基于锚词对的英汉双语语段对齐模型*

吴尉林 屈刚 陆汝占

上海交通大学计算机科学与工程系 上海 200030

E-mail: wu-wl@cs.sjtu.edu.cn

摘 要：对双语语料库进行语段级对齐是基于实例的机器翻译（EBMT）的基础。本文提出了基于锚词对的英汉双语语段对齐模型并给出了相应的对齐算法，解决了中、小规模语料库的数据稀疏问题。系统把语段切分的歧义推迟到语段对齐时排除，提高了语段切分的正确率。试验表明模型具有较高有效性。

关键词：语段对齐，机器翻译

Anchor-based English-Chinese Bilingual Chunk Alignment Model

Wu Weilin, QuGang, Lu RuZhan

Department of Computer Science and Engineering, Shanghai JiaoTong University, Shanghai 200030
E-mail: wu-wl@cs.sjtu.edu.cn

Abstract: Bilingual chunk alignment on bilingual corpus is the base of Example-based machine translation. This paper presents an anchors-based English-Chinese bilingual alignment model as well as the corresponding algorithm of alignment, addressing the sparse data problem in medium or small bilingual corpora. In this system, the disambiguation of chunk segmentation is postponed to the alignment process to increase the correctness of chunk segmentation. Experimental results demonstrate the feasibility of this model.

Key words: chunk alignment, machine translation

1. 引言

近年来，随着语料库语言学的发展，基于实例的机器翻译[1,2]（Example-based MT）方法成为机器翻译的新思路之一。EBMT 系统事先存储大量语段级对齐的双语句子对，即双语语料库。翻译时，系统仅对被翻译句进行浅层分析，把它切分成语段，然后根据上下文从双语语料库中找出各语段的最佳翻译，再把它们按一定的顺序排列起来，最后生成译句。对双语语料库进行语段对齐是基于实例的机器翻译的需求，同时，对齐的双语语段本身也可以作为一种翻译知识独立使用。

双语语段对齐的方法主要有两类：基于评分的方法[3,4]和基于翻译模型的方法[5,6,7]。前者使用评分函数来计算对齐分数，评分函数的参数可以人工设定[3]，这种参数不能根

*本文承上海市科委重点项目多语种自然语言信息处理应用系统（项目号 025115038）的资助

据真实语料进行训练；也有把双语句子对的相关信息作为评分函数的参数，如在目标句没有对译词汇的源句中的词数等[4]，这种评分参数可以自动计算得到。1993 年 IBM 公司的 Brown[5]等人提出了基于词的（word-based）统计机器翻译模型，提出了隐藏对齐（Hidden-alignment）的思想，认为翻译的过程中隐含了对齐的过程。Ye-Yi Wang 提出了基于结构的（Structure-based）统计翻译模型[6]，把语段看成是翻译的基本单位，并给出了基于统计的解码（decoding）翻译算法[7]。

2. 理论

2.1 基于语段的噪声信道模型的随机过程

设待翻译的英语句为 $\bar{e} = e_1 e_2 ... e_l$，翻译生成的汉语句为 $\bar{c} = c_1 c_2 ... c_m$。基于语段的噪声信道模型由汉语句 $\bar{c}$ 输出英语句 $\bar{e}$ 随机过程如下：

(1) 把汉语句切分成语段，并称切分的结果为“汉语语段串”。

设切分后汉语语段串长度为 n，则可表示为 $\bar{C} = C_1 ... C_n$，其中 $C_i = C_{i1}, C_{i2}, ... C_{i,l_i}$ $(1 \le i \le n)$ 是汉语语段，而 C_{ij} 是 $\bar{c}$ 中的某个词。并设汉语句子 $\bar{c}$ 切分为语段串 $\bar{C}$ 的概率为 $\Pr(\bar{C} \mid \bar{c})$。

(2) 根据汉语句 $\bar{c}$ 和语段串 $\bar{C}$，选择英语语段串 $\bar{E}$ 的长度（即它所含语段数）。

设英语语段串长度为 q 的概率是 $\Pr(q \mid \bar{c}, \bar{C})$，则可表示为 $\bar{E} = E_1 E_2 ... E_q$，其中 $E_i = E_{i1}, E_{i2}, ... E_{iq_i}$ $(1 \le i \le q)$ 是英语语段，而 E_{ij} 是 $\bar{e}$ 中的某个单词。

