融合特征约束模型的纳西-汉语双语词语对齐算法

张涛[1]，余正涛[1,2]，郭剑毅[1,2]，曹先彬[3]
(1.昆明理工大学信息工程与自动化学院，650051，昆明；2.昆明理工大学智能信息处理重点实验室，650051，昆明；
3.北京航空航天大学电子信息工程学院，100191，北京)

摘要：针对纳西语、汉语因句法结构差异较大而导致双语词语自动对齐较为困难的问题，提出一种融合特征约束模型的纳西-汉语双语词语对齐算法.首先在语料中统计纳西-汉语词语区间扭曲和位置转换特性，并由此建立2个双语词语对齐的特征约束模型；然后将提出的特征约束模型融入词语对齐的对数线性模型框架，并结合最小错误率算法训练模型参数；最终搜索出最佳的词语对齐结果.实验以IBM Model3为词语对齐比较模型，结果表明，该双语词语对齐算法可以使纳西-汉语词语的对齐准确率提升21.9%.

关键词：词语对齐；纳西；汉语；特征约束模型
中图分类号：TP391 **文献标志码**：A **文章编号**：0253-987X(2011)10-0048-06

A Bilingual Word Alignment Algorithm of Naxi-Chinese Based on Feature Constraint Models

ZHANG Tao[1], YU Zhengtao[1,2], GUO Jianyi[1,2], CAO Xianbin[3]
(1. The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China;
2. The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming 650051, China;
3. School of Electronics and Information Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: A bilingual word alignment algorithm of Naxi-Chinese based on feature constraint models is proposed to reduce the difficulty of bilingual word alignment for Naxi-Chinese which has huge difference in syntactic structure. Two feature constraint models- interval distortion model and position transformation model are established by counting the traits of interval distortion and position transformation in corpus, and are integrated into a log-linear framework of word alignment. Then parameters in the models are trained using the minimum error rate algorithm and the best alignment results are eventually searched. Experimental results on IBM Model3 show that the proposed algorithm increases the word alignment accuracy of Naxi-Chinese about 21.9%.

Keywords: word alignment; Naxi; Chinese; feature constraint model

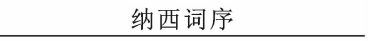
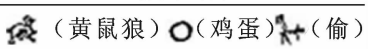
我国是一个多民族国家，少数民族语言和文字极大地丰富了我国传统文化，对少数民族语言进行的信息化处理也越来越多.通过比较可以发现，很多民族语言在语法上和汉语相比很有特点，如在壮族语言中定语后置，塔塔尔族语言中宾语在谓语的前面，而在水族语言中表示修饰和领属关系的词语顺序和汉语相反，这些特点可以为少数民族语言的智能计算和处理提供特定的研究条件.基于此种思路，本文以实现纳西-汉语词语自动对齐为目的，在深入分析语言特性的基础上，提出一种结合语言特性的词语对齐算法，有效提高了词语对齐的准确率.

纳西文是由云南丽江纳西族先民创造并使用的

收稿日期：2011-03-08. 作者简介：张涛(1986—)，男，硕士生；余正涛(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(60863011)；云南自然科学基金重点资助项目(2008CC023).
网络出版时间：2011-09-06 网络出版地址：http://www.net/kcms/detail/61.1069.T.20110906.1656.002.html

文字,是目前世界上唯一仍在使用中的象形文字,与汉语相比,在语法上有很多不同之处. 例如,纳西语言基本语序是主-宾-谓,即纳西语具有动词置后的特点. 在一些情况下,纳西-汉语句对还有动词成分和名词成分在位置上对应的特点,表 1 为汉语和纳西语法语序区别示例.

表 1　汉语和纳西语法语序区别示例

汉语词序	纳西词序
黄鼠狼 偷 鸡蛋	
爸爸 打 儿子	

另外,纳西象形文中没有助词、介词、叹词、副词等等. 这些句法上的巨大差异导致纳西-汉语词语自动对齐相对其他语言更为困难.

