词语对齐的对数线性模型[1]

刘洋 刘群 林守勋
中国科学院计算技术研究所
北京市海淀区科学院南路 6 号 2704 信箱，100080
{yliu, liuqun, sxlin}@ict.ac.cn

摘要

基于对数线性模型，我们为词语对齐提出一种框架。所有的知识源被视作依赖于源语言句子、目标语言句子以及可能的其他变量的特征函数。对数线性模型使统计对齐模型易于扩展，方便加入更多的语言学信息。在本文，我们使用 IBM 模型 3、词性信息和双语词典作为特征。实验表明，对数线性模型显著优于 IBM 翻译模型。

1. 引言

词语对齐的目标在于指明平行文本中词之间的对应关系，最早是作为统计翻译模型的中间产物而被提出(Brown et al., 1993)。由于经过词语对齐的语料是重要的与翻译相关的资源，词语对齐对统计机器翻译而言十分关键。

研究人员提出各种各样的方法在平行文本中计算词语对齐，这些方法大体上可分为两类：统计方法和启发式方法。统计方法往往试图通过建立模型来描述平行文本之间的关系，模型参数可以从训练语料库中学习(Brown et al., 1993; Vogel and Ney, 1996)。启发式方法通过根据语言对设计各种各样的相似度函数来计算词语对齐(Smadja et al., 1996; Ker and Chang, 1997; Melamed, 2000)。统计方法和启发式方法的主要区别在于统计方法是基于概率模型而启发式方法则依赖于相似度函数。研究表明，统计对齐模型要优于简单的 Dice 系数方法(Och and Ney, 2003)。

然而，由于自然语言的多样性，词语对齐问题还远未达到充分解决的地步。比如，习惯表达、随意翻译以及内容词或功能词省略等语言现象给词语对齐带来很大的困难。当两种语言在词语顺序上差异很大时，词语对齐尤为困难。因此，通过整合所有有用的语言学信息来缓解这些问题是很有必要的。

Tiedemann (2003)提出整合关联线索（association clue）的词语对齐方法。线索被定义为关联的概率，线索整合是通过单个线索的分离实现的。线索整合的一个关键假设是线索之间是相互独立的，然而这个假设通常并不能保证为真。Och and Ney (2003)提出模型 6，该模型是 IBM 翻译模型和 HMM 模型的线性整合。虽然模型 6 要比通常的 IBM 模型取得更好的结果，它不能够引入除 IBM 翻译模型和 HMM 模型之外的依赖关系。Cherry and Lin(2003)提出一种易于整合与上下文相关的特征的统计模型。

---

[1] 这篇文章是一篇译稿，由刘洋翻译，可供中文读者参考，正式的论文是由英文写作。译稿相对正式的论文有少量改动，纠正了一些错误。正式的论文参见：
Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear Models for Word Alignment. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), pages 459-466, Ann Arbor, USA, June.

对数线性模型易于合并附加的依赖关系，并且在统计机器翻译中得到成功的应用(Och and Ney, 2002)。在本文，我们为词语对齐提出基于对数线性模型的框架。所有的知识源被视作依赖于源语言句子、目标语言句子以及可能的其他变量的特征函数。对数线性模型使统计对齐模型易于扩展，方便加入更多的相关信息。我们使用 IBM 模型 3、词性信息和双语词典作为特征。实验表明，对数线性模型显著优于 IBM 翻译模型。

我们首先描述词语对齐的对数线性模型，然后讨论特征函数设计，接着介绍对数线性模型的训练方法和搜索算法。最后是实验结果和分析，以及未来的研究方向。

2. 对数线性模型

下面我们将给出词语对齐的正式定义。已知源语言句子 $\mathrm{e}=e_1^I=e_1,...,e_i,...,e_I$ 和目标语言句子 $\mathrm{f}=f_1^J=f_1,...,f_j,...,f_J$，我们定义 $l=(i,j)$ 是一个连线如果 e_i 和 f_j 互为翻译（或者部分翻译）。对齐 a 被定义为词语位置的笛卡尔集的子集：

$$\mathrm{a}\subseteq\{(i,j):i=1,...,I;\ j=1,...,J\} \tag{1}$$

我们定义对齐问题为：已知源语言句子 e 和目标语言句子 f，求使 $\Pr(\mathrm{a}|\mathrm{e},\mathrm{f})$ 取得最大值的对齐 a。