(3) 对每个英语语段，选择和它对齐的汉语语段的位置。

设第 j 个英语语段和第 a_j 个汉语语段对齐。本文中，记汉语语段串 $\bar{C}$ 的子串 $C_1 C_2 ... C_k$ 为 C_1^k，记英语语段串 $\bar{E}$ 的子串 $E_1 E_2 ... E_k$ 为 E_1^k，记 $a_1 a_2 ... a_k$ 为 a_1^k。并设第 j 个英语语段和第 a_j 个汉语语段对齐的概率为 $\Pr(a_j \mid j, a_1^{j-1}, \bar{C}, \bar{c})$。

(4) 对每个汉语语段 C，选择一个翻译 E。设其概率为 $t(E \mid C)$，其中 t 对于每个 C 满足归一条件：$\sum_E t(E \mid C) = 1$。

(5) 把英语语段串 $\bar{E}$ 合并为英语句 $\bar{e}$，设其概率为 $\Pr(\bar{e} \mid \bar{E})$，则有：

$$\Pr(\bar{e} \mid \bar{E}) = \begin{cases} 1 & \text{当} \bar{E} \text{和} \bar{e} \text{表示相同的句子时} \\ 0 & \text{其他} \end{cases}$$

2.2 基于锚词对的参数估计

上面的随机过程基于一个隐藏语段、隐藏对齐（hidden-Alignment）的模型。翻译和对齐的关系包括：1. 翻译隐含对齐；2. 最佳翻译对应一个最佳对齐。噪声信道由输入 $\bar{c}$ 输出 $\bar{e}$ 的概率等于它在所有可能的语段切分和所有可能的对齐的情况下输出 $\bar{e}$ 的概率之

和，即：$\Pr(\bar{e}\mid\bar{c})=\sum_{\bar{E}}\sum_{\bar{a}}\sum_{\bar{C}}\Pr(\bar{e},\bar{E},\bar{C},\bar{a}\mid\bar{c})$。

假设：$\bar{E}$的产生只与$\bar{C}$、$\bar{a}$有关，而与$\bar{c}$无关；$\bar{e}$的产生只与$\bar{E}$有关，而与其他参数无关。则：$\Pr(\bar{e},\bar{E},\bar{C},\bar{a}\mid\bar{c})=\Pr(\bar{e}\mid\bar{E})\Pr(\bar{E},\bar{a}\mid\bar{C})\Pr(\bar{C}\mid\bar{c})$。

由此可得：$\Pr(\bar{e}\mid\bar{c})=\sum_{\bar{E}}\sum_{\bar{C}}\{\Pr(\bar{e}\mid\bar{E})\Pr(\bar{C}\mid\bar{c})[\sum_{\bar{a}}\Pr(\bar{E},\bar{a}\mid\bar{C})]\}$

$$\Pr(\bar{E}\mid\bar{C})=\sum_{\bar{a}}\Pr(\bar{E},\bar{a}\mid\bar{C})。$$

以下，$\sum_{\bar{E}}\sum_{\bar{C}}$ 都是对所有合法的语段切分路径而言，则有：$\Pr(\bar{e}\mid\bar{E})=1$和

$$\Pr(\bar{E},\bar{a}\mid\bar{C})=\Pr(q\mid\bar{C})\prod_{j=1}^{q}[\Pr(a_j\mid a_1^{j-1},E_1^{j-1},q,\bar{C})\times\Pr(E_j\mid a_1^{j-1},E_1^{j-1},q,\bar{C})]。$$

假设：(1) $\Pr(q\mid\bar{C})\equiv\varepsilon$；(2) $\Pr(E_j\mid a_1^{j-1},E_1^{j-1},q,\bar{C})\equiv t(E_j\mid C_{a_j})$。并记$\Pr(a_j\mid a_1^{j-1},E_1^{j-1},q,\bar{C})\equiv A_j^{a_j}$。$A_j^{a_j}$和$a_j$一样，与句子对$<\bar{c},\bar{e}>$相关，于是有：

$$\Pr(\bar{E},\bar{a}\mid\bar{C})=\varepsilon\prod_{j=1}^{q}t(E_j\mid C_{a_j})A_j^{a_j}$$

