多策略汉维句子对齐

田生伟[1] 吐尔根·依布拉音[1] 禹 龙[2] 加米拉·吾守尔[1] 杨飞宇[3]
(新疆大学信息科学与工程学院 乌鲁木齐 830046)[1] (新疆大学网络中心 乌鲁木齐 830046)[2]
(新疆大学国际文化交流学院 乌鲁木齐 830046)[3]

摘 要 提出了一种错误抑制的多策略算法对齐汉维语句子。针对长度对齐算法无法避免错误蔓延的特点,提出了一种新的错误蔓延抑制策略:利用双语语料的词汇共现信息,自动抽取汉维语词汇搭配,结合句子长度特征,寻找1:1模式的句对作为锚点,将错误蔓延抑制在锚点内;在锚点之间,利用标点符号和长度混合方法进行句子对齐。算法实验结果验证了该多策略算法寻找的锚点的精度高,有效抑制了对齐错误的蔓延;采用的混合对齐算法,避免了基于词汇对齐算法的高时间复杂度的弱点,比传统的对齐算法性能有了较大提高,对齐准确率由95.0%提高到97.6%,召回率由96.8%提高到98.2%,采用的对齐正确性评价算法可以有效发现自动对齐中的噪音对齐。

关键词 双语语料,错误抑制,句子对齐,混合策略,汉维句子
中图法分类号 TP391 **文献标识码** A

Chinese-Uyhur Sentence Alignment Based on Hybrid Strategy

TIAN Sheng-wei[1] TURGUN Ibrahim[1] YU Long[2] JAMILA Wushouer[1] YANG Fei-yu[3]
(Information Science and Engineering Technology Institute, Xinjiang University, Urumqi 830046, China)[1]
(Net Center, Xinjiang University, Urumqi 830046, China)[2] (International Cultural Exchange College, Xinjiang Unversity, Urumqi 830046, China)[3]

Abstract This paper proposed a hybrid algorithm of sentence alignment in Chinese-Uyhur parallel corpora. Aiming at the shortcoming of mistake spread in alignment algorithm based on length, this paper presented a new kind of suppression strategy for mistake spread. By using csentence length and Chinese-Uyhur correspondence information, the anchor points with 1:1 pattern sentence pairs are identify to suppress mistakes spread. Among anchor points, a approach based on both length and punctuation is used to align sentences. Experimental results verify the high precision of identifying anchor points and the effective restraint of the spread of mistakes; Hybrid alignmentd algorithm avoids the weakness of high time complexity algorithms based on words. In addition, its performance is improved more compare with traditional alignment algorithms, and increase alignment accuarcy from 95.0% to 97.6% and recall from 96.8% to 98.2%, and the validity evaluation method can find the noised alignment efficently.

Keywords Bilingual corpora, Error curb, Hybrid strategy, Sentence alignment, Chinese-Uyhur sentence

近年来,双语语料库的建设和研究蓬勃发展。双语语料库由于含有两种语言的对应信息,广泛应用于机器翻译[1]、双语词典的编纂[2]、自动问答、信息检索、信息抽取等领域[3-5]。

句子对齐是建设双语语料库的一个重要组成部分,是短语、词对齐的工作基础,也是机器翻译的基础。迄今,句子对齐的方法主要有3种:基于长度的方法[6,7]、基于词汇的方法[8]和混合的方法[9-11]。基于长度的方法,依据是长句子的译文也是长句子,短句子的译文也是短句子,它们的长度满足一定的比例关系,该方法适用于同一语系如印欧语系的句子对齐。长度对齐方法效率高,缺陷是错误容易蔓延,导致大范围错误。基于词汇的方法通常利用双语词典和词汇信息来对齐句子,但由于该方法频繁在大规模词典中查找匹配,因此造成算法时间复杂度很高。混合方法则同时使用两种以上的方法。