我们直接对概率 $\Pr(\mathrm{a}|\mathrm{e},\mathrm{f})$ 建立模型，而最大熵是非常合适的框架(Berger et al., 1996)。在此框架下，我们可设计一组特征函数 $h_m(\mathrm{a},\mathrm{e},\mathrm{f})$，其中 $m=1,...,M$。对于每个特征函数，存在相应的模型参数 λ_m，其中 $m=1,...,M$。因此：

$$\Pr(\mathrm{a}|\mathrm{e},\mathrm{f})=\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\mathrm{a},\mathrm{e},\mathrm{f})\right]}{\sum_{\mathrm{a}'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\mathrm{a}',\mathrm{e},\mathrm{f})\right]} \tag{2}$$

(Papineni et al., 1997)将这种方法在自然语言理解中使用，并且由(Och and Ney, 2002)成功地应用于统计机器翻译中。

因此，我们获得下面的决策规则(decision rule)：

$$\hat{\mathrm{a}}=\mathrm{argmax}_{\mathrm{a}}\left\{\sum_{m=1}^{M}\lambda_m h_m(\mathrm{a},\mathrm{e},\mathrm{f})\right\} \tag{3}$$

一般而言，源语言句子 e 和目标语言句子 f 是词语对齐的两个基础知识源，而那些可以确定词汇之间关联的语言学知识，往往会被传统的词语对齐方法所忽略。一些语言学工具，如词性标记器、句法分析器、命名实体识别器，已经越来越成熟并且可用于越来越多的自然语言。利用这些语言学信息来提高词语对齐是很有必要的，而对数线性模型非常适合把这些知识以特征函数的形式整合到模型中来。

为了加入一个新的有别于源语言和目标语言句子的依赖关系，我们在公式 2 的基础上添加一个新变量 v：

$$\Pr(\mathrm{a}|\mathrm{e},\mathrm{f},\mathrm{v})=\frac{\exp[\sum_{m=1}^{M}\lambda_m h_m(\mathrm{a},\mathrm{e},\mathrm{f},\mathrm{v})]}{\sum_{\mathrm{a}'}\exp[\sum_{m=1}^{M}\lambda_m h_m(\mathrm{a},\mathrm{e},\mathrm{f},\mathrm{v})]} \tag{4}$$

相应的决策规则为：

$$\hat{\mathrm{a}}=\mathrm{argmax}_{\mathrm{a}}\left\{\sum_{m=1}^{M}\lambda_m h_m(\mathrm{a},\mathrm{e},\mathrm{f},\mathrm{v})\right\} \tag{5}$$

需要注意的是，我们的对数线性模型与(Och and Ney, 2003)提出的模型 6 是不同的，后者将词语对齐问题定义为：已知源语言句子 e，求使 $\Pr(\mathrm{f},\mathrm{a}|\mathrm{e})$ 取得最大值的对齐 a。