同[5]，为了求满足约束$\sum_{E}t(E\mid C)=1$的$\Pr(\bar{e}\mid\bar{c})$的最大值，建立如下辅助函数：

$$h(t,\lambda)\equiv\sum_{\bar{E}}\sum_{\bar{C}}\{P(\bar{C}\mid\bar{c})[\varepsilon\sum_{\bar{a}}\prod_{j=1}^{q}t(E_j\mid C_{a_j})A_j^{a_j}]\}+\sum_{E}\lambda_E(\sum_{C}t(E\mid C)-1)$$

它取得极值应满足的条件为：$\frac{\partial h}{\partial t(E\mid C)}=0$，可求得：

$$t(E\mid C)=\lambda_E^{-1}\sum_{\bar{E}}\sum_{\bar{C}}\{\Pr(\bar{C}\mid\bar{c})\times\sum_{\bar{a}}\Pr(\bar{E},\bar{a}\mid\bar{C})\sum_{j=1}^{q}\delta(E,E_j)\delta(C,C_{a_j})\}$$

其中δ是Kronecker delta函数。

定义语段E和C在翻译$(\bar{e}\mid\bar{c})$中的有效共现次数为：

$$c(E\mid C;\bar{e},\bar{c})=\sum_{\bar{E}}\sum_{\bar{C}}\{\Pr(\bar{C}\mid\bar{c})\Pr(\bar{a}\mid\bar{E},\bar{C})\times\sum_{j=1}^{q}\delta(E,E_j)\delta(C,C_{a_j})\}$$

由$\Pr(\bar{a}\mid\bar{E},\bar{C})=\Pr(\bar{E},\bar{a}\mid\bar{C})/\Pr(\bar{E}\mid\bar{C})$和$\Pr(\bar{E}\mid\bar{C})$可表示为$\prod_{j=1}^{q}\sum_{i=0}^{n}[t(E_j\mid C_{a_j})A_j^{a_j}]$，

可得：$t(E\mid C)=\lambda_E'^{-1}C(E\mid C;\bar{e},\bar{c})$ （1）其中 $\lambda_E'=\lambda_E/\Pr(\bar{E}\mid\bar{C})$

$$c(E\mid C;\bar{e},\bar{c})=\sum_{\bar{E}}\sum_{\bar{C}}\{\Pr(\bar{C}\mid\bar{c})\times\sum_{j=1}^{q}\sum_{i=0}^{n}\frac{t(E\mid C)\delta(E,E_j)\delta(C,C_i)A_j^i}{t(E\mid C_0)A_0^i+\cdots+t(E\mid C_n)A_n^i}\}\quad(2)$$

为了估计A_j^i，在模型中引入了锚词(anchor)对。目前，锚词对尚无统一的定义，它总是跟具体的应用相关，它的精确定义因系统而异。通常，锚词对是指一对有较高对译可能的源语言、目标语言词对。在本系统中，从实用的角度出发，词对$<w_E,w_C>$定

义为锚词对当且仅当它们不是冠词或介词、在双语词典中出现而且在对译句中只有一词与之对译。可以借助对齐好的锚词对来确定同一句子中其他词（或语段）之间的对齐。本文中令第 j 个英语语段和第 k 个汉语语段对齐的概率为：

$$A_j^k = \Pr(a_j = k \mid a_1^{j-1}, E_1^{j-1}, q, \vec{C}) = \frac{1}{\lambda} e^{-\frac{(k-\mu)^2}{2dist^2}}$$

$$\mu = y_0 + (center(j) - x_0)(y_1 - y_0)/(x_1 - x_0)$$

$$dist = \min(|center(j) - x_0|, |center(j) - x_1|)$$

$$center(j) = (BeginPos(j) + EndPos(j))/2$$

其中：$< x_0, y_0 >$ 是语段 j 左侧离语段 j 最近（根据 $center(j) - x_0$）的锚词对，而 $< x_1, y_1 >$ 是语段 j 右侧离语段 j 最近的锚词对；$BeginPos(j)$ 是指语段 j 中第一个词的位置，而 $EndPos(j)$ 则指语段 j 中最后一个词的位置。如图 1：

