基于标志词的EBMT翻译单元抽取方法[1]

刘海洁　时晓升　姚建民　李生

（哈尔滨工业大学计算机科学与技术学院，黑龙江　哈尔滨　150001）

E-mail: (hjliu, xshshi, james, shengli)@mtlab.hit.edu.cn

摘要： EBMT技术的关键问题之一是如何能取得较好的翻译单元。传统的方法有通过双语语料的词语对齐来取得对译片断，也有通过对句子的分析（如依存分析）、根据句子分析结果来获取翻译单元。然而词语对齐或句子分析结果都存在一定误差，这对翻译单元的抽取产生了负面的影响。本文试图通过汉英双语标志词(Marker)找到一种新的翻译单元的抽取的方法。该方法不需辅助的自然语言处理技术，在关于奥运的体育领域的上获得了较好的效果。在此方法上，我们实现了一个EBMT系统，该系统取得了较好的翻译效果。

关键词： 基于实例的机器翻译；翻译单元抽取；标志词假设；

引言

在当前的EBMT研究中，如何取得大量的高质量翻译单元是研究的主要内容之一。传统的对译碎片抽取方法通常依托于自然语言处理技术（刘占一，2003[1]；Hideo Watanabe and Sadao Kurohashi and Eiji Aramaki，2003[2]；Kaoru Yamamoto, Yuji Matsumto[3]）等。这样就造成翻译单元的抽取过程的复杂性，同时还会出现累计误差。为此，我们提出一种不需句法分析的基于标志词的翻译单元抽取技术。

本文内容安排如下：第1节讨论标志词的一些基本概念；第2节介绍基于标志词的EBMT系统；第3节讨论翻译单元的抽取策略，第4节给出了我们的实验结果及相关讨论，第5节给出了实验结论。

1 标志词的基本概念

1.1 标志词的描述

“标志词假设（Marker Hypothesis）”[Green, 1979][4]是一个语言学限制，是指语言被由词位（lexemes）和词素（morphemes）中的特定词语组成的一个封闭集合“标记”的表层语法结构。其中起“标记”作用的词就是本文所指的标志词。在若干以前的机器翻译系统中，标志词已经作为语言学的提示性信息被加以利用，如METLA [Juola，1994，1997] [5], Gaijin [Veale & Way, 1997] [6]，和wEBMT [Gough et al., 2002; Way & Gough, 2003; Gough & Way, 2003, 2004][7]。

在我们的实验中，英文的标志词主要是如下这些词性的词语：

be动词（BE），连词（CONJ），限定词（DET），英文标点（PNCT），物主代词（POSS），人称代词（PPRON），介词（PREP），数词（QUANT），疑问代词(WRB)。例如以下英文句子：

My father <BE>01 is also <DET>02 a football fan <PNCT>03 . <CONJ>04 But it seems that <PPRON>05 he has never liked Argentina.　　（例句一）

上面例子中的所有前面标有“<词性>”的词语都是标志词。我们可以看到，这个例子中有如下标志词：is(系动词)、a（限定词）、.（标点符号句号）、but（连词）、he（代词）。这些词把句子分割成了较为独立的语法语块：“My father”，“is also”，“a football fan”，“But It seems that”，“he has never like Argentina”。这些语块对EBMT有巨大的实用价值。

同样，我们把汉语中和英语相对应的词性的词语作为标志词。主要是指如下词性的词语：

[1]本论文受到以下资助：

1. 基于双语信息的英汉译文消歧技术研究，项目批准号：60375019；
2. 中国-爱尔兰合作研究项目：基于大规模双语WEB资源的EBMT研究，项目号：CI-2003-03

“的、之”(usde)，时态助词（ut)，其他助词（ur)，方位词（f)，连词（c)，介词（p)，数词（m)，代词（r)，系动词（vx)，介词（p)。如以下例句：

<r>01我 爸爸也<vx>02是 个足球迷<wo>03，<c>04但 <r>05他 似乎不怎么喜欢<r>06这个队。 （例句二）

这句话中的标志词如下：我（代词)，是（系动词)，标点符号逗号（标点)，但（连词)，他（代词)，这个（代词)。这些标志词把句子分割成：“我爸爸也”，“是个足球迷”，“他似乎不怎么喜欢”，“这个队”。