3. 特征函数

在本文，我们采用 IBM 翻译模型 3 作为我们线性对数模型的基本特征。此外，我们还会利用到词性标记和双语词典。

3.1 IBM 翻译模型

Brown et al. (1993)为翻译过程建立了一系列统计模型。IBM 翻译模型试图对翻译概率 $\Pr(f_1^J|e_1^I)$ 进行建模，以描述源语言句子 e_1^I 和目标语言句子 f_1^J 之间的关系。在统计对齐模型 $\Pr(f_1^J,a_1^J|e_1^I)$ 中，词语对齐 $\mathrm{a}=a_1^J$ 作为隐变量引入，描述了目标语言词位置 j 到源语言词位置 $i=a_j$ 的映射关系。翻译模型和对齐模型之间的关系可由下面的公式得到：

$$\Pr(f_1^J|e_1^I)=\sum_{a_1^J}\Pr(f_1^J,a_1^J|e_1^I) \tag{6}$$

虽然 IBM 模型被认为在逻辑上比启发式方法更有条理，它们也有两个缺点。第一，IBM 模型限制每个目标语言词 f_j 只能连向一个源语言词 e_{a_j}。更普遍的方法应当是建立一个对齐模型，使得源语言和目标语言词位置之间可以任意连线。第二，IBM 模型是与具体语言无关的，这样就无法处理一些与具体语言相关的语言现象。

在本文，我们使用模型 3 作为我们的基础特征函数[2]：

$$\begin{aligned}&h(\mathrm{a},\mathrm{e},\mathrm{f})=\Pr(f_1^J,a_1^J|e_1^I)\\&=\binom{m-\phi_0}{\phi_0}p_0^{m-2\phi_0}p_1^{\phi_0}\prod_{i=1}^{l}\phi_i!\,n(\phi_i|e_i)\prod_{j=1}^{m}t(f_j|e_{a_j})\,d(j|a_j,l,m)\end{aligned} \tag{7}$$

我们将不同翻译方向的模型 3 区分为不同的特征：将英语作为源语言、法语作为目标语言或者将法语作为源语言、英语作为目标语言。

[2] 如果一个目标语言词连向多个源语言词，则设定 $h(\mathrm{a},\mathrm{e},\mathrm{f})=0$

3.2 词性标记转换模型

除了源语言和目标语言句子，我们采用的第一个语言学信息是词性标记。(Toutanova et al., 2002) 使用词性标记来提高基于 HMM 的模型的对齐质量。他们为两种语言引入词语标记的附加词汇概率。

在 IBM 模型和 HMM 模型中，如果想要容纳新的信息，必须设计一个扩充的模型使之能够利用前面的模型参数。而对数线性模型却可以很容易地容纳新信息。

我们使用词性标记转换模型作为特征函数。这个特征从外部数据(held-out data)通过简单计数学习词性标记转换概率，然后将学习到的概率分布应用到评价词语对齐中。概率估计方法如下：

$$p\left(fT \mid eT\right)=\frac{N_A\left(fT,eT\right)}{N\left(eT\right)}$$

其中，$N_A\left(fT,eT\right)$ 是指词性标记 fT 连向词性标记 eT 的次数，$N\left(eT\right)$ 是词性标记 eT 出现的次数。

我们定义 $\mathrm{eT}=eT_1^I=eT_1,...,eT_i,...eT_I$ 和 $\mathrm{fT}=fT_1^J=fT_1,...,fT_j,...fT_J$ 分别是句对 e 和 f 的词性标记序列，则词性标记转换模型定义如下：

$$\Pr\left(\mathrm{fT}|\mathrm{a},\mathrm{eT}\right)=\prod_{l\in\mathrm{a}}t\left(fT_{l(j)}\mid eT_{l(i)}\right) \tag{8}$$

其中，l 是 a 中的一个元素，换言之，l 是一条连线。$l\left(i\right)$ 是 l 中的源语言词位置，$l\left(j\right)$ 是 l 中目标语言词位置。

因此，特征函数可设计为：

$$h\left(\mathrm{a},\mathrm{e},\mathrm{f},\mathrm{eT},\mathrm{fT}\right)=\prod_{l\in\mathrm{a}}t\left(fT_{l(j)}\mid eT_{l(i)}\right) \tag{9}$$