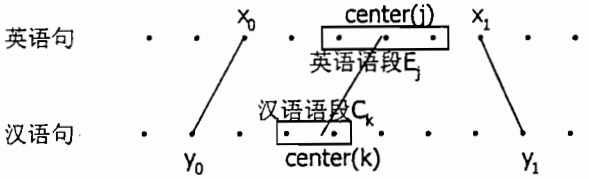

图 1 基于锚词对的语段对齐概率分布

需要说明的是，根据我们的经验，双语句子中锚词对出现的比例比较高的。即使双语句子中没有出现锚词对，上述的公式仍然适用，即退化为 $x_0 = 0, y_0 = l$ 和 $x_1 = 0, y_1 = m$ 的情况，直观上看，即句子对中双语句子的开头和结尾（句号）总是对齐的。在这种情况下，相当于所有的对齐都有相同的概率。还有，由于在本系统中锚词对的定义的简单性，锚词对齐的错误率非常低。

2.3 词性语段

一般地，语言中语段的数目是很大的，远远大于语言中词汇的数目。相对于词汇，语段有更低的使用频率，如果直接按（1）式进行训练会因为数据稀疏而导致训练数据不准确。为克服这个问题，本文用构成语段的单词的词性串之间的翻译概率估计语段之间的翻译概率，即对语段 $E = e_0 e_1 ... e_m$ 和 $C = c_0 c_1 ... c_n$，设语段中各个单词的词性标记构成的串分别为 $T^E = t_0^E t_1^E ... t_m^E$ 和 $T^C = t_0^C t_1^C ... t_n^C$，并称它们为“词性语段”。本文认为 $\Pr(E \mid C) \propto \Pr(T^E \mid T^C)$，这样假设是因为：

1. 可以克服数据稀疏问题。由于词性语段的数目少，出现频率高，因此，不会受到数据稀疏问题的干扰，而且训练数据容易收敛。

2. 尽管用 $\Pr(T^E \mid T^C)$ 来估计 $\Pr(E \mid C)$ 有不精确性，但在预先进行锚词对齐使得占句中总词数 64%的词已对齐的情况下，$dist$ 值一般较小，即有较高对齐概率的候选汉语语段数少，用上面的估计可以较为准确地从少数候选汉语语段中选出正确的候选（对齐时）或为正确的候选分配较大的概率（训练时）。

3. 在英汉双语对译中，名词-名词语段在大多情况下都被翻译为名词-名词语段等现象，为我们的假设提供了直观正面实例的支持。

记语段 E,C 对应的词性串的对译概率为 $\Pr(T^E \mid T^C)$ 或 $t(t^E \mid t^C)$，则有：

$$t(t^E \mid t^C) = \lambda_E^{-1} C(t^E \mid t^C; \bar{e}, \bar{c}) \quad (1')$$

$$c(t^E \mid t^C; \bar{e}, \bar{c}) = \sum_{\bar{E}} \sum_{\bar{C}} \{\Pr(\bar{C} \mid \bar{c}) \times \sum_{j=1}^{q} \sum_{i=0}^{n} \frac{t(t^E \mid t^c)\delta(E, E_j)\delta(C, C_i)A_j^i}{t(t^E \mid t^C)A_0^i + \cdots + t(t^E \mid t^C)A_n^i}\} \quad (2')$$

2.4 参数估计步骤

参数估计的目标是求 $t(t^E \mid t^C)$ 以及与 $\Pr(\bar{C} \mid \bar{c})$ 有关的概率参数。参数估计的步骤如下：

（1）词性语段串的获取

在实现时，本文设定英语和汉语的语段词性串为有限集。词性语段串的获取步骤如下：1. 对语料库进行切分、词性标注。这样，每个句子对应一个词性标注串。2. 选取语料库中出现超过一定次数（200 次）且长度不小于 2 的词性串，加入到语段词性串集。3. 把词性标注集的每个词性标记加入到语段词性串集。

（2）对双语语料库中的句子按词性语段进行“全切分”。这里，“切分”的对象是句子的词性标注串，“切分”使用的词典是语段标注串集。

（3）按 Bi-Gram 统计

假设汉语词性语段转移概率为 $\Pr(C_{i+1} \mid C_i)$ 和句子以语段 C_1 开头的概率为 $\pi(C_1)$，则 $\Pr(\bar{C} \mid \bar{c}) = \pi(C_1)\prod_{i=2}^{n} \Pr(C_i \mid C_{i-1})$。于是：$\Pr(\bar{a} \mid \bar{E}, \bar{C}) = \Pr(\bar{C} \mid \bar{c})\prod_{j=2}^{q} A_j^{a_j}$ （3）

