基于双语语料的单个源语词汇和目标语多词单元的对齐

陈博兴 杜利民

中国科学院声学研究所语音交互技术研究中心
北京市中关村路17号 100080
{chenbx, dulm}@iis.ac.cn

摘要: 多词单元包括固定搭配、多词习语和多词术语等。本文提供了一个基于双语口语语料库的自动对齐单个源语词汇和目标语多词单元的算法，算法一方面通过计算对应于同一个源语词汇，多个目标语词汇之间的互信息和 t 值的归一化差值的大小来衡量目标语多个词语之间的关联程度以提取多词单元，另一方面通过计算互信息和 t 值的平均值作为多词单元和单个源语词汇之间互为相互翻译的衡量程度，用局部最优、首尾禁用词过滤以及长词优先等策略很好地解决了这个问题。另外，对短语翻译词典的分级，有效地减少了高级别词典中非正确翻译项的数目，使得翻译词典具有更好的实用性。

关键词: 双语对齐、多词单元、翻译词典、平均关联值、关联值归一化差值

引言

1 双语词典自动提取的背景

双语短语（固定搭配、多词习语和多词术语等，以下同）的自动提取是双语语料自动对齐的一个重要分支，主要应用于机器翻译、机器辅助翻译、双语词典编纂、术语学、信息提取、自然语言生成等自然语言处理技术以及应用于第二语言教学等等。从上个世纪80年代以来双语语料的自动对齐技术得到了很大的发展，进入 90 年代中后期，国内外很多研究人员开始进行双语翻译词典的自动生成的研究，如 Wu 和 Xia（1995）、Hiemstra（1996）、Melamed（1996）[1]等等都进行过双语翻译词典的研究，他们的工作主要是研究单个词语的对齐。同时，单一语言的短语提取也得到了很大的发展，Church et al.（1990）[2]应用互信息来表征两个单词的关联程度，从此互信息在词典自动提取的研究中一直扮演着一个重要的角色，事实上，目前为止互信息在用统计的方法提取双语词典和提取短语的研究中是用得最多的关联度衡量参数。Smadja（1993）、Nagao 和 Mori（1994）、Kita et al.（1994）、Zhou 和 Dapkus（1995）、Johansson（1996）、Shimohata et al.（1997）、Yamamoto 和 Church（1998）都利用了互信息（或互信息的变形）作为重要参数来提取短语，但是这些方法的缺点在于低频短语容易被排除掉，因为互信息利用的是概率信息，这就有一个样本无穷大的假设，所以当语料库规模不大的情况下，效果不理想；而且提取结果很大程度上依靠于算法中循环开始时合适的 Bigram 的确定。

双语短语自动提取的工作是建立在双语单个词语对齐和单一语言短语自动提取的基础上的。这方面的研究也已经开展了不少，如 Smadja et al.（1996）[3]、haruno et al.（1996）[4]、Melamed（1997）[5]、Tanaka 和 Mtsuo（1999）[6]等等，但是他们的工作的问题在于继承了所有单语短语提取所遗留的问题，而且过分地依赖统计而忽视了语言本身的特点。所有这些研究中，基于汉英双语语料的研究不多，尤其

是口语语料的几乎很少有。

2 本文的研究对象

在这之前已经完成了单个英文单词和单个汉语词语的对齐，在召回率为93.5%的情况下，正确率达到93.389%。分析结果，可以看出对于一个英文单词，如果它对应的汉语翻译被切分成了若干个汉语词语，那么这几个汉语词语与该英文单词之间的互信息以及t值都很高，而且这几个互信息以及t值的大小比较接近，于是可以用平均互信息和归一化的互信息差值以及平均t值和t值归一化差值来提取一个英文单词对应的汉语词多词单元，这种情况最典型的就是一个英文人名单词对应多个汉语单字，因为一般的汉语分词词库都没有收录人名，所以英文名字的汉语翻译在经过分词后一般都会被切分开，还有比较新的英文术语（汉语词典还没来得及收入）也属于这种情况，如“Patterson：佩特逊”；当然还有非术语的情况也很多，如“my：我的”，一般的分词词库都是不会把“我的”作为一个词语收入词库的。同样也可以提取一个中文词语对应多个英文单词所构成的短语，如“不三不四：get funny”和“放肆：get fresh”等等。所以本文的研究就集中在一个源语词汇和多个目标语词汇所构成的短语的对齐上。源语和目标语多词单元的提取将是下一步研究的重点。