2 基于标志词的EBMT系统

标志词假设是语言学上的提示信息，对从对齐后的双语文本<source, target>中抽取子句的对齐有很好的贡献。我们已经收集了英语和汉语的标志词，如连词、介词等，这样的标志词组成了一个词典。接下来，这些标志词将把对齐的句对划分成若干碎片。每个碎片都是从一个标志词开始，到下一个标志词结束，并且每个碎片中至少包含一个以上的非标志词成分（即内容词，content word)。我们实验了若干策略以取得基于标志词的双语对齐文本的翻译碎片对应。简单来说，我们使用了标志词标记，相关的标记位置，以及每个标志词分割的语块中的内容词等信息来确定双语碎片的对应。之后可以据把翻译单元加进EBMT系统中完成翻译任务。

下面是完整的EBMT系统流程图：

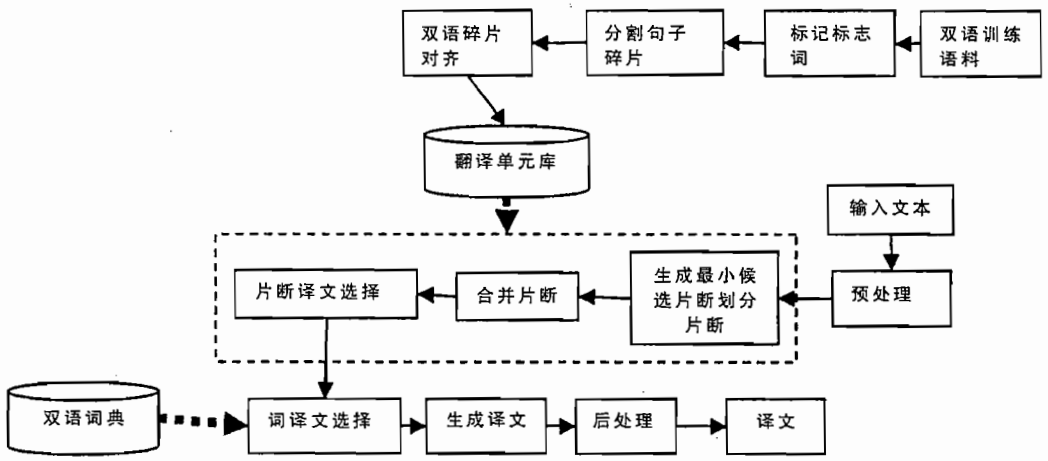

图2-1 EBMT系统流程图

3 翻译单元的抽取

3.1 语料库规模

由于本系统准备面向2008奥运会，故侧重于收集奥运领域的体育语料。目前共收集了奥运口语双语语料2718个句子，其有关信息请参见表3-2。

3.2 标志词对应

确定了标志词后，我们还尝试把这些双语文本中的标志词进行对应识别，以应用在翻译单元的抽取中。如对于第1节的例句一和例句二，有如下的标志词对应：

表3-1 一个实例的标志词对应

汉语标志词编号	英语标志词编号
02	01
03	02
04	04
05	05

标志词大都是没有很确定意义的虚词,对应的确定比较困难。我们对应标志词的基本思想是查找辞典。我们实验了两中标志词对齐的算法:

1. 标志词对应贪心算法1

如果一个英文标志词在汉英词典中是中文标志词的一个词义，就认为这两个标志词是对应的。一个句子中可能存在一个中文标志词对应多个英文标志词的现象，我们在中文标志词找到一个与之对应的英文标志词后，将这个英文标志词做上已被匹配的标记，句子中其它中文标志词不能再与它匹配，而中文标志词找到一个匹配的英文标志词后不再向后继续查找。

2. 标志词对应贪心算法2

与贪心算法1相比，在已知前一个英文标志词被匹配以后，查找下一个中文标志词对应的英文标志词时，从前一个英文标志词位置开始，找最近的能匹配上的英文标志词。

下面是标志词对齐的实验结果:

表3-2 文本信息

实验数据	文本规模
中文标志词总数	9394
英文标志词总数	9446
句子数	2718
中文平均每句标志词数	3.46
英文平均每句标志词数	3.48
中文总词数	32492
英文总词数	29042
中文每句平均词数	11.95
英文每句平均词数	10.69

表3-3 标志词对齐算法实验结果

实验数据	算法1	算法2
手工标志词匹配对数	3180对	3180
机器自动标志词匹配对数	2868对	2868
手工与机器相同匹配对数	2217对	2217
机器自动标志词匹配精确率	77.3%	77.3%
机器自动标志词匹配召回率	69.7%	69.7%