我们将不同翻译方向的词性标记转换模型区分为不同的特征：将英语作为源语言、法语作为目标语言或者将法语作为源语言、英语作为目标语言。

3.3 双语词典

双语词典也可以作为附加的知识源。给定词语对齐，我们可以统计双语词典中有多少个词条在对齐中共现。因此，双语词典的权重就可以获得。我们采用双语词典作为特征的原因在于双语词典应该比自动获得的词典更可靠，同时也应当获得较大的权重。

我们定义双语词典是一组词条：$\mathrm{D}=\left\{\left(e,f,conf\right)\right\}$。其中，$e$ 是源语言词，f 是目标语言词，$conf$ 是一个正实数（通常为 1.0）。$conf$ 是由词典编纂者设定，用来表示该词条有效性的程度。因此，使用双语词典的特征为：

$$h\left(\mathrm{a},\mathrm{e},\mathrm{f},\mathrm{D}\right)=\sum_{l\in\mathrm{a}}occur\left(e_{l(i)},f_{l(j)},D\right) \tag{10}$$

其中，

$$occur(e, f, D) = \begin{cases} conf & \text{if (e,f) occurs in D} \\ 0 & \text{else} \end{cases} \tag{11}$$

4. 训练

根据公式 4，我们使用GIS(Generalized Iterative Scaling)算法(Darroch and Ratcliff, 1972)来训练对数线性模型的模型参数 λ_1^M 。经过适当的转换，GIS算法可以用来处理实数值特征。我们采用由Franz J. Och开发的YASMET[3]来执行训练。

公式 4 中的重正化(renormalization)需要大量的、可能产生的对齐集合。如果源语言句子 e 包含 l 个词，目标语言句子 f 包含 m 个词，那么总共能够产生的词语对齐的数目是 2^{lm} (Brown et al., 1993)。当 lm 非常大时，枚举所有可能的词语对齐是不现实的。因此，我们用较大数量的高概率对齐集合来逼近所有可能的对齐集合，这样的对齐集合也称之为对齐的 n-best 列表。

我们在开发集上训练模型参数。开发集包含数百个人工对齐的双语句对。使用 n-best 列表逼近可能会导致使用 GIS 算法训练的参数在测试集上产生质量较差的对齐，甚至是在开发集上也质量较差。这是因为在训练过程中模型参数变化很大并且可能会包含训练中没有考虑到的对齐。为了避免这个问题，我们依照 Och(2002)的方法迭代训练模型参数，每次迭代都合并 n-best 列表，直至 n-best 列表不再变化为止。然而，这种训练方法是基于极大似然准则(maximum likelihood criterion)的，与最终未知双语文本的对齐质量关联很小。因此，当迭代结束时，我们有一系列模型参数，我们选择在开发集上产生最好对齐的模型参数。

5. 搜索

我们采用贪心算法从所有可能的对齐空间中搜索概率最高的对齐。空间中的一个状态是是一个部分对齐。在当前状态下增加一条连线被称之为迁移(transition)。开始状态是空对齐，源语言和目标语言的所有词都连向空。终止状态是添加任何连线都无法使概率进一步增长的状态。搜索的过程就是从开始状态开始，不断地添加连线，直至概率不再增长为止。

我们通过计算增益而不是概率来提高效率。增益是一个启发式函数，定义如下：

$$gain(\mathrm{a}, l) = \frac{\exp\left[\sum_{m=1}^{M} \lambda_m h_m(\mathrm{a} \cup l, \mathrm{e}, \mathrm{f})\right]}{\exp\left[\sum_{m=1}^{M} \lambda_m h_m(\mathrm{a}, \mathrm{e}, \mathrm{f})\right]} \tag{12}$$

其中，$l = (i, j)$ 是添加到 a 的连线。

对于一般的对数线性模型而言，贪心搜索算法如下：

[3] 可在http://www.fjoch.com/YASMET.html下载

输入：e，f，eT，fT 和 D
1. $\mathrm{a}=\phi$ 2. 对每个不属于 a 的连线 $l=(i,j)$ 计算增益 $gain(\mathrm{a},l)$ 3. 如果对于任意的连线 l $gain(\mathrm{a},l)$ 均不大于 1，则算法终止 4. 向 a 中添加 $gain(\mathrm{a},l)$ 最大的连线 $\hat{l}$ 5. 转到 2
输出：a