本文提出了一种错误抑制的多策略算法对齐汉维句子,将错误抑制在锚点内部,以防止蔓延的发生;利用汉维语标点符号特征和句子长度结合的方法进行句子对齐,避免了词汇方法的高时间复杂度,有效降低了对齐错误率。

1 句子对齐模型

假设汉语文本C和对应的维语译文文本U的一个长度为k的对齐$A=\{A_1,A_2,A_3,\cdots,A_k\}$,$A_i=\langle C_{iS},U_{iT}\rangle(i=1,2,\cdots,k)$。如果每个$C_{iS}$,$U_{iT}$分别包含汉语文本或维语译文文本的零个、一个或者多个句子,则A称为C,U的一个句子对齐。每个$|C_{iS}|:|U_{iT}|$称为句子的匹配模式。获得句子对齐A的概率为$Prob(A|C,U)$,则所要寻求的最佳句子对齐为

到稿日期:2009-05-20 返修日期:2009-08-05 本文受国家自然科学基金项目(60663006,60963017),新疆维吾尔自治区高等学校科学研究计划(XJEDU2009I05)资助。

田生伟(1973-),男,博士生,副教授,主要研究方向为计算机智能技术、计算机网络,E-mail:tianshengwei@163.com;吐尔根·依布拉音 教授,博士生导师,主要研究方向为自然语言处理、计算机智能技术;禹 龙 副教授,主要研究方向为计算机智能技术、计算机网络。

$$A=\arg\max_{A} Prob(A|C,U)$$

如果假定各翻译对之间独立并且不依赖上下文，上式转化为

$$A=\arg\max_{A}\prod Prob(A|C,U)=\arg\max_{A}\prod Prob(A_i|C_{iS},U_{iT})$$

假设 A_i 的概率只依赖于有限个属性 $\alpha_1,\alpha_2,\cdots,\alpha_m$ 的取值，则有

$$Prob(A_i|C_{iS},U_{iT})=Prob(A_i|\alpha_1(C_{iS},U_{iT}),\alpha_2(C_{iS},U_{iT}),\cdots,\alpha_m(C_{iS},U_{iT})) \quad (1)$$

1.1 基于长度的对齐模型

若 A_i 的概率只依赖于 C_{iS}，U_{iT} 的长度，则基于长度的对齐模型为

$$A=\arg\max_{A}\prod Prob(A_i|Len(C_{iS}),Len(U_{iT}))$$

由贝叶斯定理得

$$Prob(A_i|Len(C_{iS}),Len(U_{iT}))=\frac{Prob(Len(C_{iS}),Len(U_{iT})|A_i)*Prob(A_i)}{Prob(Len(C_{iS}),Len(U_{iT}))}$$

其中，分母 $Prob(Len(C_{iS}),Len(U_{iT}))$ 是一个常数，在计算时略去，因此上式转化为

$$Prob(A_i|Len(C_{iS}),Len(U_{iT}))=Prob(Len(C_{iS}),Len(U_{iT})|A_i)*Prob(A_i) \quad (2)$$

假设汉语文本对应维语译文长度的事件之间相互独立，依据概率学原理，汉语对应的维语译文长度的随机变量服从 $N(Len(C_{iS})c,Len(U_{iS})\sigma^2)$ 分布（c 是单位汉语字节长度对应的维语译文的字节数），则有

$$Prob(Len(C_{iS}),Len(U_{iT})|A_i)=Prob(\delta(LenC_{iS}),Len(U_{iT}))|A_i)=2(1-Prob(\delta))$$

其中，$Prob(\delta)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{|\delta|}e^{-\frac{x^2}{2}}dx$，$Prob(A_i)$ 是匹配模式的概率，可以从训练语料中获得。

基于长度对齐的模型，关键是评价函数 $\delta(Len(C_{iS}),Len(U_{iT}))$ 的设计。只有在 δ 近似满足标准正态分布时，该模型才可以获得较高的正确率。