从实验结果看，这两种方法的精确率和召回率都不是很高。究其原因，标志词不能表达实际语义，大部分只起句法功能，而且词典本身也不十分完备。由于测试语料的句子较短，其中包含的标志词本来就不多，故这两种方法获得的精确率和召回率是很近似的。

3.3 对译碎片的抽取策略

当取得了双语文本的标志词，我们考虑了若干策略取得翻译单元:

3.3.1 简单的抽取

由于双语句对中各自标以标志词，把各自的句子分割成了若干部分。我们简单的通过查词典计算两个句子中的所有碎片的对应，察看碎片的对应中所有对应上的词占总词数的比重，进而判断这两个碎片是否是一个对译片断。应用这样的方法，我们取得的总是最小的对译碎片。

3.3.2 基于自动确定的 marker 对应的对译碎片抽取

对于自动确定的标志词对应，如下图所示:

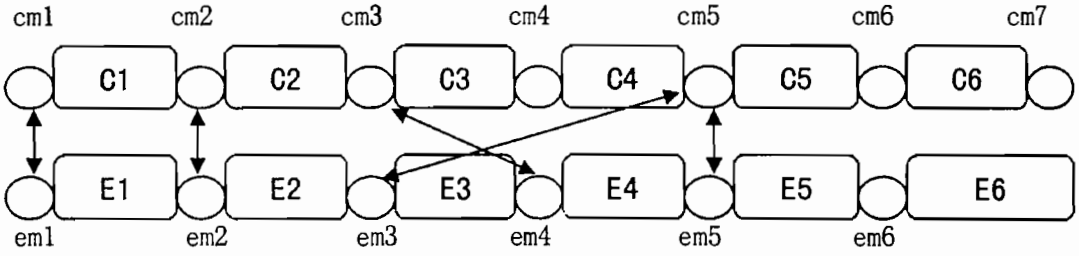

图3-1 双语marker对应图

图3.1中，cm1对应于em1，cm2对应于em2，cm3对应于em4，cm5对应于em3和em5。

则我们认为，C1和E1是一个对译片断，C2和E2是一个对译片断，C3、C4与E3、E4认为是一个对译片断，C5和C6与E5和E6是一个对译片断。

3.3.3 基于手工标注的marker对应的对译碎片抽取

与方法2很类似，区别在于，方法2中的marker的对应是通过计算机计算出来的，而在本方法中，所有的marker都是经过手工标注的，有较高的准确性。

例如根据例句一和例句二，我们可以得到如下的对译碎片：

〈我注意到， I noticed〉；〈我爸爸也， My father〉；〈是个足球迷， is also a football fan.〉；〈但， But it seems that〉；〈他似乎不怎么喜欢这个队， he has never liked Argentina.〉

3.3.4 基于自动确定的marker对应的对译碎片抽取II

这个方法中，我们简单的认为某个对应的marker其后的句子碎片也是对应的，于是，图3.1中的对译片断有：C1和E1，C2和E2，C3、C4和E4，C5、C6和E3，C5、C6和E6。

4 实验结果及讨论

试验方案即按照第二部分各个方法的策略进行碎片抽取。试验过程按照抽取翻译单元策略进行试验，共进行了四组实验。

我们使用了两种手段对翻译单元进行评价，其一为随机的抽取若干翻译单元，再手工的对这些翻译单元进行评价。其二，我们还把抽取出来的翻译单元集成进现有的EBMT系统中，察看其对翻译系统的支持程度。

4.1 翻译单元的质量评价标准

为了对抽取来的翻译单元的质量有一个量化的标准，我们把所有的翻译单元分成了如下四种类型：

4.1.1 Exact ：

汉英对译片断完全匹配（汉语和英文中的所有实词在对应的对译片断中都有相应的翻译和其对应，边界的虚词也有相应的对应）和整句的匹配（如果意思有差异，应该归结于训练语料库质量）。这样的对应应用于EBMT系统，会极大的提高系统的翻译质量。

