利用平行网页建立中英文统计翻译模型

聂建云，陈江
Lab. RALI
University of Montreal
CP. 6128, succursale Centre-ville
Montreal, Quebec
H3C 3J7 Canada
Email: nie@iro.umontreal.ca

建立翻译模型的目的是试图从平行文本(或翻译例句)中自动抽取翻译关系。本文将描述我们在建立中英文统计翻译模型上的尝试。我们所用的平行文本是从万维网上自动获得的半结构性平行文本。在训练过程中，我们尽量利用文本中的 HTML 结构信息。实验表明，所训练的翻译模型能达到 80%的准确率。对于象跨语言信息检索这样的应用，这样的准确率已经能大致满足需要。这一工作表明，对于检索引擎上的问句的翻译可以使用比机器翻译成本更低的工具。

Abstract

A statistical translation model tries to capture translation relationships from a set of parallel texts (or translation examples). This paper describes our attempt to train such translation models from a set of semi-structured parallel texts in Chinese and English. These texts are gathered from the Web by an automatic mining tool – PTMiner. Our work takes advantage of the HTML structure of the texts. Some special processing is necessary on Chinese. Our experiments show that we can obtain a translation precision of about 80% with the trained model. This performance is reasonable for less critical tasks such as cross-language information retrieval. This work shows that it is possible to construct a means of query translation at a much lower cost than a machine translation system.

1 引言

随着因特网的发展，人们越来越多地面临怎样有效地查找相关外语文件的问题。例如，一个中国用户可能希望找到英语文件，而他的英语水平又不足以使他能用英语准确地表达他的需求。跨语言信息检索正是为了满足这种需要，它的目的是通过自动翻译用户的问句来帮助用户克服语言障碍。

人们一般会认为机器翻译系统是跨语言信息检索中问句翻译的最理想系统。但是，机器翻译到现在已经 50 年了，而翻译的结果却远不能令人满意，特别是对于中文。使机器翻译质量在近期内有重大突破的可能性很小。因此，我们需要别的手段来替代或补充机器翻译系统。利用平行语料进行机器辅助翻译是现在计算语言学上一个很强的趋势。这也是可以在一些应用中取代或辅助机器翻译的一种手段。通常使用的方法是训练[3]里描述的三种翻译模型之一。这一类工作先是在英语和法语之间进行的，因为英法语之间有大量的平行语料 - 加拿大 Hansard。Hansard 包含加拿大议院七年中英法对应的

发言稿。在我们以前的试验中发现以 Hansard 训练出来的统计翻译模型对跨语言信息检索非常有用。它的效果与使用最好的机器翻译系统相近。然而其代价却要低很多。

本文将介绍我们在中英文之间使用统计翻译模型的尝试。这一工作与我们以前的工作有以下两点不同：1) 中英文之间的差别要比英法文之间大很多，因此语言的相似性不再适用(如对人名)。2) 所用的平行语料是自动从万维网上获取的，它们的平行性显然要比 Hansard 差很多。我们这一工作的目的是探索在以上两个差异下统计翻译模型是否仍然对跨语言信息检索有较好的效果。

因为万维网上的文件是半结构性的 — 它们含有 HTML 标志，所以我们在语料自动对齐中尽量利用这一特性。另外对于中文也要作一些特别处理(如切词)。

在本文中，我们将顺序介绍平行语料的自动获取和预处理，以它为基础的建模，以及对翻译模型和跨语言信息检索质量的测试。

2 平行语料及预处理

一般统计计算语言学的研究都是以语料为基础的。统计翻译模型更是建立在平行语料上。但是，目前所能得到的平行语料库却非常有限。因此，统计翻译模型的实际应用受到很大限制。但是我们发现，随着因特网的发展，越来越多的网站成为双语网站。这为平行语料提供了一个很大的来源。如果我们能自动从因特网上得到大量平行语料，那么统计翻译模型就能实际地用于许多语言。