双语词语对齐最早是由 Brown 等[1]作为机器翻译的中间隐含过程提出来的,采用最大期望(expectation-maximization,EM)算法估计词语对齐概率. Vogel 等[2]提出基于先序(hidden Markov model,HMM)的词语对齐模型,其最终的对齐概率依赖于对齐词语位置的差异而不是词语在句子中的绝对位置,该方法不能根据具体语言特性而扩展. 近年来,词语对齐的任务逐步被单独提出来,多数方法均采用了基于可集成特征的框架,在框架下针对不同的语言特性集成不同的特征模型,如文献[3-7]均提出了相应的基于特征框架的判别方法,通过集成有益词语对齐的特征来提高词语对齐的效果. 基于特征框架的方法在词语对齐上已取得了较好的效果,该方法亦可应用于纳西-汉语双语词语对齐,但采用此种方法时,需根据具体语言特点设计相应的特征函数.

本文结合纳西-汉语句法结构差异大的特点,提出区间扭曲模型和位置转换模型 2 个特征模型,在对数线性框架下,从语料库中抽取统计特征并结合最小错误率算法训练特征参数,实现了纳西-汉语句子对齐语料下词语的自动对齐,为纳西-汉语语料库建设和纳西-汉语统计翻译提供了基础支持.

1　双语词语对齐定义

对数线性模型是训练双语词语对齐的有效方法,可以将有益于词语对齐的各种特征函数综合起来,汇集成语言知识,针对不同语言对的情况,增加或者改变相应特征函数,并在统一的框架下训练,而不必考虑和其他特征函数之间的联系[8].

1.1　双语词语对齐定义及建模

首先对词语对齐做出定义[6-7]. 定义源语言句子为词序列 $s=s_{1(I)}=s_1,\cdots,s_i,\cdots,s_I$,目标语言句子为词序列 $t=t_{1(J)}=t_1,\cdots,t_j,\cdots,t_J$. I 和 J 分别为 s、t 的词语数量. 若 s_i 和 t_j 互为翻译项,即互相对齐,这种关系表示为 $l(i,j)$,亦即 s_i 和 t_j 之间有连线. 最正确的双语词语对齐关系 α 是 $l(i,j)$ 的笛卡尔子集,即 $\alpha\subseteq\{l:i=1,\cdots,I;j=1,\cdots,J\}$.

然后对双语词语对齐关系 α 建模. 由于双语句子 s 和 t 为词语对齐基本资源,则词语对齐的目标便是求某个对齐关系 α,使得对齐模型 $P_r(\alpha|s,t)$ 最大. 采用对数线性模型[9]对 $P_r(\alpha|s,t)$ 建模为

$$P_r(\alpha\mid s,t)=\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\alpha,s,t)\right]}{\sum_{\alpha'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\alpha',s,t)\right]}\tag{1}$$

式中:$P_r(\alpha|s,t)$ 为 s 和 t 的对齐关系 α 的概率;$h_m(\alpha,s,t)$ 为对 s 和 t 的对齐进行建模的特征函数,其中 $m=1,\cdots,M$,M 为特征个数;λ_m 为 $h_m(\alpha,s,t)$ 的权重参数;α' 是句子 s 和 t 中词语所有可能的对齐. 双语词语对齐的目标便是求解特征函数权重 λ_m 使双语词语对齐的概率最大,即使 $P_r(\alpha|s,t)$ 最大. 由于(1)式的分母部分对最终的搜索结果没有影响,则双语词语对齐的搜索模型为

$$\hat{\alpha}=\arg\max_{\alpha}\left\{\sum_{m=1}^{M}\lambda_m h_m(\alpha,s,t)\right\}\tag{2}$$