算法

提取双语短语的方法分为以下几步：

（1） 分词，因为中文没有词边界，所以分词是必不可少的第一步；

（2） 统计共现频率，只要候选词对出现在一个对齐的双语对中，则当该词对共现一次；

（3） 单个词关联值的计算，计算任意一个源语词汇与它所有共现的单个目标语词汇的互信息和t值；

（4） 计算每个源语词汇和与其共现的目标语的N-gram（N：2~7）之间的平均互信息、归一化互信息差值、平均t值和t值归一化差值；

（5） 局部最优算法，利用局部最优算法将非局部最优的候选短语排除掉；

（6） 首尾禁用词表过滤，某些词语是不能作为一个短语的第一个词或者最后一个词的，通过词表将这一部分候选短语过滤；

（7） 关联值最优，将前两次过滤剩下的候选短语取其中平均互信息和平均t值最大的N项作为可能的目标语翻译候选项；

（8） 长词优先策略，因为提取的是短语而不是词语，所以如果一个较长的词串C_1完全包含另外一个较短的词串C_2，则提取C_1作为源语词汇的翻译项；

（9） 按照同时满足四种参数中部分或全部参数进行四个信任值级别的词典分类。

下面以提取“Glasgow：格拉斯哥”为例来说明整个算法的过程。

1 汉语分词

所采用的分词方法是“最大概率分词方法”[13]，其中的分词词典是北大的《现代汉语语法信息词典》[14]。首先将词典整理成以下格式：词语、拼音、词性、概率，每个词的概率可以用频率来估计。词表中的概率是通过统计采用“平均词长最大法”切分94年《人民日报》所得结果而得到的。该方法的基本思想是：先根据词表把输入串中的所有可能的词都找出来，然后把所有可能的切分路径（词串）都找出来，并且从这些路径中找出一条最佳的（即概率最大的）路径作为输出结果。随机抽样1000个句子进行检验，在不考虑未登录词的情况下的切分正确率为98.88%，在考虑未登录词的情况下，切分的正确率为

88.74%。未登录词在双语语料库 DECC1.0（Daily English-Chinese Corpus）中主要是外国人名和地名的中文翻译，关于切碎词的聚合正是这里要研究的重点。

2 统计共现频率

该算法本质上也是依据双语词汇的共现概率来计算它们的关联度（互信息和 t 值），按照关联度的高低从而确定两个双语词汇是否为相互翻译。直观的来看，每一对双语句子必定是相互正确的翻译，如果一个候选词对在对齐的句子中同时出现的概率很高，那么这个词对是相互翻译的可能性也就高。而共现模型就是限制在什么情况下算是共现，只要候选词对在一个句子对中出现一次就认为它们共现一次，这是因为首先，口语句子的长度一般比书面语要短，DECC1.0 语料库每个英文句子的平均长度是 7.07 个单词，每个汉语句子的平均长度是 6.87 个词语，其次，在英汉口语句子对中对应的意群在位置上的并不总是对应。

3 计算互信息及 t 值

统计出所有的共现频率和每个词汇的出现频率后就用公式（1）和（2）计算计算任意一个源语词汇与它所有共现的单个目标语词汇的互信息 $MI(S,T)$ 和共现 t 值 $t(S,T)$。$t(S,T)$ 作为假设校验值[7]，如果 t 值越高，说明 S 和 T 的相关联的程度也越大。

$$MI(S,T)=\log\frac{\Pr(S,T)}{\Pr(S)\Pr(T)} \qquad (1)$$

$$t(S,T)\approx\frac{\Pr(S,T)-\Pr(S)\Pr(T)}{\sqrt{\frac{1}{N}\Pr(S,T)}} \qquad (2)$$

其中：S 表示源语词汇，T 表示目标语词汇，*Pr(S,T)*表示 S 和 T 的共现概率，*Pr(S)*表示 S 出现的概率，*Pr(T)*表示 T 出现的概率，N 表示句子对的总数 。根据公式（1）（2）计算出的结果以“Glasgow”为例，如图 2、3 所示。

