序列，并且支持相同子序列的共享，从初始节点 0 到终止节点的每条路径都代表一组可能的序列，因此能够将不同输入融合在同一个图结构中。

在音译问题中，假设源端为 n 个字母的词，词图上的每个节点为源端的跨度(从 0 到 n)，连接节点的边为该跨度下可能的翻译。我们的目标即为找到一条概率最大的路径，路径上的边即为生成的目标翻译。

如图 2 所示，音译“Julianne”的最佳路径为红线标出的“0-2-4-7-8”，生成的结果为“朱丽安”。

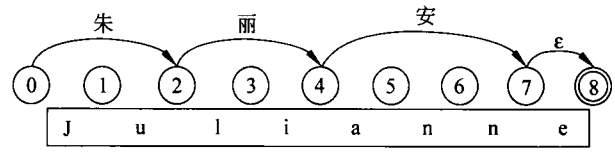

图 2　实例 Julianne 的词图及最优翻译路径

3.2　解码

Chiang[8] 提出了基于上下文无关文法(SCFG)的层次短语翻译模型。在解码过程中，不断使用翻译规则匹配源端输入串，生成翻译片段，同时在目标端生成基于 SCFG 的树结构。本文采用类似方法，从对齐语料中抽取符合上下文无关文法的音译规则进行解码。

在我们的解码算法中包含两种元素。

1. $[X\rightarrow\alpha\cdot\beta, i, j]$ 表示在跨度 (i, j) 上未匹配完成规则，“·”为位点，指示当前需要匹配的符号位置。

2. $[X, i, j]$ 表示在跨度 (i, j) 上为非终结符 X。解码的目标为找到一组覆盖整个词图跨度 $[S, 0, |V|-1]$ 的规则推导。

在解码中，我们定义如下两种规则推导。

1. 匹配一个终结符 β，位点前进一位，同时覆盖相应词图上的一条边。规则跨度变为 $[i, j+1]$，生

成新的翻译片段 $w_{j,j+1}$。

$$\frac{[X\rightarrow\alpha\cdot\beta,\ i,\ j]:\ w}{[X\rightarrow\alpha\beta\cdot,\ i,\ j+1]:\ w\times w_{j,j+1}}$$

2. 匹配一个非终结符 X，位点移位，并找到其对应后继，将两者的翻译片段合并为 $w_1\times w_2$。

$$\frac{[Z\rightarrow\alpha\cdot X\beta,\ i,\ k]:\ w_1\quad[X\rightarrow\gamma\cdot,\ k,\ j]:\ w_2}{[Z\rightarrow\alpha X\cdot\beta,\ i,\ j]:\ w_1\times w_2}$$

基于以上两种推导规则，我们使用 CKY 算法，按照自底向上的顺序，对词图进行解码。

4 多粒度切分方法

为了进行多粒度的融合，需要获得各种粒度的英文切分。本节主要介绍三种切分方法。

4.1 基于字母的切分

基于字母的切分方法[3-4]以英文字母为单位，采用统计的方法学习源端和目标端的对应关系。

假设源端序列为 f_1^j，目标端为 e_1^i。则目标即是找到最佳的对齐 a_1^j，是源端翻译到目标端的概率最大化。即 $\arg\max_{a_1^j}\Pr(f_1^j,\ a_1^j\mid e_1^i)$。采用 Och 和 Ney[9]的方法，使用 GIZA++ 对齐工具我们可以得到源端和目标端的最佳对齐 a_1^j。从对齐中即可抽取源端正确的切分。