利用文献[7]的评价函数对 $\delta(Len(C_{iS}),Len(U_{iT}))$ 进行改进，得

$$\delta=\frac{Len(C_{iT})-c*Len(U_{iS})}{\sqrt{Len(C_{iS})*S^2}}$$

其中，c 是单位汉语字节长度对应的维语译文的字节数，S^2 是 $(Len(C_{iS}),Len(U_{iT})-c*Len(C_{iS}))$ 线性回归的斜率，而不是印欧语系的 $(Len(C_{iS}),Len(U_{iT})-Len(C_{iS}))$ 线性回归的斜率，原因是印欧语系的 c 近似为 1，而汉维语的 c 值很大。改进后可以保证 δ 服从正态分布。

对人工对齐的 1000 个汉维句子对齐进行统计，得出 $c=4.43$，$S^2=19.26$。经检验，汉维句子对齐的长度相关系数达到了 0.98，在显著水平为 0.01 的假设检验中，δ 服从正态分布，说明评价函数是适合长度对齐的。δ 分布如图 1 所示。

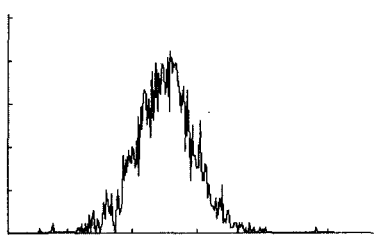

图 1 δ 分布

从训练语料中得出匹配模式 $Prob(A_i)$ 的概率如表 1 所列。

表 1 匹配模式概率

匹配模式	1-1	1-2	2-1	2-2	其它
匹配概率	0.932	0.041	0.012	0.006	0.009

匹配模式大部分是 1-1，1-2，2-1 和 2-2，同时有少量的其它匹配模式。

2 混合对齐模型

对 50000 个汉维语句子对齐的标点符号进行统计，得出表 2。

表 2 汉维语标点符号统计

维语标点	出现频率	汉语标点	出现频率
.	23823	。	20537
≪	4879	《	2538
≫	4877	》	2538
:	2031	：	2387
‘	68197	，	44942
(	6155	（	7015
)	9367	）	7019
!	108	！	94
%	1093	％	997
؛	3354	；	3184
؟	6.0	？	10
{	2	【	2
}	2	】	2
<	162	〈	5
>	163	〉	5
_	492	—	8
[	13	「	2
]	13	」	2
/	37	／	92
\	36	\	2914
-	565	-	172
\\	30	\\	0

经检测，汉维语标点符号的相关系数达到了 0.99，说明汉维语标点符号特征适合用于刻画句子对齐模型。

考虑句子长度和标点符号两种特征，由式(1)、式(2)得出混合对齐模型为

$$A=\arg\max_{A}\prod Prob(A|C,U)=\arg\max_{A}\prod Prob(A_i|C_{iS},U_{iT})$$

$$Prob(A_i|C_{iS},U_{iT})=Prob(A_i|(Len(C_{is}),Len(U_{iT})),Mark(C_{is},U_{iT}))$$

假设长度特征和句子标点符号特征相互独立，并利用二项式分布描述汉维句子的标点符号分布，则上式为

$$Prob(A_i|C_{iS},U_{iT})=Prob(Len(C_{is}),Len(U_{iT})|A_i)*Prob(A_i)*Prob(Mark(C_{is},U_{iT})) \quad (3)$$

而

$$Prob(Mark(C_{is},U_{iT}))=\binom{m}{n}probmark(C_{is},U_{iT})^n*(1-probmark(C_{is},U_{iT}))^{m-n}$$

其中，m 是 C_{is}，U_{iT} 中标点多的标点总数，n 是匹配标点的总数，$probmark(C_{is},U_{iT})$ 是汉维语标点匹配的概率，可以从表 2 中计算得到。