Glasgow：
- 哥：8.004633
- 格：6.723699
- 拉：6.669632
- 飞往：6.087710
- 斯：4.637337
- ……

Glasgow：
- 哥：1.413741
- 格：1.412514
- 拉：1.412419
- 斯：1.400519
- 飞往：0.997729
- ……

4 计算平均关联值及其归一化差值

计算与“Glasgow”共现的，连续的汉字串 n-gram (n：2~7) 之间的平均互信息、归一化互信息差值、平均 t 值和 t 值归一化差值。因为根据 Spela Vintar[8]的研究表明英语和斯拉夫语短语的长度 95%的其中在 2~6 个单词，根据经验可知，长度超过 6 个词语的汉语短语肯定也是极少数，为了减少计算复杂度，

这里也只考虑长度小于和等于6个词语构成的短语。假设一个汉字串C用下面的符号表示：

$$C = W_1 W_2 ... W_i ... W_n \qquad (7)$$

则平均互信息AMI（Average Mutual Information）、归一化互信息差值MID（Mutual Information Difference）和平均t值AT（Average T-score）、t值归一化差值TD（T-score Difference）和的计算公式分别如下：

$$AMI(C,T) = \frac{1}{n}\sum_{i=1}^{n} MI(W_i,T) \qquad (8)$$

$$MID(C,T) = \frac{1}{n * AMI(C,T)}\sum_{i=1}^{n} | MI(W_i,T) - AMI(C,T) | \qquad (9)$$

$$AT(C,T) = \frac{1}{n}\sum_{i=1}^{n} t(W_i,T) \qquad (10)$$

$$TD(C,T) = \frac{1}{n * AT(C,T)}\sum_{i=1}^{n} | t(W_i,T) - AT(C,T) | \qquad (11)$$

根据公式（8）～（11）计算出来的结果如表1所示：（每个参数计算出来的结果均有108项，为了节省篇幅，这里只选择与正确选择项“格拉斯哥”有关而且能说明算法的11项）

5 局部最优算法

到目前为止，提取短语的算法大多是基于为某个关联值（互信息、熵、相互期望值等）在全局范围内设置一个阈值，只有当被考察的词串的关联值大于这个或者小于该阈值的时候就认为该词串是一个短语。但是阈值的方法有很多局限性，因为阈值会随着语言种类的改变、语料的多少以及所选取的关联值的不同而发生改变，而且选取阈值也主要是凭观察或者小范围内的统计而得到的，所选取的阈值是否合适得不到保证。

局部最优算法[9]提供了一个鲁棒性更强、适用范围更广、更为灵活的提取短语的手段。如果每一个词串（n-gram）是一个短语，那么会有着更强的内在关联，同时它的关联值肯定也会更高，并且一个短语是一个局部的结构，在一个局部能表现出最优的关联程度，而在全局范围内可能会因为它出现的频率太低等原因而表现不出在全局范围内有优势的关联值来，所以当一个词串的关联值在一个局部表现出最优，那么可以认为该词串就是一个短语。

把一个n-gram 词串C（Chunk）包含的所有(n-1)-gram的集合用Ω_{n-1}表示，而所有包含该n-gram 词串C的(n+1)-gram的集合用Ω_{n+1}表示，假设关联值S(.)越大，结果就越优，则局部最优算法可以表述如下：

算法1. 局部最优算法

$\forall x \in \Omega_{n-1}$，$\forall y \in \Omega_{n+1}$ 如果

(length(C) = 2 and S(C) > S(y)) 或者
(length(C) > 2 and S(x) <= S(C) and S(C) > S(y))

则 词串 C 是一个短语。

其中 length(C)表示词串 C 所包含的词语的个数。

表1 与“Glasgow”共现的汉语 Ngram（N=2～7）的 AMI、MID、AT 和 TD

	AMI	MID	AT	TD
飞往格	6.405704	0.049642	1.205121	0.172093
飞往格拉	6.493680	0.041679	1.274221	0.144659
飞往格拉斯	6.029595	0.115452	1.305795	0.117961
飞往格拉斯哥	6.424602	0.132251	1.327384	0.099340
格拉	6.696666	0.004037	1.412466	0.000034
格拉斯	6.010223	0.152283	1.408484	0.003770
格拉斯哥	6.508825	0.143765	1.409798	0.003291
格拉斯哥吗	5.474901	0.363350	1.275428	0.168565
拉斯	5.653485	0.179738	1.406469	0.004230
拉斯哥	6.437201	0.186402	1.408893	0.003962
拉斯哥吗	5.162702	0.421181	1.241156	0.202718