4.2 基于音节的切分

Wei-Hao Lin 和 Hsin-His Chen[5]提出以英文音节为单位的切分方法。从发音的角度来寻找符合目标端语言习惯的最佳切分。如图 3 所示。

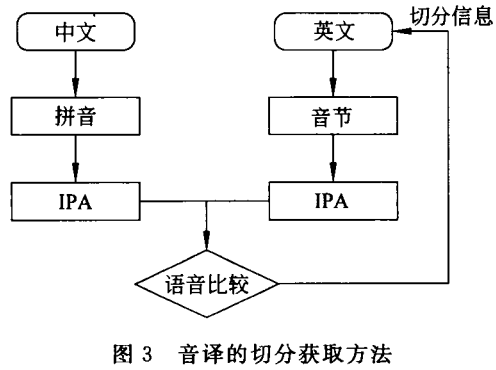

图 3 音译的切分获取方法

首先英文部分利用 CMU pronouncing dict① 将英文序列拆分成相应的音节，再通过音节词典转换为国际通用发音序列 International Phonetic Alphabet (IPA)。同时中文端将汉字转化为拼音，再转化为 IPA 序列。这样源端和目标端通过 IPA 序列进行语音比较，从而找到源端正确的切分。

4.3 基于机器学习的切分

我们将英文音节的切分看成一个序列标注的问题：以 L(音节首)，M(音节中)，R(音节尾)，S(独立音节)来标识英文字母在所在音节中的位置。这四个类别可以覆盖英文字母位置的所有情况。给定一个人工切分好音节的训练集，我们可以很容易得到英文字母的标注序列。

在估计字母位置标注的概率分布时，我们使用最大熵模型。假设 h 为该标注的上下文特征集合，t 为可能的标注集，则最终标注的概率可以表示为 H 和 T 的联合概率分布，如式(2)所示。

$$p(h,\ t)=\pi\mu\prod_{j=1}^{k}a_j^{f_j(h,\ t)} \tag{2}$$

其中 π 为归一化常数，$\{\mu,\ a_1,\ \cdots,\ a_k\}$ 为模型参数，$\{f_1,\ \cdots,\ f_k\}$ 为最大熵模型中定义的特征，$f_j(h,\ t)\in\{0,\ 1\}$。对于每一个特征 f_j，都有一个参数 a_j 与之对应，作为该特征的权重。在训练过程中，给定一个英文字母序列 $\{c_1,\ \cdots,\ c_n\}$ 和它们的标注集 $\{t_1,\ \cdots,\ t_n\}$，训练的目的是找到一组最佳的参数 $\{\mu,\ a_1,\ \cdots,\ a_k\}$，使训练数据的 P 的似然值 $L(P)$ 最大。

$$L(P)=\prod_{i=1}^{n}P(h_i,\ t_i)=\prod_{i=1}^{n}\pi\mu\prod_{j=1}^{k}a_j^{f_j(h,\ t)} \tag{3}$$

最大熵模型的效果在很大程度上取决于选择合适的特征。在 (h,t) 给定的条件下，所选特征必须包含对预测 t 有用的信息。我们在实验中使用特征见表 1。

表 1 切分特征模板

	特征表示	特征描述
1	C0	当前英文字母
2	C-2,C-1, C1, C2	前两个和后两个字母
3	C-1C0, C0C1	前(后)一个字母，与当前字母的组合
4	C-2C-1, C1C2	前两个，后两个字母的组合
5	C-1C1	前一个和后一个字母的组合
6	T-1, T-2	前一个字母的标注，和前两个字母的标注
7	D	默认特征

如上列表所示，我们定义了三类特征，第一类是

① http://www.speech.cs.cmu.edu/cgi-bin/cmudict

基于当前和上下文字母的特征(1，2，3，4，5)，第二类是基于前一二个字母的标注特征(6)，第三类是默认特征(7)，用来捕捉前两类无法表示的情况。当训练完成时，特征和它们的对应权重将可以用作计算未知数据中各种标记的出现概率，如式(4)所示。

$$P(t_1, \cdots, t_n \mid C_1, \cdots, C_n) = \prod_{i=1}^{n} P(t_i \mid h_i) \quad (4)$$

给定一个序列$\{c_1, \cdots, c_n\}$，通过 viterbi 算法可以得到概率最大的标注序列$\{t_1, \cdots, t_n\}$，进而得到切分序列。