我们正是为达到这一目的而试图从万维网上自动获取平行网页。所用的系统 - PTMiner 是按以下步骤运行的[7]：

1) 确定候选网站：候选网站是指可能含有平行网页的网站。确定候选网站的目的是将查找限制在可能的网站上。我们的假设是，如果某一网站有一个中文网页含有象“English version”，“ in English”之类的锚文(anchor text)，则这一网站就有可能含有平行网页。同样，如果英语网页里有“中文版”这样的锚文，则该网站也是候选网站。在万维网上已有许多大型检索引擎索引了大量的网站。通过它们我们可以很快确定某一网站是否可能含有平行语料。例如，我们可以向 AltaVista 发以下问句：

anchor: English Version

而将语言设为中文，则可得到含此锚文的中文网页。从它们的网址我们可以确定候选网站。

2) 寻找候选网站上的文本。从某一候选网站上，我们尽量多地找到所含的文本。第一步仍是通过检索引擎(如 AltaVista, NorthernLight)找到部分文本。这只要向它们发送下面的问句：

host: <host_address>

但是检索引擎只索引了部分网页。为了找到更多的网页，我们从已知的网页出发顺着它们的链结搜索网站上的其它网页。

3) 确定平行文本。如果假设每一网页都是另一网页可能的平行网页，那么这一步骤的复杂性是 $O(n^2)$，其中 n 是网页数。如果 n 很大，所需的时间会很长。为了缩短这一时间，我们运用以下的观察：平行网页通常都有相似的文件名。如，file_e.html 和 file_c.html。这使我们能通过文件名的相似性快速地确定哪些文本可能是平行的。

4) 平行文本的过滤。在上一步确定的文本中可能有不平行的文本。所以我们使用别的准则进一步过滤。这些准则包括：文本长度，文本的语言，文本内 HTML 标志的相似性。其中文本的语言的确定是由语言自动检测系统[11]完成的。

最后我们得到 14 820 被认为平行的网页。 通过对这些网页的抽查发现其中 90% 是真正平行的。这些网页就组成了训练翻译模型的语料库。在训练之前还需要对文本作一些预处理。首先，HTML 文本要被切成句子。这不但要用到标点符号，还有用到 HTML 符号。其次，中英文句子要对齐(alignment)。这除了用到句子长度以外也要用到 HTML 符号。最后还有中文编码统一转换和切词问题。预处理的结果是以下三个文件：

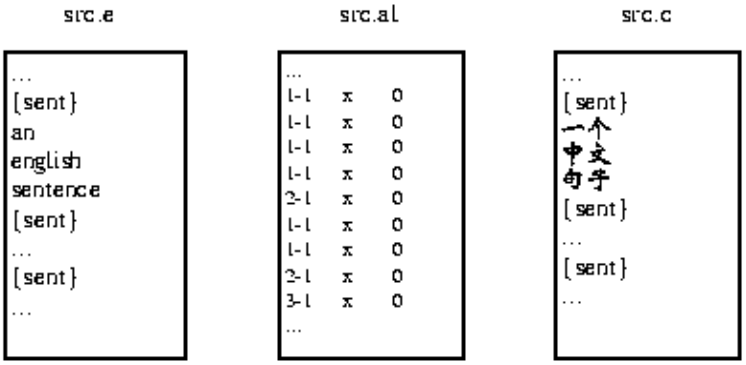

图 1. 中英文句子 (src.e and src.c) 及它们的对应关系 (src.al).

在下面的几节中，我们将进一步描述预处理中几个关键问题：句子对齐，中文切词和英文组词。

3 句子对齐

句子对齐对于统计模型训练是至关重要的一步。它的目的是使中英文词之间的对应限制在最小范围内，从而增加统计模型的准确率。当然我们希望这一范围越小越好。但同时对齐的准确率也要相当地高。在句子之间作对齐是综合这两个因素的最佳选择，因为我们能很容易地确定句子，而句与句对齐的几率也很大。

即使在句子这一层上对应也不是一件容易的事。首先，句与句之间不总是一一对应的关系。一句句子可能翻译成几句，几句可能被合并起来。有时一些句子还被删除或添加。这给确定句子的对应关系增加很多困难。特别是我们的语料库是自动建立的，含有噪音。另外，对中英文句子的对应困难更大，因为中文的书写习惯通常将几句并为一句，而在对应过程中又不能用到语言的相似性。