具体到该算法，AMI 和 AT 越大表示越优，MID 和 TD 越小表示越优，与“Glasgow”共现的各 n-gram 中局部最优的各项在表 1 中已用黑体标明。

通过局部最优算法提取出来的短语还存在以下两个主要问题：（1）一小部分提取的短语不正确，最主要的现象是一些象“的传球”和“没法把”这样的短语，不合适的词语出现在短语的开始或者结尾；同样的情况在提取的英语短语中也出现，比如“and、or”等单词出现在短语的开始，“the、my、if”等单词出现在短语的结尾。（2）对于一个源语词汇，提取的短语有多个，但是不全部是正确的翻译。

针对第一个问题，通过首尾禁用词表过滤来解决。第二个问题通过关联值最优和长词优先的策略来解决。

6 首尾禁用词表过滤

所谓首尾禁用词就是某些不能作为短语的开始和结尾的词语。通过分析词性和具体词语的搭配特点手工生成了一个首尾禁用词表。一共分为四个表，分别是汉语短语非开始词表、汉语短语非结尾词表、英语短语非开始词表、英语短语非结尾词表。每个表的大致构成如下：

汉语短语非开始词表：量词（次）、助词（的）、语气词（吗）等267个词语

汉语短语非结尾词表：连词（和、或者）、部分介词（从）等189个词语

英语短语非开始词表：部分副词(not)、部分连词（and、or）等23个词语

英语短语非结尾词表：冠词(the)、连词（when）、情态动词（ought to）、部分代词（my）等78个词语。

其中汉语禁用词表还涉及到英文人名、地名的翻译中用到一些属于上述被禁的词性，比如“克”属于量词，但是在人名中出现的比较多，于是把“克”从禁用词表中删除，英语姓名译名词表参照了刘开瑛的研究结果[15]。

将所提取的短语中首尾词语出现在首尾禁用词表中的短语过滤掉，就能很好地解决上面提到的第一个问题。

7 关联值最优过滤

因为关联值（互信息和t值）是源语词汇和目标语短语互为相互翻译的一个衡量标准，所以如果一个源语词汇所对应的目标语短语有多个，那么关联值较高的目标语短语更可能是源语词汇的翻译。所以将前两次过滤剩下的候选短语取其中平均互信息和关联校验平均值最大的N项作为可能的目标语翻译候选项。根据抽样检验，经过局部最优过滤剩下的目标语中正确的翻译项的关联值一般是最大的前三位，所以这里取N等于3。

8 长词优先原则

虽然词长短的更可能是词语[10]，但是因为（1）采取的算法决定了两个词语构成的短语，尤其是与源语词汇关联值最大的两个词语所构成的短语，有更大的平均关联值和更小的关联差值，正如在前面的例子中可以看到“格拉”比“格拉斯哥”在四个参数上都有着更优的结果。（2）提取的是短语而不是词语，如果一个更长的词串有着局部最优的结果，说明这个词串是一个相对稳定的结构。所以如果一个较长的词串C_1完全包含另外一个较短的词串C_2，则提取C_1作为源语词汇的翻译项。

9 信任值分级

至此，针对每个源语词汇所提取相应短语翻译项的工作已基本完成，根据四个参数一共生成了四个双语词典，可以按照不同的应用要求对四个词典进行合并处理或者交集处理，将同时出现在四个表当中的候选词对设置信任分数值为4，将所有信任分数值为4的候选词对提取出来作为一个词典，命名为“4级词典”，出现在三个表当中的候选词对的信任分数值为3，将所有信任分数值为3的候选词对提取出来作为一个词典（“3级词典”），依此类推。也可以把只要在四个词典中出现过的词条都收入一个词典，将该词典命名为“0级词典”。如果一个源语词汇提取的目标语翻译有多项，则统计每个翻译项在语料中与源语词汇共现的频率，然后对每个翻译项进行概率归一化。