上面的搜索算法对于我们所采用的对数线性模型（以IBM模型 3、词性标记转换模型和双语词典作为特征）而言效率并不高。当添加新的连线时，为每个特征计算特征值非常耗时间，特别是当句子十分长的时候。因此，针对我们所采用的对数线性模型，下面的增益计算方法[4]会使得搜索效率更高：

$$gain(\mathrm{a},l)=\sum_{m=1}^{M}\lambda_m \log\left(\frac{h_m(\mathrm{a}\cup l,\mathrm{e},\mathrm{f})}{h_m(\mathrm{a},\mathrm{e},\mathrm{f})}\right) \tag{13}$$

需要注意的是，我们限制所有特征函数的值均不小于 0。

前面所描述的贪心搜索算法的终止条件是：

$$gain(\mathrm{a},l)=\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\mathrm{a}\cup l,\mathrm{e},\mathrm{f})\right]}{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\mathrm{a},\mathrm{e},\mathrm{f})\right]}\le 1.0$$

即：

$$\sum_{m=1}^{M}\lambda_m\left[h_m(\mathrm{a}\cup l,\mathrm{e},\mathrm{f})-h_m(\mathrm{a},\mathrm{e},\mathrm{f})\right]\le 0.0$$

我们引入特征增益 t，从而获得新的终止条件：

$$gain(\mathrm{a},l)=\sum_{m=1}^{M}\lambda_m \log\left(\frac{h_m(\mathrm{a}\cup l,\mathrm{e},\mathrm{f})}{h_m(\mathrm{a},\mathrm{e},\mathrm{f})}\right)\le t$$

其中，

$$t=\sum_{m=1}^{M}\lambda_m\left\{\log\left(\frac{h_m(\mathrm{a}\cup l,\mathrm{e},\mathrm{f})}{h_m(\mathrm{a},\mathrm{e},\mathrm{f})}\right)-\left[h_m(\mathrm{a}\cup l,\mathrm{e},\mathrm{f})-h_m(\mathrm{a},\mathrm{e},\mathrm{f})\right]\right\}$$

需要注意的是，我们仍然限制所有特征函数的值均不小于 0。特征增益 t 是一个实数，可在开发集上优化。

因此，针对本文所采用的对数线性模型，搜索算法如下：

[4] 我们仍将新的启发式函数称之为gain来避免引入更多的符号，虽然公式 13 中的gain与公式 12 中的并不等价。

输入：e，f，eT，fT，D 和 t
1. $\mathrm{a}=\phi$ 2. 对每个不属于 a 的连线 $l=(i,j)$ 计算增益 $gain(\mathrm{a},l)$ 3. 如果对于任意的连线 l $gain(\mathrm{a},l)$ 均不大于 t，则算法终止 4. 向 a 中添加 $gain(\mathrm{a},l)$ 最大的连线 $\hat{l}$ 5. 转到 2
输出：a

特征增益 t 依赖于添加的连线 l。在搜索过程中，我们没有考虑这种依赖关系，而是把 t 设定为一个固定的实数。

6. 实验结果

在本节，我们将给出在汉英平行语料库上的实验结果。在实验中，我们使用了训练集、双语词典、开发集和测试集。表 1 给出了它们的一些统计数据。

		英语	汉语
训练集	句子数	108 925	
	词语数	3 784 106	3 862 637
	词汇量	49 962	55 698
双语词典	词条数	415 753	
	词汇量	206 616	203 497
开发集	句子数	435	
	词语数	11 462	14 252
	词汇量	26.35	32.76
测试集	句子数	500	
	词语数	13 891	15 291
	词汇量	27.78	30.58