5　实验

5.1　实验准备

实验的语料来源于 Chinese-English Name Entity Lists v1.0 (LDC2005T34)，该语料库包括 565 935音译对。我们从中过滤掉一些其他语种音译对，得到 4 万英汉人名对。从中随机挑选 500 对作为开发集，500 对作为测试集，其余作为训练集。汉语语言模型以汉字为单位使用训练集进行训练。

我们分别使用第 4 节所述的三种切分方法获得不同粒度的英文切分，再使用 GIZA ++ 工具对中英文两端进行对齐。音译规则抽取及词汇化模型的训练在生成的对齐数据上进行。

我们使用最小错误率训练方法来优化线性模型的参数。所使用的解码器是层次短语解码的 C++ 重实现版本。该解码器采用 CKY 方式进行解码，并使用 cube-pruning 的方法进行剪枝，以减少搜索空间。实验所使用的栈大小为 100。

我们使用层次短语的经典特征进行解码：

1. 英文序列 e 音译为汉语序列 c 的概率$P(c|e)$
2. 汉语序列 c 音译为英文序列 e 的概率$P(e|c)$
3. 英文序列 e 音译为汉语序列 c 的词汇化概率 $lex(c|e)$
4. 汉语序列 c 音译为英文序列 e 的词汇化概率 $lex(e|c)$
5. 语言模型特征 $lm(c)$
6. 汉语译文长度 $L(c)$
7. 音译规则使用数量 n
8. 黏着规则使用数量 m

5.2　实验结果

在实验中，我们比较了基于各种粒度的音译效果，评价的标准如下。

1. 准确率：音译结果完全匹配的结果百分比。
2. BLEU：机器翻译中常用评价指标，表征音译结果片段的准确率。

由于准确率只考虑完全匹配的情况，从而忽视了某些音译片段的效果提升。因此我们加入 BLEU 作为评价标准，从更细的粒度来考察音译的准确率。

如表 2 所示，以字母为粒度的音译方法准确率为 49.2%，BLEU 值为 0.532 5，在所有方法中效果较差。主要原因是因为以字母为粒度导致对齐边的增多，从而引入了很多对齐错误，导致许多错误的切分。在没有其他发音信息的辅助下，生成过多无用的音译规则，使规则表达到 29MB。而以音节为粒度的方法充分利用发音信息，系统性能有所提升准确率为 54.2%，BLEU 值为 0.551 3。但此方法生成的音译规则过少，导致覆盖率不足，并且不具备处理音译歧义现象的能力。通过机器学习方法得到切分的音译模型在性能上有了进一步的提升，准确率为 61.2%，BLEU 值为 0.572 1。该方法通过标注语料自动学习切分，但由于人工标注语料较为稀少，所以覆盖率有限。

表 2　不同粒度实验效果比较

评价指标 / 模型	规则表大小	准确率 /%	BLEU
字母粒度(ch)	29MB	49.2	0.532 5
音节粒度(sy)	15MB	54.2	0.551 3
机器学习(ml)	24MB	61.2	0.572 1

本文采用词图的方法混合以上三种粒度进行音译规则抽取。实验结果如表 3 所示。

ch＋sy：融合字母和音节粒度

ch＋ml：融合字母和机器学习粒度

sy＋ml：融合音节和机器学习粒度

ch＋sy＋ml：同时融合三种粒度

表 3　不同粒度融合实验效果比较

评价指标 / 模型	规则表大小 /MB	准确率 /%	BLEU
ch＋sy	38	56.4	0.567 3
ch＋ml	41	63.9	0.582 2
sy＋ml	26	61.4	0.572 4
ch＋sy＋ml	45	64.3	0.594