在西方语言之间的对应问题上，有几个常用的算法，其中包括基于长度和基于同构词(cognate)的算法。

3.1. 基于长度的算法

Brown et al. [2] 和 Gale & Church [9]平分别提出了两个基于长度的算法。其基本思想是：对应的句子的长度是相似的，而且它们的顺序也基本相同。据此，两个算法使用动态规划的方法将句子对应调到最佳点。两个算法之间的差别是 Brown et al. 用词数来计算句子长度，而 Gale & Church 用字符数。

3.2. 基于同构词的算法

所谓同构词是指词形大致相同的词。这在许多西方语言中很常见。如在英语和法语中“信息”都是 information，而“对齐”是 alignment 和 alignement。如果在某一对英语和法语句子中有很多同构词，那么它们对应的可能性就更高。

Simard et al. [18] 和 Chen [6]正是运用这一原理对基于长度的算法进行改进。试验表明这种方法比基于长度的算法要更准确、鲁棒。[18]所用的方法是将两种语言中前四个字母相同的词看作同构词。这使得基于同构词的算法适用于大部分西方语言。

3.3. 基于翻译词的方法

除了使用同构词外，我们还可以使用一本翻译词典。如果两个句子中的词相互有翻译关系，那么它们之间的对应也会很高。所用翻译词典可以是人工建立的[20]，也可以是以前训练的结果[6, 12]。如[12]就是迭代的方法，把上一循环所得的结果确定为可能的翻译用在本循环中作为翻译词。

在[20]的试验中，他发现纯基于长度的方法用在中英文上效果比英法文差很多。这是可以预料的，因为中英文之间有更大的差异。而同构词的概念又不适用于中英文。因此，[20]使用了一本小翻译词典以此确定“对应信号”(lexical cue)。他用的词典只是包含一些常用词和功能词。

3.4. 我们的方法 - 以 HTML 标志为同构词

我们的语料质量相对于人工建立的语料库差很多。所确定的对应网页并不完全对应，翻译质量因网页而异，文本与图形相混合，这都给句子对齐增加很多困难。好在我们的目的并不是希望所建的翻译模型能准确地自动翻译全文，而是为问句确定最合适的翻译词和相关词。因此它对翻译模型的容错性也大一些。

因为我们的平行语料含有 HTML 结构，相对于纯文本，句子对齐也多了一些有用的信息。一般情况下，中英文版网页的 HTML 结构应该是对应的。如果结构对应，那么句子对应的可能性就更高。这和同构词的作用相当。所以，我们可以将网页里的 HTML 标志作为一种特殊的同构词，因而使用已有的基于同构词的算法[18]。相对于与长度为基础的算法，这种算法确实能提高对齐的准确度。这可以从图 2 显示的例子观察。图 2.1 显示的是 Gale & Church 的基于长度的算法的结果。图 2.2 是使用 HTML 标志的方法。我们可以看到，第二种方法能纠正第一种方法的一些错误(如<s id="0003">被“ 强制性地” 相互对齐)。

<s id="0000"> <HTML> <HEAD> <META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=iso-8859-1"> <META HTTP-EQUIV="Content-language" CONTENT="Western"> </s>	<s id="0000"> <HTML> <HEAD> <META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=big5"> <META HTTP-EQUIV="Content-language" CONTENT="zh"> </s>
<s id="0001"> <TITLE>Journal of Primary Education 1996, Vol., No. 1&2, pp. 19-27 </TITLE> </HEAD> </s>	<s id="0001"> <TITLE> Journal of Primary Education 1996, Vol., No. 1&2, Page 19-27 </TITLE> </HEAD> </s>
<s id="0002"> <BODY BACKGROUND="../gif/pejbg.jpg" TEXT="#000000" BGCOLOR="#ffffff"> <CENTER> </s> <s id="0003"> <H1>Journal of Primary Education </H1> </s> <s id="0004"> <HR> <B>Volume 6, No 1&2, pp. 19-27 (May, 1996) </B> <HR> </s>	<s id="0002"> <BODY BACKGROUND="../gif/pejbg.jpg" TEXT="#000000" BGCOLOR="#ffffff"> <A HREF="/en/pej/b2g_pej.phtml?URL=%2fen%2fp ej%2f0601%2f0601019c.htm"> <IMG SRC="/en/gif/kan.gif" ALT="簡体" BORDER=0 ALIGN=R IGHT> </A> <CENTER> </s>
<s id="0005"> <H3>Principles for Redesigning Teacher Education </H3> Alan TOM </CENTER> </s>	<s id="0003"> <H2>初等教育學报</H2> </s> <s id="0004"> <HR>（一九九六年五月） 第六卷. </s>
<s id="0006"> <P> <B> <I> Abstract </I> </B> </s>	<s id="0005"> 第一及二期19-27頁 <HR> </s>