2.1 锚点发现模型

长度对齐模型虽然效率高，但由于仅考虑长度特征，忽略了句子词汇的互译特征，其对齐错误往往会蔓延，导致大范围的对齐错误；而利用词汇对齐，又会造成时间效率急剧降低，同时词汇对齐的数据稀疏问题也不容忽视。

从表 1 中可以看出，汉维语句子对齐的匹配模式大部分都是 1∶1，利用这个特点，可以事先在待对齐文本中寻找和确定若干个 1∶1 对齐的句对作为锚点，再利用混合对齐方法在锚点之间进行对齐。对齐过程如图 2 所示。

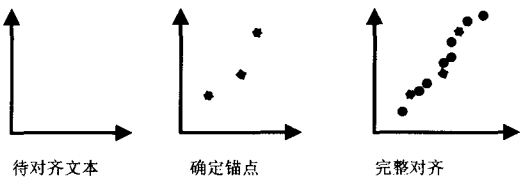

图 2　对齐过程

由于在对齐文本中设置了锚点，因此可以将对齐错误的蔓延抑制在锚点内部，有效避免了更大范围的扩散。

锚点的精确度是该模型的关键。如果锚点自身选错，不但不能抑制错误蔓延，反而会成为错误蔓延的发生源。因此，模型应当从多角度综合判别，以确保锚点句对的准确。

本文设计的锚点发现算法综合考虑了句子的长度、互译词汇信息和标点符号因素。

2.1.1　句子长度因子

给定汉维语两个句子 Cs 和 Us，判定其是锚点句对的句子长度因子为

$$factorLen(Cs,Us)=\delta=\frac{Len(Us)-c*Len(Cs)}{\sqrt{Len(Cs)*s^2}} \tag{4}$$

其中，c，S^2 的取值为 1.1 节中所述。

2.1.2　基于共现的汉维互译词汇信息抽取

实验发现，即便使用规模较大的电子词典，但由于词汇在词典的解释中的表述不一定和句子中的词汇完全一样，在汉维语互译的词汇数据仍然存在稀疏现象，使得在句子间的互译关系不强。如维语单词ئالدىراڭ对应的汉语词汇应该是“紧迫”，然而在电子词典中，查到的解释是“不拖延，急切，立即，马上”，虽能够从意义上看出词典解释和“紧迫”是相似的，但对应的汉语句子中并没有这些解释词汇，最终判定该互译词汇不存在。同时，由于电子词典的规模较大，造成搜索的时间效率也比较低。因此，可以采用统计的方法来获得汉维语互译词汇信息。

互信息(Mutual Information)

$$I(Cword,Uword)=\log_2\frac{Prob(Cword,Uword)}{Prob(Cword)Prob(Uword)}$$

其中，$Prob(Cword)$，$Prob(Uword)$分别表示汉维语词汇在语料中出现的概率，$Prob(Cword,Uword)$表示汉维语词汇在语料中对应出现的概率。上式表示源文和译文中两个词汇互译的可能性，值越大，互译的可能性越大。为了避免语料中个别词汇数目极少而 $I(Cword,Uword)$值较大，出现噪音，采用 t 检验[12]来消除。

$$t\approx\frac{Prob(Cword,Uword)-Prob(Cword)Prob(Uword)}{\sqrt{\frac{1}{M}Prob(Cword,Uword)}}$$

其中，M 是语料的总词汇量。取 $t=1.65$，即只有汉维词汇的 t 检验值大于 1.65 才作为互译词汇。

2.1.3　互译词汇信息因子

通过计算源文和译文句子互译信息来估计句子互为翻译的概率，作为判断源文和译文句子是锚点句对的一个因素。

令 Cs 和 Us 设长度分别为 cl 和 ul，汉语词汇 ch 的维语译文单词为 u，chi 和 uj 是没有在译文和源文中出现对应词的汉语和维语单词，则互译的概率为

$$factorMI(Cs,Us)=Probtrans(Cs,Us)=\frac{\sum_{t=1}^{l}(\sum(len(ch)+len(u)))^2}{\sum_{i=1}^{cl-l}len(Ch_i)^2+\sum_{j=1}^{ul-l}len(U_j)^2+\sum_{t=1}^{l}(\sum(len(ch)+len(u)))^2} \tag{5}$$