从表 3 中可以发现，基于词图的多粒度融合方法取得了明显的性能提升。值得注意的是字母粒度

虽然自身的性能较差，但是和其他两种粒度融合都取得了明显的效果。而音节粒度和机器学习粒度的融合却没有取得明显的效果提升。造成这种现象的原因是音节粒度和机器学习粒度生成规则的相似性较高，且粒度都较大，因此规则数量较少。所以两者融合后规则表数量并无明显提升，性能上也没有显著增长。而字母为粒度的方法生成规则和其他两种方法差异较大，从某种意义上提升了规则的多样性，从而在融合中取得了良好的效果。最终，我们将三种粒度进行混合，得到最佳的性能，准确率为64.3%，BLEU值为0.594，比单粒度的最好性能准确率提升3.1%，BLEU提升2.2%。

表4列出了不同音译粒度下英文人名，"Julianne"的音译结果。我们可以发现单粒度的结果都存在着不同程度上的问题。而多粒度融合的方法能够得到正确的结果。

表4 不同粒度方法音译"Julianne"的结果

	英文切分	音译结果
ch	Ju li an ne	朱丽安妮
sy	Julian ne	朱莲妮
ml	Ju li an ne	朱丽安娜
ch+sy+ml	Ju li ann e	朱丽安

6 相关工作

近些年来，研究者们在人名音译领域进行了广泛的研究，Knight和Graehl[2]在日英人名音译中，提出以英文音素为粒度，通过发音相似性寻求转换的方法。Al-Onaizan和Knight[3]，Sherif[4]提出以字母为单位，跳过发音过程，直接进行翻译。Wei-Hao Lin和Hsin-His Chen[5]使用音节相似度模型进行人名音译。Long Jiang[10]通过人工定义规则的方法进行了有益的尝试，将英文字母划分为元音和辅音，在切分时遵循元音和辅音配对的原则。邹波，赵军[6]将音节切分问题转换为序列标注问题，将机器学习和统计机器翻译模型用于音译。本文的模型融合以上方法的优势，通过词图融合生成多粒度的音译规则，缓解了因切分错误带来的翻译错误，提高了系统的鲁棒性。

在使用机器学习方法进行切分时，本文使用了最大熵模型[11]。该方法在NLP其他领域都有广泛的应用，如Ratnaparkni[12]将其用于处理词性标注问题，Nianwen Xue[13]在处理中文分词问题时也用到类似方法，取得良好的效果。

在词图解码算法上，Christopher Dyer[14]将其使用在机器翻译上，融合源端的多种分词结果，提升机器翻译的性能。

7 结论

本文提出了一种基于多粒度的英汉人名音译方法，融合多种粒度的切分信息，生成更鲁棒的音译规则。实验结果在准确率上比单粒度效果提升3.1%，BLEU提升2.2%。在后续的研究中，我们将探索更多的切分方法的融合，并改进解码算法，争取进一步提升音译系统的性能。

参考文献

[1] Li Haizhou, Zhang Min, Su Jian. A Joint Source-Channel Model for Machine Transliteration[C]//Proceedings of ACL,2004: 159-166.

[2] Kevin Knight, J. Graehl. Machine Transliteration[J], Computational Linguistics, 1998, 24(4): 599-612.

[3] Yaser Al-Onaizan, Kevin Knight. Translating named entities using monolingual and bilingual resources [C]//Proceedings of ACL, 2002: 400-408.

[4] Tarek Sherif, Grzegorz Kondrak. Bootstrapping a stochastic transducer for Arabic-English transliteration extraction[C]//Proceedings of ACL, 2007: 864-871.

[5] Wei-Hao Lin, Hsin-His Chen. Backward Machine Transliteration by Learning Phonetic Similarity[C]// Proceedings of the 6th CoNLL, 2002: 139-145.

[6] 邹波，赵军. 英汉人名音译方法研究[C]//第四届全国学生计算语言学研讨会论集，2008: 24-30.

[7] Brown P F, Pietra S A D, Pietra V J D. The mathematics of statistical machine translation: parameter estimation[J]. Computational Linguistics, 1993: 19(2): 263-311.

[8] David Chiang. Hierarchical phrase-based translation [J]. Computational Linguistics, 2007, 33(2): 201-288.

[9] Franz Josef Och, Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models[J]. Computational Linguistics, 2003, 29(1): 19-51.