表 1：训练集、双语词典、开发集和测试集的统计数据

开发集和测试集中的汉语句子采用 ICTCLAS(Zhang et al., 2003)进行分词和标注。我们自己开发了一个简单的 tokenizer 处理英语句子，然后用一个由 Eric Brill 开发的基于规则的标记器(Brill, 1995)做词性标记。我们对 935 个句对进行人工对齐，从中挑选 500 句作为测试集，其余 435 句作为开发集，用来优化模型参数和增益阈值。

给定人工标注的词语对齐，我们采用准确率 precision、召回率 recall 和对齐错误率 AER(Och and Ney, 2003)作为评价标准：

$$precision = \frac{|A \cap P|}{|A|}$$

$$recall = \frac{|A \cap S|}{|S|}$$

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

其中，A 是词语对齐系统输出的连线集合，S 是人工标注者标为“确定”的连线集合，P 是人工标注者标为“可能”的连线集合，S 是 P 的子集。在实验中，我们只采用了一种标记类型，因此 $S = P$ 。

我们使用 GIZA++(Och and Ney, 2003)训练 IBM 翻译模型。训练方案是 $1^5 H^5 3^5$，即模型 1 训练 5 次，HMM 模型训练 5 次，模型 3 训练 5 次。除了改变模型的迭代次数，我们使用 GIZA++的默认配置。之后，我们使用了三种 IBM 模型的平衡化方法：intersection，union 和 refined method(Och and Ney, 2003)。

给定 GIZA++的输出参数，我们将其用于对数线性模型的基本特征 IBM 模型 3。换言之，除了词性标记转换概率表和双语词典，我们的对数线性模型和 GIZA++使用完全相同的参数。

表 2 给出了我们的对数线性模型和 IBM 模型 3 的结果。其中，第 3 至 7 行是 IBM 模型 3 的结果，第 8 至 12 行是对数线性模型的结果。第 9 行“+Model 3 C->E”的意思是对数线性模型采用两个特征：Model 3 E->C 和 Model 3 C->E，依此类推。

		训练语料库规模				
		1K	5K	9K	39K	109K
IBM 模型	Model 3 E->C	0.4497	0.4081	0.4009	0.3791	0.3745
	Model 3 C->E	0.4688	0.4261	0.4221	0.3856	0.3469
	Intersection	0.4588	0.4106	0.4044	0.3823	0.3687
	Union	0.4596	0.4210	0.4157	0.3824	0.3703
	Refined Method	0.4154	0.3586	0.3499	0.3153	0.3068
对数线性模型	Model 3 E->C	0.4490	0.3987	0.3834	0.3639	0.3533
	+ Model 3 C->E	0.3970	0.3317	0.3217	0.2949	0.2850
	+POS E->C	0.3828	0.3182	0.3082	0.2838	0.2739
	+POS C->E	0.3795	0.3160	0.3032	0.2821	0.2926
	+Dict	0.3650	0.3092	0.2982	0.2738	0.2685

表 2：IBM 模型 3 和对数线性模型的 AER 值比较

从表 2 可以看出，我们的对数线性模型在所有的训练语料库规模上都比 IBM 模型取得更低 AER 值(对齐错误率越低表示对齐质量越高)。单独考虑 Model 3 E->C，即以英语为源语言、汉语为目标语言的模型 3，第五节所描述的贪心算法比 GIZA++所采用的爬山算法(hillclimbing algorithm）取得更好的结果。

表 3 给出了我们的对数线性模型和 IBM 模型 5 的结果。训练方案是 $1^5 H^5 3^5 4^5 5^5$ 。对数线性模型同样使用 GIZA++的输出参数。