实验结果及分析

1 双语语料

采用的双语语料库DECC1.0（Daily English-Chinese Corpus）的内容主要是日常生活对话用语，包括14974对已经对齐的双语句子，总字节数为1,039,183 bytes。

2 词与词对齐的结果

在先前完成的词与词的对齐工作中，首先利用释义词典从双语文本中过滤得到一部分翻译词典，继而通过统计共现概率，计算出候选词对的互信息和t值得到正确率和召回率不同的4个级别的词典，其中“过滤词典+4级词典”在召回率为93.5%的情况下，正确率达到93.389%。

3 短语翻译词典的性能评估

以英语为源语，汉语为目标语，给出了一个“4级词典”和“0级词典”的样例，如图4和图5所示：

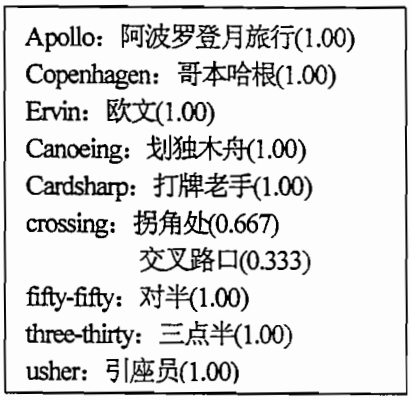

图4　4级词典样例

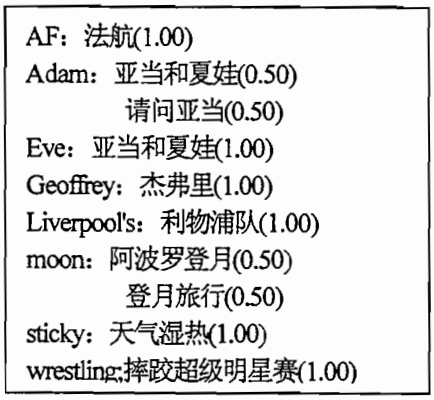

图5　0级词典样例

翻译词典的正确率（precision）没有统一的计算方法，这里采取的方法是：如果一个英文单词它所对应的一个中文翻译在语料中出现，如单词“fifty-fifty”在英汉释义词典中释义项有“平分为二的、对半地、平分为二分地”，但是在语料库中出现“fifty-fifty”的时候它所对应的翻译就是“对半”，则认为“fifty-fifty：对半”是正确的，但是如上面例子中给出的“Adam：亚当和夏娃”则认为是不正确的。召回率（recall）为每个词典中出现的英文单词数除以整个语料库中出现的单词数。

F值是一个衡量正确率和召回率之间平衡的重要参数[11]。分别从各种词典中随机提取了100个词条（包括一个源语词汇词和它的所提取出来的目标语翻译项）统计了正确率、召回率和F值。表2为英汉、汉英0～4级词典的正确率和召回率和F值。

$$F = 2\frac{recall \times precision}{recall + precision} \qquad (7)$$

其中recall为召回率，precision为正确率。

表2　各级词典的正确率、召回率

词典	正确率（%）	召回率（%）	F-measure
英汉0级词典	41.394	98.63	0.583
英汉1级词典	23.535	84.22	0.368
英汉2级词典	52.388	31.56	0.394
英汉3级词典	78.323	5.18	0.097
英汉4级词典	94.900	1.36	0.027
汉英0级词典	38.266	96.94	0.549
汉英1级词典	18.943	82.58	0.308
汉英2级词典	47.564	29.92	0.367
汉英3级词典	75.092	7.54	0.137
汉英4级词典	88.293	2.83	0.055

4 短语翻译词典的结果分析

分析各级词典的正确率和召回率和具体的词条，可以得出以下一些结论：

1．满足一个条件（即1级词典）的词条很多，差不多每个英文单词和中文词语都能有一个或一个以上对应的目标语词串能满足局部最优和其他条件，但是1级词典的正确率很低，这说明（1）光靠单一条件很难提取正确率高的双语词典（2）不是每个源语词汇都有一个目标语短语与其对应的；

2. 级词典和 1 级词典相比正确率提高幅度很大，据统计，2 级词典中正确的一部分绝大多数满足的两个条件是互信息的两个条件或者是 t 值的两个条件，这说明对于一种参数（互信息或者 t 值），同时引入差值和平均值能大大改善结果；