文献[10-11]等的研究说明,增加一些和特定语言相关的特征有助于提高对齐的准确率,如词性、词典等特征. 增加的特征变量以 x 来表示,在本文中即为位置转换和区间扭曲特性,那么式(1)变为

$$P_r(\alpha\mid s,t,x)=\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\alpha,s,t,x)\right]}{\sum_{\alpha'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\alpha',s,t,x)\right]}\tag{3}$$

式(3)对应的搜索模型为

$$\hat{\alpha}=\arg\max_{\alpha}\left\{\sum_{m=1}^{M}\lambda_m h_m(\alpha,s,t,x)\right\}\tag{4}$$

1.2　融入新的特征函数的必要性

IBM 模型因其通用性而通常作为双语词语对齐的基础性特征. 然而,由于纳西语言和其他语言之间结构的较大差异性,导致单纯采用 IBM 模型为特征函数很难取得很好的效果,这主要是因为 IBM 模型在计算词语扭曲概率时是基于整个句子范围的,而纳西语言在词语扭曲范围上相比于其他语言要小

得多. 基于此,本文根据纳西语言的特点,提出2个基于纳西语言特性的特征约束模型,并将其融入到对数线性模型框架中,以实现纳西-汉语双语词语自动对齐.

2 纳西-汉语词语对齐特征模型

当前在词语对齐和翻译领域中,IBM模型多作为基础工作或者比较模型,而其中尤以IBM Model3更为常用,故本文以IBM Model3为基础特征,同时逐步增加针对纳西语言特点来设计特征函数.

2.1 IBM Model3模型

若$s_{1(I)}$和$t_{1(J)}$分别为源语言和目标语言句子词语序列,其中I和J为s、t的词语数量,Brown等[1]将词语对齐α作为隐含变量引入,用以描述源语言句子s和目标语言句子t之间的关系,即源语言词语位置j到目标语言词语位置$i=\alpha_j$的对齐关系,建模[1]如下

$$P_r(s_{1(I)} \mid t_{1(J)}) = \sum_{\alpha_{1(J)}} P_r(s_{1(I)}, \alpha_{1(J)} \mid t_{1(J)}) \quad (5)$$

本文使用常用的IBM Model3为基础模型[1],进行对数处理后形成词语对齐的基础特征函数

$$h_m(\alpha, s, t) = \log(P_r(s_{1(I)}, \alpha_{1(j)} \mid t_{1(J)})) \quad (6)$$

不同对齐方向的IBM Model3可视作不同的特征:源语言可以是汉语或者纳西语,目标语言可以是纳西语或者汉语.

IBM Model3不关心源语言和目标语言的具体语种,在计算源语言词语到目标语言词语的扭曲分布时,是基于源语言和目标语言整个句子的,这样在理论上具有普适性,但也带来了计算量大、扭曲概率计算不够精确等问题,并导致实际排序效果不好. 因此,本文提出2个基于纳西语言特点的特征函数,以对双语词语对齐产生更好的约束效果.

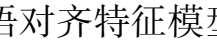
2.2 基于纳西语言特点的特征函数

为了提高纳西-汉语双语词语对齐的准确率,提出以下2个基于纳西语言特点的特征模型:区间扭曲模型和位置转换模型.

2.2.1 区间扭曲模型 在纳西语言与汉语对齐中,最明显的句法特点是区间扭曲,也就是词语对齐具有区间性特点. 对纳西-汉语双语的翻译来说汉语句子翻译成纳西语言句子时,纳西语言中动词总是置后,而名词和形容词等放在动词之前. 表2为纳西-汉语区间扭曲特点示例.

因此,可以把纳西语句子分成2块,第1块是由名词和形容词等,即非动词组成的前半部分,记为n_t. 第2块是由动词组成的句子后半段,记为v_t.