4.1.2 Almost ：

对译碎片中汉语英语的实词有相应的解释，边界的虚词没有或不全有相应的对译词语。这样的翻译单元同样能提高系统的翻译质量。但从人的角度看译文可能会有不恰当的地方。

4.1.3 Somewhat ：

汉语碎片在英文中不能恰好的对应，不能完整的或确切的表达出。只能部分地、不确切的表达。其应用于EBMT系统后的对翻译质量的提高贡献不大。

4.1.4 Bad ：

汉语和英语的表达的意思有较大的差异，完全不能作为对译片断。在系统中应避免使用。

4.2翻译单元的质量评价

通过随机的选取部分翻译单元，之后对其进行手工的分类，我们得到了如下的结果：

表4-1 各方法抽取的翻译单元质量表

类别	方法i（%）	方法ii（%）	方法iii（%）	方法iv（%）
Exact	12	18	26	78
Almost	15	15	15	8
Somewhat	42	33	32	9
Bad	31	34	27	16

从试验中我们可以看出，第一种方法中，Exact和Almost所占的比例很低，而Somewhat和Bad的比例都很高。究其原因，由于marker并没有对齐的信息，只是把句子分割成了若干部分，而汉语和英语的碎片并不能很好的一一对应，所以在确定汉语碎片和相应英语碎片的边界上产生了问题。我们的做法是选取一个阈值，如果这些碎片的相互对应超过这个阈值则认为是一个对

应。这样就产生了问题，如果阈值过高，则对翻译语料的要求比较高(否则可能产生不了很好的对应)，并且产生的对译碎片对训练语料的覆盖率也比较低。如果阈值比较低，则又可能取到不是很好的对译碎片。方法ii和方法iii所抽取的翻译单元质量仍然不高。对于方法iii，其取得的碎片通常会比较大，使得在开放语料的测试时对译碎片对测试文本的覆盖率不高，并且对译碎片的质量仍然不高。方法ii除了有方法三的缺点外，还存在着marker对应的正确率和召回率的问题。在第四种方法抽取的翻译单元中，Exact类的比重很大，能够取得比较好的对译碎片。但这样抽取的翻译单元也有问题，就是整句的对应占了相当大的比重(整句的对应占总体的59%)。这样的结果并非翻译单元抽取策略的问题，而应归结于汉英marker的对应不是很充分。语言学上这两种语言的结构差异还是很大的。我们相信在同源语言上可取得更好的效果。

4.3系统集成试验

我们同样把这些抽取的单元应用到现有的EBMT系统中，观察其对现有系统翻译质量的贡献。下面的表4-2和表4-3是试验的结果：

表4-2 EBMT实验结果表（封闭测试）

封闭测试	答案自身得分	EBMT_HIT	EBMT_dic	方法i	方法ii	方法iii	方法iv
BLEU1元	1	0.8638	0.4866	0.5275	0.5057	0.5748	0.9339
BLEU3元	1	0.7300	0.1582	0.2467	0.3701	0.4045	0.8706
BLEU5元	1	0.6825	0.0854	0.1698	0.3340	0.3608	0.8453
NIST1元	7.3125	6.2611	2.9305	3.2408	3.4416	3.8560	6.8715
NIST3元	10.7026	8.5235	3.3489	3.9229	4.4294	4.8225	9.7005
NIST5元	10.9503	8.6444	3.3578	3.9443	4.4723	4.8544	9.8728

表4-3 EBMT实验结果表（开放测试）

开放测试	答案自身得分	EBMT_HIT	EBMT_dic	方法i	方法ii	方法iii	方法iv
BLEU1元	1	0.5107	0.4924	0.4743	0.3561	0.4112	0.4481
BLEU3元	1	0.1916	0.1735	0.1738	0.1286	0.1453	0.1764
BLEU5元	1	0.1211	0.1013	0.1065	0.0841	0.0940	0.1172
NIST1元	7.3479	3.3053	3.1559	2.9637	2.1718	2.5188	2.7609
NIST3元	11.2061	3.8942	3.6671	3.4408	2.4942	2.8764	3.2758
NIST5元	11.5639	3.9148	3.6786	3.4559	2.5013	2.8858	3.2939

以上两个表中EBMT_HIT是基于词对齐的翻译单元抽区方法的EBMT系统的得分，EBMT_dic是通过只查词典可获得的分数，作为本次实验的baseline。

从系统集成试验的结果看，方法i的结果确实很不理想，造成这种情况的原因我们已经做过分析，是由于双语碎片边界的不确定性。方法iii和方法ii取得的翻译单元质量不高，这与其翻译的质量评价也是相符的。该方法中的抽取碎片的策略并不成功。方法iv的封闭测试结果已经比较高，说明该方法对受限领域已经能取得很好的翻译碎片，并且其方法还有容易实现的优点。同时，该方法所取得的对译碎片对训练语料有较好的覆盖率。