3. 级词典相比于 2 级词典，正确率提高的幅度也同样很大，同时召回率大大下降，这说明在满足一种参数后如果还能满足另外一个参数的一个条件，那么就很有可能是正确的了，在其他研究人员从事的相近的工作中，很多工作都是考虑多个参数的，但是参数的选择很重要，通过前期的词与词的对齐工作以及现在短语的提取工作，发现 t 值和互信息是两个配合得很好的参数；

4. 级词典只要经过小量的人工校对工作就具有实用性了，其中英汉 4 级词典中只提取了 98 个词条，除了一些出现频率较高的固定短语外，基本上都是人名、地名和比较专业的术语，这些术语出现的频率都很小，很多甚至在语料中只出现过一次，这说明该算法对于低频短语的提取同样有很好的效果。汉英 4 级词典提取了 205 个词条，对比这两个词典中词条的数量，可以看出：相对于英语单词，更多的汉语词语需要一个英语短语来翻译，从另一个侧面证明了英语相对于汉语有更大的冗余度。

5. 级别越高的词典召回率过低，这一方面，说明单个源语词汇对应一个目标语词串的情况相对来说是有限的，另外一方面与语料库太小有关系，如果语料库再适当地加大，会取得更佳的结果。

6. 各级词典中都存在有短语扩大的情况，如 4 级词典中“Apollo：阿波罗登月旅行”，“Apollo”对应的应该是“阿波罗”，但是由于语料中出现“Apollo”只出现一次，而且在例子中“阿波罗登月旅行”是作为一个意群存在的，所以在提取短语时根据长词优先原则，该算法选择了“阿波罗登月旅行”，应该看到，虽然“Apollo：阿波罗登月旅行”这个词条是不正确的，但是可以在这个基础上提取源语和目标语均是多词单语的翻译词典，而这正是下一步工作的重点。

结论及下一步研究

1 结论

本文提供了一个基于双语对齐口语语料的自动对齐单个源语词汇和目标语单语的算法，这对于提高翻译词典的实用性有很大的帮助，因为一方面确实存在很多单个源语词汇和多词目标语短语对应的情况，最典型的就是英文人名、地名，解决了这个问题对于机器翻译，尤其是汉英翻译有很大的帮助，另外一方面也可以作为提取双语多词短语的一个前期工作。该算法和其他同类工作相比不同之处在于：

（1） 采用了关联值的归一化差值作为提取短语的判断依据；

（2） 对于短语的提取同时综合了局部最优、首尾禁用词过滤、长词优先等策略；

（3） 根据满足参数的多少进行分级，不同级别的词典可以分别用于实用翻译词典或者作为下一步研究的基础。

互信息在其他同类研究中利用得比较多，但是在利用互信息的过程中大多是重复统计 Bigram 来完成的，这种方法的结果很大程度上取决于初始 Bigram 的统计结果，任何一种统计结果都几乎不可能达到 100%的正确率，于是在后面的重复统计 Bigram 的过程中就会使错误率成指数级的增长，严重地影响了多词短语提取的正确率[12]，而该算法通过计算对应于同一个源语词汇目标语词汇之间的关联值的归一化差值很好地解决了这个问题。另外，t 值使用大大地提高了短语翻译词典的正确率，对短语翻译词典的分级，有效地减少了高级别词典中非正确翻译项的数目，使得翻译词典具有更好的实用性。

2 下一步研究计划

该算法的结果中短语“扩大”是一个出现得比较多的问题，下一步的研究中将会利用这个结果来提取双语均是多词的短语。另外一个问题就是对于一个源语词汇，在前期的词对词的对齐工作中，分别给每个源语词汇找到了一个或者多个关联值最大的单个目标语词汇，在现在的工作中，同样给每个源语词汇找到了一个或者多个关联平均值局部最大而且关联差值局部最小的目标语词串，但是问题在于对于该源语词汇的正确翻译是选择目标语的单个词汇还是一个词串还没有得到解决，因为词串的平均关联值肯定小于单个词汇的最大关联值，而且单个词汇没有关联差值。在以往的研究工作中，很多研究者都回避了这个问题或者是没有意识到这个问题，这也是今后将要重点考虑解决的问题。

参考文献：

[1] Melamed I. D., Automatic Construction of Clean Broad-Coverage Translation Lexicons. In: Conference of the Association for Machine Translation in Americas, Montreal, Canada, 1996.