表2 纳西-汉语区间扭曲特点示例

汉语词序	纳西词序
我看见一只猫	
在抓老鼠	

对于汉语-纳西句对来说,定义汉语句子词性标记序列为$s=s_{1(n)}=sT_1,\cdots,sT_i,\cdots,sT_n$,纳西句子词性序列为$t=t_{1(m)}=tT_1,\cdots,tT_i,\cdots,tT_m$,其中$n$和$m$为$s$、$t$的词语数量. 于是根据纳西语言动词置后的语法特点,可将纳西词性序列拆分为$n_t=tT_1,\cdots,tT_i,\{i \mid i>=0, i<m\}$和$v_t=tT_j,\cdots,tT_m,\{j \mid j=i+1, j<m\}$两部分,其中$v_t$表示句子中的动词部分,$n_t$为在$v_t$之前的部分,即非动词部分.

那么一个汉语动词扭曲到纳西句子v_t块的概率是

$$P_r(v_t \mid s_i) = \frac{\mathrm{count}(s_i, v_t)}{\mathrm{count}(s_i)} \quad (7)$$

式中:$\mathrm{count}(s_i)$表示汉语动词词性标记s_i出现的次数;$\mathrm{count}(s_i, v_t)$表示s_i扭曲到纳西句子v_t区间的次数. 汉语动词扭曲到纳西语言句子v_t的扭曲模型定义为

$$P_r(v_t \mid \beta, s) = \prod_{l \in \beta} f(v_t \mid s_{l(i)}) \quad (8)$$

式中:l表示对齐关系;$l(i)$表示源语言汉语中第i个动词扭曲到纳西的v_t区间;$\beta=\{i: s(i)->v_t\} \subseteq \alpha$表示第$i$个词语扭曲到$v_t$区间. 同样地,若汉语词语为非动词,则扭曲模型为

$$P_r(n_t \mid \gamma, s) = \prod_{l \in \gamma} f(n_t \mid s_{l(i)}) \quad (9)$$

式中:$\gamma=\{i: s(i)->n_t\} \subseteq \alpha$表示第$i$个词语扭曲到$n_t$区间. 因此,可以在IBM Model3中的扭曲模型的基础上增加基于目标语言是纳西语言的区间扭曲模型,用以约束IBM对齐结果,则对于源语言汉语动词的特征函数可以设计为

$$h(\beta, s, t, v_t, s_{1(n)}) = \log\Big(\prod_{l \in \beta} f(v_t \mid \beta, s_{l(i)})\Big) \quad (10)$$

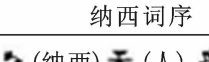
同时,对纳西非动词区间来说有

$$h(\gamma, s, t, v_n, s_{1(n)}) = \log\Big(\prod_{l \in \gamma} f(v_n \mid \gamma, s_{l(i)})\Big) \quad (11)$$

这里所提出的区间扭曲模型,在词语对齐时可以根据词语类别,将词语对齐的位置分布限定在一个区间之内,从而提高对齐的准确率.

2.2.2　位置转换模型　目前在对纳西-汉语双语文本进行分析时,由于语料句子长度相对较短(多数集中在 15 个字以下),在一些情况下会有源语言中第 N 个名词往往对应于目标语言的第 N 个名词,而第 M 个动词对应于目标语言中第 M 个动词的现象. 表 3 为纳西-汉语位置转换特点示例,其中动词和名词在位置上是对应的.

表 3　纳西-汉语位置转换特点示例

汉语词序	纳西词序
纳西 人 喜欢 唱歌 跳舞	 (纳西)(人)(唱歌) (跳舞)(喜欢)

根据这种现象,可以增加纳西-汉语双向共 4 个特征函数以反映这种统计特性,也即位置转换模型. 对于非动词表示为 $P_r(n_t|n_s)$,而动词表示为 $P_r(v_t|v_s)$. 这里的源语言(纳西或者汉语)词语可以是动词或者名词.