但同时我们还注意到，尽管封闭测试中方法iv的方法取得的翻译单元在系统的集成测试上取得了很高的分数，但在开放测试中，其得分比EBMT_HIT要低，甚至比纯粹查词典的分数还低（EBMT_dic)。由于EBMT_HIT的翻译单元抽取是基于词对齐结果的，所以这还是符合我们的直觉，单纯的标志词对齐的准确率比不上信息更丰富的词对齐结果。毕竟这里的词对齐是对齐了大部分的内容词（content word)，相对于标志词对齐，其粒度更小，准确率也更高。而反过来观察，标志词对齐取得的翻译单元为什么甚至比纯粹查词典还要低的分数呢？我们观察是由于语料的不“洁净”造成的。口语语料与奥运语料的翻译都是基于句意的翻译，并非按词翻译而成。从

例句一和二我们就可以看其概貌。我们相信随着双语语料的质量的提高，基于标志词的翻译单元抽取的优势会进一步得到加强。

5 试验结论与下一步工作

本文提出了一种基于标志词的新的翻译单元抽取方法，验证了标志词的句法作用，探讨了基于此的翻译单元抽取新方法。该方法依托于标志词在句法结构上的作用，通过标志词的分割句子成分的作用以若干策略取得翻译单元。在受限的奥运领域取得了较好的翻译效果，使翻译系统的性能得到了较大的提升。接下来，如何利用混合信息（如词对齐信息、句子的依存分析结果等）以取得句法上较为独立的更高质量的对译碎片应该是可能的研究方向。

参考文献:

[1] 刘占一，2003，哈尔滨工业大学机器翻译实验室，基于实例的机器翻译研究和实现，工学硕士论文。

[2] Hideo Watanabe and Sadao Kurohashi and Eiji Aramaki, 2003, Finding Translation Patterns from Paired Source and Target Dependency Structures, (A. Carl and A. Way(eds.), Recent Advances in Example_Based Machine Translation, 397-420.)

[3] Kaoru Yamamoto and Yuji Matsumto, 2003, Extracting Translation Knowledge from Parallel Corpora, (A. Carl and A. Way(eds.), Recent Advances in Example_Based Machine Translation, 365-395.)

[4] Thomas R.G. Green. 1979. The Necessity of Syntax Markers. Two experiments with artificial languages. Journal of Verbal Learning and Behavior 18:481—496.

[5] Patrick Juola. 1994. A Psycholinguistic Approach to orpus-Based Machine Translation. In CSNLP 1994; 3rd International Conference on the Cognitive Science of Natural Language Processing, Dublin, Ireland, [pages not numbered].

[6] Tony Veale and Andy Way. 1997. Gaijin: A Bootstrapping, Template-Driven Approach to Example-Based Machine ranslation. In International Conference, Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, pp.239—244.

[7] Andy Way and Nano Gough. 2003. wEBMT: Developing and Validating an Example-Based Machine Translation System using the World Wide Web. Computational Linguistics 29(3).

第一作者简介: 刘海洁，男，北京人。目前就读于哈尔滨工业大学计算机学院语言技术研究中心，硕士研究生，师从李生教授。研究方向为自然语言处理，机器翻译。

A Experimentation on Marker-Based EBMT Translation-Units Extraction

HaiJie Liu, XiaoSheng Shi, James Yao, Sheng Li

(School of computer science and Technology, Harbin Institute of Technology, Harbin 150001, China)

E-mail: (hjliu, xshshi, james, shengli)@hit.edu.cn

Abstract: As for EBMT, a key problem is that how to get good translation units. They are obtained from word sense alignment in bilingual sentence pairs traditionally, or from the result of sentence analyzing(for example, Dependency-linked parsing). But as errors accumulated in word sense alignment or analyzing result, the quality of the translation units are not very high. In this paper we try to find a new method for the extraction of translation units by the marker words in Chinese-English bilingual corpus. This method doesn't need technologies in NLP, and get pretty good result in sports domain in Olympic Games. Based on this method, we developed an EBMT system which obtains good translation result.

Key words: Example-Based Machine Translation; Extraction of Translation Unit; Marker Hypothesis;