[2] Church K.W. and Hanks, Word association norms, mutual information and lexicography. In: Computational Linguistics 16(1):22~29 1990.

[3] Smadja F., McKeown K.R. and Hatzivassiloglou V., Translation collocations for bilingual lexicons: a statistical approach. In: Computational Linguistics 22(1):1~38 1996.

[4] Haruno M., Ikehara S. and Yamazaki T., Learning bilingual collocations by word-level sorting. In: COLING96 (pp. 525~530) 1996.

[5] Melamed I. D., Automatic Discovery of Non-Compositional Compounds. In: Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, Providence, RI 1997.

[6] Takaaki Tanaka and Yoshihiro Matsuo, Extraction of compound noun translation from non-parallel corpora. In: Proc. of the 5th Annual Meeting of the ANLP, Japanese 1999.

[7] Fung P., Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In: proceedings of the 33th Annual Meeting of tha Association for Computational Linguistics, Boston, USA. 1995.

[8] Vintar, Spela, Using Parallel Corpora for Translation-Oriented Term Extraction. In: Babel Journal, John Benjamins Publishing, 2001.

[9] Silva J.F., Dias G., Guillor S. and Lopes J.G.P., Using Localmaxs Algorithm for Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In: 9th Portuguese Conference in Artificial Intelligence, Lecture Notes, Springer-Verlag, Universidade de Evora, Evora, Portugal. 1999.

[10] Tanapong Potipiti, Virach Sornlertlamvanich and Thatsanee Charoenporn, Towards Building a Corpus-based Dictionary for Non-word-boundary Language. In: Workshop on Terminology Resources and Computation, Workshop Proceedings of the Second Internationl Conference on Language Resources and Evaluation (LREC2000), Athens, Greece, 2000.

[11] Langlais P., Simard M. and Véronis J., Methods and Practical Issues in Evaluating Alignment Techniques. In: Proceedings of COLING-ACL, Montréal, Canada. 1998.

[12] G. Dias, S. Guilloré and J.G. Pereira Lopes, Normalization of Association Measures for Multiword Lexical Unit Extraction. In: International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications (ACIDCA'2000), Monastir, Tunisia, 2000.

[13] 陈小荷. 现代汉语自动分析. 北京:北京语言文化大学出版社，1999.

[14] 俞士汶等. 现代汉语语法信息词典详解. 北京:清华大学出版社，1998.

[15] 刘开瑛. 中文文本自动分词和标注. 北京：商务印书馆，2000.

作者简介： 陈博兴（1975－），男，湖南人，博士生，主要研究领域为自然语言处理；杜利民（1957－），男，四川人，博士，中国科学院声学研究所主任研究员。多年从事语音信号与信息处理技术的研究，在汉语话者无关的连续语音鲁棒识别和连续语音关键语检测、语音交互助理、自然口语语音对话、语音翻译、强噪声环境下语音增强和语音提取、低速率语音压缩等方面有深入研究。

Alignment of Single Source Words and Target Multi-word Units from Parallel Corpus

Boxing Chen Limin Du

Center for Speech Interaction Technology Research

Institute of Acoustics, Chinese Academy of Sciences

17 Zhongguancun Rd. Beijing 100080, China

{chenbx , dulm}@iis.ac.cn

Abstract: Multi-word unit includes steady collocation, multi-word phrase and multi-word term, this paper we provide an algorithm for automatic alignment of single source words and target multi-word units from sentence aligned parallel spoken language corpus. Mutual information has been used to extract multi-word units by many other researchers, but the retrieval results mainly depend on the identification of suitable bigrams for the initiation of the iterative process. This algorithm utilizes normalize mutual information difference and normalize t-scores difference between multi target words correspond to the same single source word to extract the multi-word units, then utilizes the even mutual information and even t-score to align the single source words and target multi-word units. In this algorithm, we have applied the Local Bests algorithm, stopword filter and long-length units preference methods et al. The grading of the lexicon can deduce the number of the incorrect entries in the high level lexicon effectively, which makes the translation lexicon more practicably.

Key words: bilingual alignment; multiword unit; translation dictionary; even association score; normalize association score difference;