其中，$\sum(len(ch)+len(u))$是源文和译文互译的连续单词序列长度和，l 是连续单词序列的数量。可以看出，由于平方运算，连续单词随序列长度的增加而增长得快，反之则会较慢，突出了连续单词序列的区分能力。极端情况下，句子 Cs，Us 完全一致，则 $Probtrans(Cs,Us)=1$；若句子 Cs，Us 没有互译的单词，则 $Probtrans(Cs,Us)=0$。

2.2　锚点判定函数

综合考虑上述 3 个因素，一对汉维句子(Cs，Us)的锚点判定函数为

$$anchorRatio=penaltyDisFactor*(\lambda_1*factorLen(Cs,Us)+\lambda_2*factorMI(Cs,Us)) \tag{6}$$

其中，λ_1，λ_2 是影响系数，且 $\lambda_1+\lambda_2=1$；且 $factorLen(Cs,Us)\geqslant t_1$，$factorMI(Cs,Us)\geqslant t_2$；$t_1$ 和 t_2 是阈值；$penaltyDisFactor$ 是两个句子的距离惩罚因子，其值为

$$penaltyDisFactor=1-\frac{\left|\frac{chNo}{chSenCount}*uSenCount-uNo\right|}{uSenCount}$$

其中，$chNo$ 是待对齐文本的汉语句子编号，uNo 是维语句子编号，$chSenCount$ 是待对齐文本中的汉语句子总数，$uSenCount$ 是维语句子总数。

对齐算法采用动态规划算法实现。

3　噪音对齐发现

对齐算法将语料进行句子对齐后，会存在一些噪音对齐，如一个对齐中汉语或维语句子部分缺失。分析其原因，是生语料本身在翻译中存在翻译不完全的现象。还存在一些翻译质量低劣的汉维语对齐，这些都属于噪音对齐。如果不把这些噪音对齐消除掉，势必会对语料库在具体应用中造成干扰。因此，需要对自动对齐的汉维语句子进行正确性判断。

3.1　统计与维语词根抽取技术的词汇互译信息

大量实验发现，人工评价对齐的正确主要依靠汉维语句子中的词汇互译信息。根据这个原理，可以采用词汇互译信息作为评价对齐的重要因素。

除了第 2 节中提到的词汇解释造成的词汇数据稀疏现象，还发现在大规模汉维语电子词典中存在维语形态变化数据稀疏现象，如英语中单词 work 可以有 works，worked，working，worker 等形式。维语的单词形态变化比英语还要复杂，变化更多，一个单词在句子中不同的环境下存在多种不同的形式。由于这种原因，实验发现许多汉语词汇只能在电子词典中找到维语单词的词根，而译文中的维语单词是该词根的变化形态。这种现象进一步加剧了互译词汇信息的数据稀疏现象。

因此，采用本课题组开发的维语词根抽取技术来提高互译词汇信息的互译率：先对维语句子的每个单词进行词根抽取，再以抽取到的词根到电子词典中查找相应的汉语解释词汇。

为了进一步提高词汇互译率，可以利用第 2 节中的统计互译词汇信息，即先使用维语词根在电子词典中进行查找，若失败，则到统计互译词汇信息中查找。经 50000 个句子对齐试验，本方法词汇互译率在 40%以上的句子占总数的 78%，说明统计与维语词根抽取技术的词汇互译信息覆盖度完全可以满足实用需要。