图 2(1) 基于长度的对齐结果

<s id="0000"> <HTML> <HEAD> <META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=iso-8859-1"> <META HTTP-EQUIV="Content-language" CONTENT="Western"> </s>	<s id="0000"> <HTML> <HEAD> <META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=big5"> <META HTTP-EQUIV="Content-language" CONTENT="zh"> </s>
<s id="0001"> <TITLE>Journal of Primary Education 1996, Vol., No. 1&2, pp. 19-27 </TITLE> </HEAD> </s>	<s id="0001"> <TITLE> Journal of Primary Education 1996, Vol., No. 1&2, Page 19-27 </TITLE> </HEAD> </s>
<s id="0002"> <BODY BACKGROUND="../gif/pejbg.jpg" TEXT="#000000" BGCOLOR="#ffffff"> <CENTER> </s>	<s id="0002"> <BODY BACKGROUND="../gif/pejbg.jpg" TEXT="#000000" BGCOLOR="#ffffff"> <A HREF="/en/pej/b2g_pej.phtml?URL=%2fen%2fp ej%2f0601%2f0601019c.htm"> <IMG SRC="/en/gif/kan.gif" ALT="簡体" BORDER=0 ALIGN=R IGHT> </A> <CENTER> </s>
<s id="0003"> <H1>Journal of Primary Education </H1> </s>	<s id="0003"> <H2>初等教育學報</H2> </s>
<s id="0004"> <HR> <B>Volume 6, No 1&2, pp. 19-27 (May, 1996) </B> <HR> </s>	<s id="0004"> <HR> （一九九六年五月） 第六卷. </s> <s id="0005"> 第一及二期19-27頁 <HR> </s>
<s id="0005"> <H3>Principles for Redesigning Teacher Education </H3> Alan TOM </CENTER> </s>	<s id="0006"> <H3>革新教師教育的原則 </H3> Alan TOM </CENTER> </s>
<s id="0006"> <P> <B> <I> Abstract </I> </B> </s>	<s id="0007"> <P> <I> <B> 摘要 </B> </I> <P> </s>

图 2(1) 加了同构词的对齐结果

图 2 基于长度与同构词的对齐结果的比较

4. 中英文预处理 - 中文切词和英语组词

句子对齐后，我们要将句子内的词对应起来 - 这就是训练翻译模型的目的。在此之前，需要确定中英文里可以对应的单元：词或词组。中文就有切词问题，而英语也有将词组合成词组的问题。

4.1. 中文切词

中文切词在过去十几年间有许多研究。如基于词典和规则的方法 [5, 14] 和基于统计的方法 [4, 19]。通常词典只存词形不变的词，如一般的常用词。有一些词不适于存在词典里，如数目。这一类词能更方便的用规则来确定。因此，基于词典的方法通常辅以规则。最常用的方法是最大匹配法。统计方法则是通过计算词串的概率来确定最佳切分的。这通常需要许多切分好的语料以估算字串成词的概率。也要一些方法将上两种方法综合起来，将没有统计数字的词以一定“概率”与统计而来的词相结合 [15].