对于某个源语言动词或者名词,其位置扭曲到目标语言词语的位置的概率计算如下

$$P_r(n_{tj}\mid n_{si})=\frac{\mathrm{count}(n_{tj},n_{si})}{\mathrm{count}(n_{si})} \tag{12}$$

式中:$\mathrm{count}(n_{si})$ 表示在纳西-汉语双语语料库中源语言第 i 个名词出现的次数;$\mathrm{count}(n_{tj},n_{si})$ 表示语料库中源语言第 i 个名词对齐到目标语言第 j 个名词的次数. 对于动词来说,情况亦同.

定义 $s=s_{1(I)}=sT_1,\cdots,sT_i,\cdots,sT_I$,和 $t=t_{1(J)}=tT_1,\cdots,tT_j,\cdots,tT_J$ 分别为源语言(汉语或者纳西)句子 s 和目标语言(纳西或者汉语)句子 t 的词性标记序列,I 和 J 为 s、t 的词语数量,则纳西-汉语名词词性转换模型定义如下

$$P_r(n_t\mid\beta,n_s)=\prod_{l\in\beta}f(n_{tl(j)}\mid n_{sl(i)}) \tag{13}$$

式中:β 表示纳西-汉语双语句对词语为名词时的词语对齐 $\beta\subseteq\{g:T(i)=N;T(j)=N;i=1,\cdots,I;j=1,\cdots,J\}\subseteq\alpha$ 的 2 个方向抽取合并,即

$$\beta\subseteq\{g:s(i)=t(j)=N;\ i=1,\cdots,I;\ j=1,\cdots,J\}\subseteq\alpha$$

那么纳西-汉语双语双向名词位置转换的特征函数可以设计为

$$h(\beta,s,t,n_t,n_s)=\log\Big(\prod_{l\in\beta}f(n_{tl(j)}\mid n_{sl(i)})\Big) \tag{14}$$

同样地,纳西-汉语双语双向动词位置转换的特征函数可以设计为

$$h(\gamma,s,t,v_t,v_s)=\log\Big(\prod_{l\in\gamma}f(v_{tl(j)}\mid v_{sl(i)})\Big) \tag{15}$$

γ 与 β 意义相同,只是对齐词语为动词. 位置转换模型在纳西-汉语句对中的动词、名词的不同翻译方向共可形成 4 个特征函数. 文献[6-7]在对汉-英作词语对齐研究时提出双语词性的转换模型,本文提出的位置转换模型在思想上与之较为类似,但与文献[6-7]的词性转换模型不同的是,纳西-汉语位置转换模型是与纳西语言相关的,不具有普适性,但会有更好的效果,这得益于纳西语言本身的特点. 位置转换模型是个双向特征,即将双语不同方向分别作为不同的特征函数,源语言可以是汉语或者纳西语,目标语言可以是纳西语或者汉语.

3　参数训练及搜索

对于上节提出的纳西-汉语双语词语对齐特征模型,本节采用最小错误率算法对其参数进行训练,并给出搜索过程.

3.1　模型参数训练

在人工标注纳西-汉语双语训练集的基础上,采用最小错误率训练模型权重参数[12]. 按照 Och 等[9]的思想,定义源语言句子集合 S_s 和目标语言句子集合 S_t,则在训练集上总错误数量定义为 $E(S_s,S_t)=\sum_{i=1}^{\mathrm{count}(S)}E(S_{si},S_{ti})$,其中 S_{si} 和 S_{ti} 为集合 S_s 上的一对句子,$E(S_{si},S_{ti})$ 为该句对词语对齐的错误数. 求解目标是在这个训练集上,迭代模型参数 λ 使总的纳西-汉语词语对齐错误数最小,即

$$\lambda_M=\arg\min_{\lambda_M}\Big\{\sum_{i=1}^{\mathrm{count}(s)}E(\hat{S}_{ti}(\hat{S}_{si};\lambda_M),\hat{S}_{si})\Big\} \tag{16}$$

同时有

$$S_{ti}(S_{si};\lambda_M)=\arg\max_s\Big\{\sum_{m=1}^{M}\lambda_m h_m(a,s,x)\Big\} \tag{17}$$

式中:$s\in S_s$.