3.2 对齐正确性评价算法

$$evaluationFun(Cs,Us)=\eta_1 * factorLen(Cs,Us)+\eta_2 * factorWordMI(Cs,Us)$$

其中，$factorLen(Cs,Us)$是长度因子，$factorWordMI(Cs,Us)$是统计与维语词根抽取技术的词汇互译信息影响因子，计算方法同第 2 节中的 $factorMI(Cs,Us)$，其中 η_1，η_2 是影响系数，且 $\eta_1+\eta_2=1$。

当一个汉维语对齐的评价函数 $evaluationFun(Cs,Us)$值低于一个阈值 Threshold 时，则判断该对齐是噪音对齐。

虽然基于统计与维语词根抽取技术的词汇互译信息会在电子词典和统计互信息表中进行频繁的查找，但由于本方法仅对一个对齐评价，因此不会发生在自动对齐中动态规划算法产生高时间复杂度。

4 实验与分析

从新疆大学多语种研究中心整理的汉维平行语料中，随机抽取 4 篇人工对齐句对的文献，分别采用基于长度对齐、混合策略对齐和错误抑制的混合策略 3 种方法进行实验。实验结果如表 3—表 5 所列。

分析实验数据，基于长度的对齐在汉维语句子的长度关系符合相应的比例，效果好。相反则很容易引发对齐错误，并导致错误蔓延。

混合对齐策略利用长度和标点两种特征，在一定程度上更精确地刻画了对齐的汉维语句子，因此对齐的错误率比长度对齐有所降低。

错误抑制的混合策略的对齐方法，由于采用了锚点发现算法，有效地防止了错误的蔓延，因此其错误率是 3 种方法中最低的。

图 3 是包含 6 个汉语句子、7 个维语句子的文本对齐。基于长度对齐的结果如图 3(a)所示，而正确的对齐如图 3(b)所示。分析原因发现，汉语文本的第 2 个句子后半部分缺失，导致句子长度变短；基于长度的对齐算法仅考虑句子的长度信息，因此错误地将汉语第 1 句和维语第 1、2 句对齐，最终导致该错误蔓延到后续的所有句子。而采用错误抑制的混合策略的对齐方法，首先寻找锚点句对，发现 2 个汉维锚点句对(1-1)和(5-5)，如图 3(c)所示。然后在这两个锚点其它部分采用混合策略对齐，对齐结果完全正确。

表 3 长度对齐

测试文件(句对)	正确率(%)	召回率(%)
文件 1(233)	94.6	97.4
文件 2(286)	100	100
文件 3(253)	93.0	94.9
文件 4(269)	92.4	94.8
合计	95.0	96.8

表 4 混合对齐

测试文件(句对)	正确率(%)	召回率(%)
文件 1(233)	96.2	98.3
文件 2(286)	100	100
文件 3(253)	94.2	95.7
文件 4(269)	94.1	95.6
合计	96.1	97.4

表 5 错误抑制的混合对齐

测试文件(句对)	正确率(%)	召回率(%)
文件 1(233)	97.4	98.3
文件 2(286)	100	100
文件 3(253)	98.0	98.4
文件 4(269)	94.9	95.9
合计	97.6	98.2

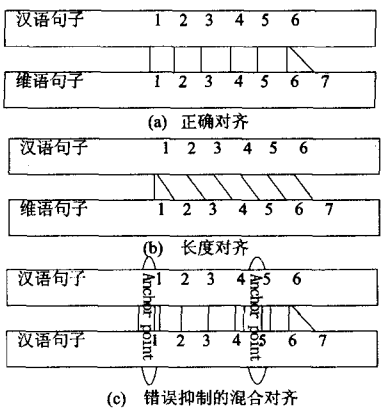

图 3 对齐结果

利用噪音对齐发现算法对 4 个文件的自动对齐进行评价，结果如表 6 所列。

表 6 噪音对齐发现

噪音对齐(句对)	发现的噪音对齐(句对)	正确率(%)	召回率(%)
文件 1(3)	2	100	66.7
文件 2(2)	2	100	100
文件 3(0)	0	100	100
文件 4(3)	3	66.7	66.7
合计	7	91.68	83.35