（4）对语料库中的双语句子进行锚词对齐。

（5）用 EM 算法，进行迭代，得到 $t(t^E \mid t^C)$。

3. 实验结果

系统选用的双语语料库是汽车方面的专业语料库。双语语料库是按句子级对齐的。我们取出所有英语句子包含 4 到 20 个单词的双语句子对，构成双语语料库。该语料库共有 26,825 个句子对，396,866 个汉字，241,485 个英语单词。英语句平均长度 9.1266 个单词，汉语句子平均长度 14.4014 个汉字。实验所采用的汉语词典（包括专业术语）有 75,352 个词，英语单词词典（包括专业术语）有 54,979 个单词，英语短语词典有 6,597 条短语，英汉双语词典（包括专业术语）共有 205,884 个词条。

对双语语料库进行汉语分词、英/汉语词性标注后，通过串频统计得到出现次数超过 200 次的词性标注串。其中，英语词性标注串有 235 个，汉语词性标注串有 245 个。提取所得到的词性标注串包含常见的结构，如形名结构和名名结构等，符合人的直观。出现频率较高的词性语段串如表 1。

使用双语词典进行锚词对齐后，平均每个句子对有 5.8656 个锚词对，即 64.27%的英语词是锚词对齐的。句子中出现这么高比例的锚词对使得可以有效地估计语段对齐概率。

通过训练，得到语段词性串之间的转移概率以及翻译概率，分别给出转移和翻译概率较高的词性语段，如表 2 和表 3。然后，进行语段对齐，语段对齐的正确率为

75.315%。

英语词性语段	出现次数	汉语词性语段串	出现次数
N	70,519	N	77,863
Vi	19,610	Vi	12,595
A n	7,554	a n	2,676
A n n	1,763	a n n	501
P n	6,875	p n	4,272
N n	17,315	n n	20,968
N n n	3,647	n n n	5,445
art n n	5,496	vt a n	786

表 1 词性语段的出现次数

C_{i-1}	C_i	转移概率
Vt	N	0.3914432341
Vt	n n	0.1102709546
P	a n n	0.0044784173
Vt n	p n	0.0035747672

表 2 部分词性语段转移概率

汉语词性语段	英语词性语段	对齐概率
n	N	0.1134335293
N n	n n	0.1225525460
vt n	Vt	0.1692302336
vt n	Vi	0.2159959771
vt n	vt n	0.2714522281

表 3 部分英汉词性语段翻译概率

4. 结束语

本文提出了一个基于锚词对的英汉双语语段对齐模型，利用了锚词对信息提高语段对齐的准确度。针对在中小规模语料库上进行训练会产生的数据稀疏问题，提出并利用了“词性语段”概念平滑概率参数，克服了数据稀疏问题的影响。同时，系统没有把语段的切分和对齐分开进行，而是在语段对齐的同时排除语段切分时产生的歧义，提高了模型的精确性。实验结果表明模型具有较高有效性。

参考文献

[1] Satoshi Sato and Makoto Nagao. Toward Memory-based Translation. Proc. of COLING-90, Helsinki, Finland, Vol.3. pp247-252, 1990

[2] E. Sumita and H. Iida, Experiments and Prospects of Example-Based Machine Translation. Proc. 29th Annl. Meeting Assn. Computational Linguistics, Berkeley, Ca, 1991.

[3] Satoshi Sato. CTM: An Example-Based Translation Aid System. Proc. of COLING-92, Vol. IV, pp1259-1263, 1992.

[4] Ralf D. Brown, Example-Based Machine Translation in the Pangloss System. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp. 169-174. Copenhagen, Denmark, August 5-9, 1996.

[5] Peter Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19: pp263--312, 1993.

[6] Y. Wang and A. Waibel. Decoding Algorithm in Statistical Machine Translation. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics Madrid, Span. July 1997.

[7] Y. Wang and A. Waibel. Modeling with Structures in Statistical Machine Translation. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Montreal, Canada. August 1998.