在这一工作中，我们采用的是全切分的方法，即如果ＡＢＣＤ，ＡＢ，ＣＤ都是词，那么这些词都将从“ＡＢＣＤ”字串中被切分出来。这样做是从信息检索的角度考虑的：如果我们只切出最长的词，训练出来的翻译模型对长词所包含的短词的翻译会乏

力。另外在信息检索中，对问句的翻译并不要求只是翻译词，我们也希望在翻译中有相关的词。而含ＡＢ的问句的翻译也可以包含ＡＢＣＤ的翻译词，因为ＡＢＣＤ是与ＡＢ相关的。反之亦然。因此，将所用的词都切分出来能使所训练的翻译模型找到部分相关词。

在这种情况下，用统计的方法确定最佳切分就意义不大。我们只要使用词典和规则即可。我们使用的词典含词量是 187182（其中不乏词组）。另外还有一组规则以确定数字串，数量词等[15]。

对于英语，情况则相反：英语词已经分开，而我们希望将词组组合起来，如 in order to，World-Wide Web， computer science，等。而同时我们也象对中文一样将构成词(computer, science)留下。

5. 翻译模型的训练

我们训练的模型是 IBM 模型 1 [3]。在[3]中共有三个模型。在模型 1 的训练中词的顺序和上下文都未被考虑。在模型 2 和 3 中这些因素被考虑了。对于机器翻译而言，词序和上下文是必须考虑的。对于跨语言信息检索，翻译时词序不重要，因为翻译好的问句最终要被信息检索系统看成是没有顺序的词集。至于上下文，它对于翻译词的确定很重要，特别是在原语言词歧义的情况下。但同时，考虑上下无会对问句的翻译有一定限制。如果某一原语言词所在的上下文在训练语料中没有遇到过，那翻译模型可能就不能给出它的翻译。同时，在信息检索中我们也希望在翻译中不光有翻译词，而且也有相关词。这样检索系统可以查出更多相关文件。而这一目的可以在一定程度上通过使用模型 1 来实现，因为它允许对应句中任意一对词相互翻译，不论它们出现在什么上下文中。因此，作为我们尝试的第一步，我们只是训练模型 1。在以后的工作中我们将探讨别的模型的使用。

IBM 模型 1 是根据以下原理建立的：给定两组对应的句子，如果一对中英文词经常在对应句子中同时出现，那么它们之间的相互翻译的概率就高。在具体实现时要用到动态规划方法(使用 EM 算法)，将词与词之间的翻译概率调到使用于训练的句对的对应概率最大的点。最终结果是一组概率 p(t|s)。对于一对语言(如中英)，我们可以从同一语料中训练出两个有方向性的翻译模型：中英和英中。

由 PTMiner 自动获得的训练语料共有 117.2M 中文和 136.5M 英文网页。表 1 列出了预处理 (包括去除 HTML 标志)以后训练语料的数据。在这些语料中共有 1,048,156 组对应。其中一对一的对应有 870,414。在我们的训练中只用了一对一的对应，因为它们通常要比一对多的对应准确得多。

	大小	词汇量	用词数

src.e	74.1M	76,969	9,816,859
src.c	51.1M	48,528	9,916,416

表 1. 训练语料的统计数据

6. 测试和分析

测试统计翻译模型有不同方式。Wu [22] 的测试方法将翻译模型看作是一个翻译词典以检验翻译词的准确率。另一种方法是将翻译模型用于某一应用，检验应用的效果。我们将使用这两种方法。我们的应用领域是跨语言信息检索。

6.1. 翻译模型作为词典

较为简单的方法是对每一个词选取前几个概率最高的翻译词看它们是否正确。另一种方法是检测使用翻译模型的困惑性(perplexity)。我们只是使用前一种方法，测试第一个翻译词是否正确。为此，我们分别从中文和英文中随机选取 200 个词用于测试。对于正确翻译是一个单词时，很容易判断正确与否(t 或 f)。当正确翻译词是词组时，我们检验词组中的词是否是翻译概率最大的词。如果在这些词之间加入了别的词，我们仍认为翻译不正确。

以上测试的结果表明，在最好的情况下中英翻译模型的准确率是 77%，而英中为 81.5%。这一质量对于跨语言信息检索可以认为是合理的。这在下面通过第二种试验可以反映出来。