如图 3 所示，由于汉语文本的第 2 个句子后半部分缺失造成的噪音对齐(2:2)，噪音对齐发现算法发现并给予了标记。

结束语 本文提出了一种改进的方法来处理汉维句子对齐。该方法利用双语语料的词汇共现信息，自动抽取汉维语词汇搭配，结合句子长度等特征，寻找 1∶1 模式的句对作为锚点，将错误蔓延抑制在锚点内；利用标点符号和长度特征的混合方法进行句子对齐。经实验验证，该算法寻找的锚点的精度高。混合对齐算法能够避免基于词汇对齐算法的高时间复杂度的弱点，提高了时间效率和对齐准确度。

将来的工作是继续引入通用词典、领域文献的特征等，作为对齐的依赖属性，进一步优化对齐算法。

参 考 文 献

[1] Dolan W B, Pinkham J, Richardson S D. The Microsoft Research Machine Translation System[J]. AMTA, 2002: 237-239

[2] Wu D, Xia X. Large-scale automatic extraction of an English-Chinese translation lexicon [J]. Machine Translation, 1995, 9 (3/4): 285-313

(下转第 292 页)

(e)—图 6(h)充分地展示了本文实时积雪场景效果的剖面图。

图 6(e)—图 6(h)是地表类型为植被、泥土分别在地表温度是—1℃,1℃时积雪场景的模拟情况。对比可看出在相同温度和等降雪量情况下,由于泥土导热率较高,热融化量较大,故地表积雪密度较小,场景雪层较稀薄;而纵向对比在相同地表类型中,地表温度越低,等降雪量的情况下融雪量变化微乎其微,积雪密度越大,场景中的雪层越厚实,符合上述的结论和预期效果。实验表明,积雪可由地表温度和地面类型得到实时控制。

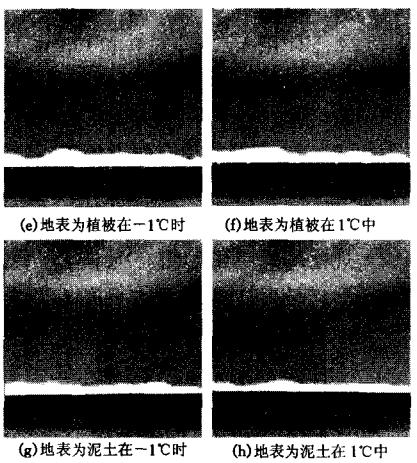

图 6 不同温度与地表植被雪场景剖面图

结束语 用衡量自然景物模拟的重要指标即真实性和实时性来考察所做的绘制工作:

1)将类衰减正弦曲线方程应用到视景飘雪模拟中,增强了整体效果的细致感,与粒子系统的完美结合,真实地再现了自然飘雪场景。与传统计算每个粒子的属性相比,提高了系统的实时性;

2)为更好模拟出不同地表以及不同温度对地表积雪密度的影响,引入积雪-融雪无损热传递方程,使场景仿真更加细致、真实,提高了系统的真实性。实验表明,本方法是合理且可行的。

3)GPU 的性能和可编程能力为计算机图形学的模拟提供了良好的平台,使系统在保证真实性的同时提高了实时性。

本方法便于推广,不仅可用于场景的模拟,更可用于真实自然灾害的预测和模拟,有效防范于未然。

进一步考虑飘雪场景中复杂的风场对雪粒子运动的影响,同时对积雪场景引入光照模型,考虑太阳辐射对温度的影响,能更有效地增强雪场景模拟的真实感。这些尚有待进一步探讨和研究。

参 考 文 献

[1] 宋成芳,于洋,杨颖振. 数据驱动的大规模森林场景真实感动画[J]. 计算机辅助设计与图形学学报,2008,20(8):1015-1022

[2] 任鸿翔,尹勇,金一丞. 大规模海浪场景的真实感绘制[J]. 计算机辅助设计与图形学学报,2008,20(12):1617-1622

[3] Rriter C. A local cellular model for snow crystal growth [J]. Chaos, Solitons & Fractals, 2005, 23(4): 1111-1119