		训练语料库规模				
		1K	5K	9K	39K	109K
IBM 模型	Model 5 E->C	0.4384	0.3934	0.3853	0.3573	0.3429
	Model 5 C->E	0.4564	0.4067	0.3900	0.3423	0.3239
	Intersection	0.4432	0.3916	0.3798	0.3466	0.3267
	Union	0.4499	0.4051	0.3923	0.3516	0.3375
	Refined Method	0.4106	0.3446	0.3262	0.2878	0.2748
对数线性模型	Model 3 E->C	0.4372	0.3873	0.3724	0.3456	0.3334
	+ Model 3 C->E	0.3920	0.3269	0.3167	0.2842	0.2727
	+POS E->C	0.3807	0.3122	0.3039	0.2732	0.2667
	+POS C->E	0.3731	0.3091	0.3017	0.2722	0.2657
	+Dict	0.3612	0.3046	0.2943	0.2658	0.2625

表 3：IBM 模型 5 和对数线性模型的 AER 值比较

对比表 2 和表 3，我们发现对数线性模型使用训练方案$1^5H^53^54^55^5$的输出参数的对齐质量要略高于使用训练方案$1^5H^53^5$，这归功于附加的模型 4 和模型 5 的训练。

对数线性模型采用了词性标记信息和双语词典，而 IBM 模型没有采用。然而，如果把对数线性组合(Model 3 E->C + Model 3 C->E)视作一种平衡化的方法，它依然比 intersection、union 和 refined method 要好。