在迭代求解过程中,根据纳西-汉语特征函数的权重 λ,取搜索到的前 N 个最好的纳西-汉语词语对齐结果,即 N-best 列表. 在纳西-汉语词语对齐特征函数权重 λ 的 M 维空间中,固定 $M-1$ 维,迭代单维权值,直到 N-best 列表不变,具体过程如下:

(1)人工给纳西-汉语词语对齐模型参数 λ 赋予初值,然后以此初始模型,搜索得到纳西-汉语词语

对齐 N-best 列表；

(2)迭代纳西-汉语词语对齐模型参数为 λ_H，并以新的模型参数 λ_H 产生新的 N-best 列表，并与 1 中产生的 N-best 列表合并，产生新的扩展的 EN-best；

(3)重复(1)、(2)步骤，直到 EN-best 列表不变，纳西-汉语词语对齐模型参数迭代结束.

在参数训练完毕后，在测试集即可进行搜索.

3.2 搜索过程

采用基于栈的搜索算法从纳西-汉语词语对齐空间中搜寻概率最大的对齐结果. 搜索过程就是不断往初始栈中增加纳西-汉语词语对齐关系，并记录增加后的概率，最终选取概率最大的对齐关系.

具体计算方法为：根据公式(3)，对某个纳西-汉语词语对齐 α_i，其对齐概率为

$$P_r(\alpha_i \mid s,t,x)=\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\alpha_i,s,t,x)\right]}{\sum_{\alpha_i'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\alpha_i,s,t,x)\right]} \tag{18}$$

那么，增加一条纳西-汉语词语对齐 l，则有

$$P_r(l\cup\alpha_i \mid s,t,x)=\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(l\cup\alpha_i,s,t,x)\right]}{\sum_{\alpha_i'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\alpha_i,s,t,x)\right]} \tag{19}$$

重复以上过程，不断增加新的纳西-汉语对齐关系 l，直到所有词语均对齐，最终选取概率最大的对齐结果.

在对齐结果中，纳西和汉语均需经过分词处理. 由于纳西文没有词法分析系统，只以空格分开每个纳西字符. 图 1 为纳西-汉语词语对齐结果示例.

图 1 纳西-汉语词语对齐示例

词语下标从 0 开始. 由于纳西语言中没有副词、叹词、介词、助词等，故在汉语词法分析过程中如遇到此类词语时，则自动对齐到空格，在最终的对齐结果中用 # 表示.

4 实验与结果分析

本节以 IBM Model3 为基础比较模型，在逐步增加对齐特征后计算词语对齐的错误率 R(R 为错误词语对齐数除以总词语对齐数)，并给出在纳西-汉语平行语料库上词语对齐的实验结果和分析. 实验使用了开发集、测试集和训练集，其中开发集和测试集为人工标注的纳西-汉语词语对齐语料. 表 4 为本文的训练集、开发集和测试集中包含词语数量的统计数据.

表 4 训练集、开发集和测试集中包含词语的数量

数据集类型	纳西-汉语句对数	汉语词数	纳西词数
训练集	9 413	37 581	35 374
开发集	500	4 703	3 987
测试集	500	4 681	4 015

在纳西-汉语词语双语对齐过程中，以 IBM Model3 特征的模型为 Baseline，并在以 IBM Model3 作为特征函数的基础上逐步增加本文定义的 6 个特征函数(位置转换模型为双向特征). 若用 IBM 表示 IBM Model3，S 表示区间扭曲模型，L 表示位置转换模型，CN 表示从汉语到纳西方向，NC 表示从纳西到汉语方向，N 表示非动词，V 表示动词，那么，“+IBMCN”则表示增加 IBM Model3 的汉语(C)到纳西(N)的方向为特征函数，其他类推. 实验以 IBM Model3 为基础特征，并作为比较对象. 表 5 给出了在实验过程中每增加一个特征后，其对齐错误率 R_i 的变化情况，i 为总特征数量.