[10] Long Jiang, Ming Zhou, Lee-Feng Chien, et al. Named entity translation with web mining and transliteration[C]//Proceedings of IJCAI, 2007: 1629-1634.

[11] Adam L Berger, Stephen A. Della Pietra, Vincent J.

Della Pietra. A Maximum Entropy approach to Natural Language Processing[J]. Computational Linguistics, 1996,22: 156-242.

[12] Ratnaparkhi, Adwait, A maximum entropy part of speech tagger[C]//Proceedings of EMNLP, 1996: 133-124.

[13] Nianwen Xue. Chinese Word Segmentation as Character Tagging [J]. Computational Linguistics and Clinese Language Processing, 2003, 8(1): 29-48.

[14] Christopher Dyer, Muresan, Philip Resnik. Generalizing Word Lattice Translation [C]//Proceedings of ACL, 2008: 1012-1020.

《中文信息学报》征稿简则

一、《中文信息学报》主要刊登中文信息的基础理论、应用技术、中文信息处理系统及设备、中文信息的自动输入和人工编码输入、汉字字形信息、自然语言处理、计算语言学及民族语言文字信息处理及网上信息处理等方面的研究论文、技术报告、综述、通讯、简报、国内外学术活动等。

二、来稿要求和注意事项

1. 来稿内容力求正确，论点明确，文字简练，数据可靠，图表清晰，字数不超过 8 000 字。

2. 文章题目不超过 20 个字，须有 200 字中文摘要和英文摘要。英文文摘应符合英文语法，概括论文内容，包括研究目的、方法、结果和结论。中英文摘要均应包括题目、作者姓名、单位名称、城市名、邮编、摘要、关键词。写明中图分类号。

有基金项目支持的写明基金名称、编号。

给出前三作者信息，包括姓名，出生年，性别，学位或职称，主要研究方向。

3. 文中图、表放在文稿中相应位置，并注明图号、图注。图中文字用六号宋体。

4. 文中外文字母、符号要分清大小写、正斜体；上下角标的位置高低应区别明显；全文计量单位要一致，或中文，或符号。

5. 参考文献只列最主要的，必须是已公开发行的书刊才能列入，最少不得少于 5 条。文献按文中出现先后次序编排，书写格式为：

专著：[序号] 作者. 题名[M]. 出版地：出版者，出版年 ：起止页码.

期刊：[序号] 作者(多作者用逗号分开，超过 3 个者用“，等”代替). 文章题目[J]. 刊物名称，年代，卷数(期数)：起止页码.

论文集：[序号] 作者. 题名[C]//编者. 论文集名. 出版地：出版者，出版年：起止页码.

学位论文：[序号] 作者. 题名[D]. 保存地点：保存单位，年份.

报告：[序号]作者. 题名[R]. 保存地点：保存单位，年份.

报纸文章：[序号] 作者. 题名[N]. 报纸名，出版日期(版次).

标准：[序号]制定单位. 标准编号，标准名称[S]. 出版地：出版者，出版年.

专利：[序号]专利所有者. 专利题名：专利国别，专利号[P]，公开日期.

电子文献：主要责任者. 电子文献题名[电子文献标识/载体类型]. [发表或更新日期]. 电子文献的出处或可获得地址.

电子文献标识：[DB]—数据库 [CP]—计算机程序 [EB]—电子公告

电子文献载体类型：[OL]—联机网络 [MT]—磁带 [DK]—磁盘 [CD]—光盘

6. 来稿请勿一稿二投，文责自负。不录用稿件概不退还，请自留底稿。来稿一经发表，按规定付给稿酬，并赠送单行本 2 册。

通信地址：北京 8718 信箱《中文信息学报》编辑部收，邮政编码 100190，电话：010-62562916。

本刊接收电子投稿，请以附件方式，将 WORD 文档发送至：cips@iscas.ac.cn。请写明作者工作单位、通信地址(邮政编码)、电话(手机)、E-mail。