图 1 给出了在模型参数固定的情况下增益阈值对准确率、召回率和对齐错误率的影响

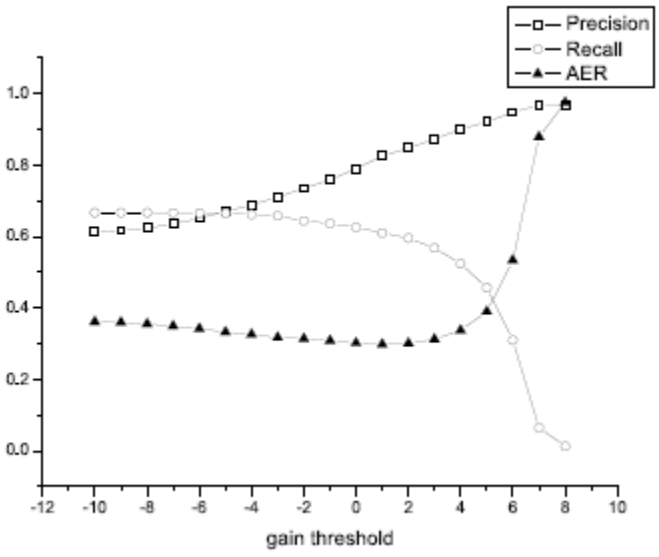

图 1：模型参数固定的情况下增益阈值对准确率、召回率和对齐错误率的影响

图 2 给出了特征数量和训练语料库规模对于搜索效率的影响。可以看出，特征越多，训练语料库规模越大，搜索时间越长。

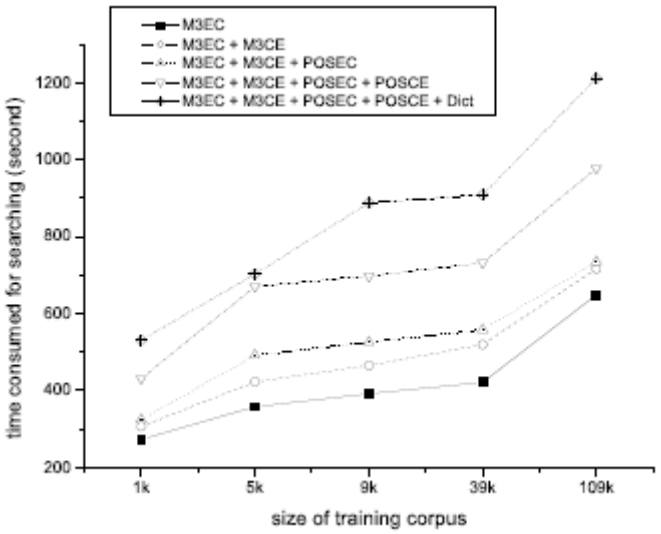

图 2：特征数量和训练语料库规模对于搜索效率的影响

表 3 给出了我们在开发集上训练得到的模型参数。我们注意到加入新的特征会影响到其它特征的模型参数。

	MEC	+MCE	+PEC	+PCE	+Dict
λ_1	1.000	0.466	0.291	0.202	0.151
λ_2	-	0.534	0.312	0.212	0.167
λ_3	-	-	0.397	0.270	0.257
λ_4	-	-	-	0.316	0.306
λ_5	-	-	-	-	0.119

表 3：模型参数。λ_1：Model 3 E->C (MEC)；λ_2：Model 3 C->E (MCE)；λ_3：POS E->C (PEC)；λ_4：POS C->E (PCE)；λ_5：Dict。模型参数被正规化使得 $\sum_{m=1}^{5}\lambda_m = 1$

7. 结论

我们为平行语料库之间的词语对齐提出基于对数线性模型的框架。该框架使得统计对齐模型易于加入新的语言学信息。我们以 IBM 模型 3 作为基础特征，同时采用了词性标记和双语词典作为特征。实验结果表明，对数线性模型要优于 IBM 翻译模型。但是，需要强调

的是，我们的方法是有监督的(supervised)，依赖于人工对齐的开发集来调参数，而 IBM 模型的训练方法是无监督的(unsupervised)。

目前，我们只采用了三种知识源作为特征函数。基于句法的翻译模型，如树到串的模型(Yamada and Knight, 2001)和树到树的模型(Gildea, 2003)，可能非常适合加入到对数线性模型中来。将在统计机器翻译中得到成功应用的最小错误率训练(Och, 2003)用来直接优化 AER 也是很有意义的。

致谢

本文的工作得到 863 计划项目“中文平台评价体系研究与基础数据库建设”（编号：2004AA114010）的支持。

参考文献

Adam L. Berger, Stephen A. Della Pietra, and Vincent J.DellaPietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39-72, March.

Eric Brill. 1995. Transformation-based-error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistic*, 21(4), December.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263-311.

Colin Cherry and Dekang Lin. 2003. A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.

J. N. Darroch and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470-1480.

Daniel Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.

Sue J. Ker and Jason S. Chang. 1997. A class-based approach to word alignment. *Computational Linguistics*, 23(2):313-343, June.

I. Dan Melamed 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221-249, June.

Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295-302, Philadelphia, PA, July.

Franz J. Och. 2002. Statistical Machine Translation: From Single-Word Models to Alignment Templates. *Ph.D. thesis*, Computer Science Department, RWTH Aachen, Germany, October.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages: 160-167, Sapporo, Japan.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19-51, March.

Kishore A. Papineni, Salim Roukos, and Todd Ward. 1997. Feature-based language understanding. In *European Conf. on Speech Communication and Technology*, pages 1435-1438, Rhodes, Greece, September.

Frank Smadja, Vasileios Hatzivassiloglou, and Kathleen R. McKeown. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1-38, March.

Jörg Tiedemann. 2003. Combining clues for word alignment. In *Proceedings of the 10th Conference of European Chapter of the ACL (EACL)*, Budapest, Hungary, April.

Kristina Toutanova, H. Tolga Ilhan, and Christopher D. Manning. 2003. Extensions to HMM-based statistical word alignment models. In *Proceedings of Empirical Methods in Natural Language Processing*, Philadelphia, PA.

Stephan Vogel, Hermann Ney, and Christoph Tillmann.1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Int. Conf. on Computational Linguistics*, pages 836-841, Copenhagen, Denmark, August.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical machine translation model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages: 523-530, Toulouse, France, July.

Huaping Zhang, Hongkui Yu, Deyi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. *In Proceedings of the second SigHan Workshop affiliated with 41th ACL*, pages: 184-187, Sapporo, Japan.