表 5 逐步增加特征后的错误率 R_i 变化情况

增加的特征函数	R_i
+IBMCN	0.450 1
+IBMNC	0.436 7
+SCNV	0.399 1
+SCNN	0.391 3
+LCNV	0.379 5
+LCNN	0.362 7
+LNCV	0.321 1
+LNCN	0.314 0

从实验结果来看，在增加了 LNCV 和 SCNV 这 2 个特征后，对齐错误率 R_i 有较为显著的降低，这

也体现了位置转换模型中的纳西到汉语中动词的位置特性. 这说明区间扭曲模型中汉语到纳西语中的动词区间特性对纳西-汉语双语词语对齐起到了重要作用，其他方向的特征亦对对齐结果起到积极作用. 以 IBM Model3 为比较对象，在增加基于纳西语言特点的特征函数后，对齐错误率降低了 0.095 7，对齐准确率提升了 21.9%.

图 2 为在测试集上的搜索时间曲线. 从图 2 可以看出，随着特征函数数量的逐步增加，需要更长的搜索时间.

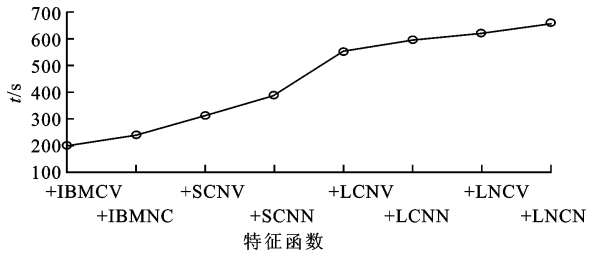


图 2　在测试集上的搜索时间曲线

5　结　论

本文基于纳西语言特点提出 2 个特征模型，用以实现纳西-汉语双语词语自动对齐. 以 IBM Model3 为比较对齐模型，逐步将提出的特征函数加入到对数线性模型框架下，在纳西-汉语平行语料库上采用最小错误率算法训练模型参数，并搜索最佳对齐结果. 最终的实验结果表明，本文提出的对齐算法可以明显地改善纳西-汉语词语对齐的效果.

参考文献：

[1] BROWN P F, PIETRA D V J, PIETRA D S A, et al. The mathematics of statistical machine translation: parameter estimation [J]. Computational Linguistics, 1993, 19(2):263-311.

[2] VOGEL S, NEY H, TILLMANN C. HMM-based word alignment in statistical translation [C]// Proceedings of the 16th International Conference on Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996: 836-841.

[3] TASKAR B, LACOSTE-JULIEN S, KLEIN D. A discriminative matching approach to word alignment [C]// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Porcessing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005: 73-80.

[4] MOORE R. A discriminative framework for bilingual word alignment [C]// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005: 81-88.

[5] CHERRY C, LIN D. A probability model to improve word alignment [C]// Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003: 88-95.

[6] LIU Yang, LIU Qun, LIN Shouxun. Log-linear models for word alignment[C]// Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005: 459-466.

[7] LIU Yang, LIU Qun, LIN Shouxun. Discriminative word alignment by linear modeling [J]. Computational Linguistics, 2010, 36(3): 303-339.

[8] AYAN N F, DORR B J. A maximum entropy approach to combining word alignments [C]// Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006: 96-103.

[9] OCH F J, NAY H. Discriminative training and maximum entropy models for statistical machine translation [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 295-302.

[10] 刘洋. 树到串统计翻译模型研究[D]. 北京：中国科学院计算技术研究所，2007.

[11] TOUTANOVA K, TOLAG I H, MANNING C D. Extensions to HMM-based statistical word alignment models [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 87-94.

[12] OCH F J. Minimum error rate training in statistical machine translation [C]// Proceedings of the 41st Annual Meeting of the ACL. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003: 160-167.

（编辑　刘杨）