[4] Gravner J, Griffeath D. Modeling snow crystal growth II: A mesoscopic lattice map with plausible dynamics [J]. Physica D: Nonlinear Phenomena, 2008, 237(3): 385-404

[5] Langer M S. A Spectral-particle Hybrid Method for Rendering Falling Snow [M]. Canada: McGill University, 2004

[6] 陈华杰,余小清. 基于粒子系统与 LOD 技术的实时雨雪效果模拟[J]. 计算机仿真,2008,25(4):491-494

[7] Fearing P. Computer Modelling of Fallen Snow [A] // SIGGRAPH 2000[C]. 2000: 37-46

[8] 陈彦云,孙汉秋,郭百宁. 自然雪景的构造和绘制[J]. 计算机辅助设计与图形学学报,2003,22(1):916-922

[9] Ohlsson P. Real-time Rendering of Accumulated Snow [D]. Sweden: Uppsala University, 2004: 1-19

[10] Saltvik I. Parallel Methods for Real-time Visualization of Snow [D]. Norway: Norwegian University of Science and Technology, 2006: 218-227

[11] 张东阳. 基于温度变化与粒子系统的雪景模拟算法研究与实现[D]. 河北:燕山大学,2008:26-40

[12] 黄庚,苏正军. 冰雪晶碰并勾连增长的实验与观测分析[J]. 应用气象学报,2007,18(4):561-567

[13] Stanley H E. Two-dimensional polymers and conformal invariance [J]. Physic A, 1990, 163(1): 158-182

[14] 范爱武,刘伟. 不同环境条件下土壤温度日变化的计算模拟[J]. 太阳能学报,2003,24(2):167-171

(上接第 218 页)

[3] Fattah M A, Ren Fuji, Shingo K. Adaptive Threshold Parameters for Bilingual Dictionary Extraction from the Internet Archive[J]. International Journa Information, 2005, 8(1): 165-175

[4] Dejean H, Gaussier E, Sadat F. Bilingual Terminology Extraction: An Approach based on a Multilingual thesaurus Applicable to Comparable Corpora[C] // Proceedings of the 19th International Conference on Computational Linguistics COLING. Taipei, Taiwan, 2002: 218-224

[5] Chuang T C, Yeh K C. Aligning Parallel Bilingual Corpora Statistically with Punctuation Criteria[J]. Computational Linguistics and Chinese Language Processing, 2005, 10(1): 95-122

[6] Brown P F, Lai J C, Mercer R L. Aligning sentences in parallel corpora[A] // Proceedings of 29th Annual Meetingof the Association for Computational Linguistics Berkeley[C]. CA: ACL, 1991: 169-176

[7] Gale W A, Church K W. A program for aligning sentences in bilingual corpora[J]. Computational Linguistics, 1993, 19(1): 75-102

[8] Kay M, Roscheisen M. Text-translation alignment [J]. Computational Linguistics, 1993, 19(1): 121-142

[9] Wu D. Aligning a parallel English-Chinese corpus statistically with lexical criteria[A] // Proceedings of the 32th Annual Conference of the Association for Computational Linguistics. Las Cruces[C]. NM: ACL, 1994: 80-87

[10] 张艳,柏冈秀纪. 基于长度的扩展方法的汉英句子对齐[J]. 中文信息学报,2005,19(5):35-36

[11] 李维刚,刘挺,张宇,等. 基于长度和位置信息的双语句子对齐方法[J]. 哈尔滨工业大学学报,2006,38(5):689-692

[12] Haruno M, Yamazaki T. High-performance bilingual text alignment using statistical and dictionary information[C] // Proceedings of the 34th Annual Meeting on Association for Computational Linguistics. 1996: 131-138