6.2. 无用词表对反映模型的影响

无用词是指在信息检索中对查询文件用处不大的词。通常这是一种语言的功能词。它们不代表有用的语义，因而放在无用词表中不作为索引词。对于在跨语言信息检索中使用的翻译模型来说，可以有两种选择：1) 在翻译时保留无用词，然后由信息检索系统将它们去除；2) 在建立翻译模型以前就将无用词去除。使用第一种方法的问题是，无用词的出现频率通常很高。因此在训练的模型中它们作为翻译词的概率也会很高。如在英语中，of 等词在大部分句子中都有。它们因此而被错误地认为是许多词的翻译词。同样，中文中的确“ 的” 也一样。因此，从直观上看，将无用词从训练语料中去除将有利于提高有用词翻译的质量。这一点可以从试验结果中看出。在表 2 中，我们可以看到将无用词从训练语料中去除后，中英和英中的翻译质量分别由 72.5%和 63%提高到 75%和 79.5%。这在英中翻译中效果尤其明显。

	英语无用词表	中文无用词表	英中	中英
1	不用	不用	63%	72.5%

2	用	用	79.5%	75%
3	不用	用	81.5%	n/a
4	用	不用	n/a	77%

表 2. 无用词表的影响

但提高幅度上这样明显的差异显然与无用词去除的干净程度与否有关。中文无用词表显然要比英语更不完整。在中文中，不少词都有时是无用词，有时是有用词，如“总”(用作“总是”或“总和”)，“所”(用作介词或作为“场所”，“研究所”)。词的词性不如英语中那样固定。这给确定无用词表增加了很多困难。同时也可以肯定，中文无用词的去除不如英语干净。但以上的结果至少可以肯定将无用词去除比不去除要好。

是否应该同时将无用词从目标语和原语言中去除？为了测试这一问题，我们只是从目标语言中去除无用词而在原语中保留无用词。从表 2 中我们可以看到，这样翻译模型的质量反而更好。

怎样解释这一现象？一种可能的解释是：在去除目标语言中的无用词时总会有一些被遗留下来。如果从原语言中把无用词去除，目标语中遗留的无用词可能就找不到对应。它们被迫与有用词对应，因而使这些有用词作为有用词翻译的概率相对减弱。相反，如果在原语言中保留无用词，则目标语言中遗留的无用词就有可能对应到这些无用词，从而使有用词之间的对应关系相对提高。

这一解释与我们观察到的翻译概率相吻合。在图 3(1)和(2)中，第一列是在目标语和原语言中同时去除无用词的结果。第二列是只在目标语中去除。我们可以看到第二列的翻译概率要比第一列高。对一些词(如 census)，第二列能纠正第一列的错误。当然，这两种情况比完全不去除无用词(第三列)明显要好。

words	model1 (with both stoplists)			model2 (with Chinese stoplist)			model4 (with no stoplists)		
a.m.	t	上午	0.201472	t	上午	0.204787	f	的	0.146161
access	f	公开	0.071705	f	公开	0.095658	f	的	0.158034
adaptation	t	适应	0.179633	t	适应	0.185545	t	适应	0.171756
add	t	补充	0.317435	t	补充	0.362725	t	补充	0.289871
adopt	t	采用	0.231637	t	采用	0.271118	t	采用	0.21532
agent	t	代理人	0.224902	t	代理人	0.209154	t	代理人	0.169783
agree	t	同意	0.36569	t	同意	0.40312	t	同意	0.32433
airline	t	航空公司	0.344001	t	航空公司	0.372547	t	航空公司	0.30955
amendment	t	修订	0.367518	t	修订	0.38425	t	修订	0.284755
appliance	t	用具	0.136319	t	用具	0.159941	f	的	0.146441
apply	t	适用	0.19448	t	适用	0.243746	t	适用	0.167856
attendance	t	到席	0.171769	t	到席	0.178672	t	到席	0.158967
auditor	f	审核	0.15011	f	审核	0.180631	f	的	0.146167
average	t	平均	0.467646	t	平均	0.486271	t	平均	0.430765
base_on	f	计算	0.107304	f	计算	0.130322	t	根据	0.187724
block	f	大厦	0.155039	f	大厦	0.17868	f	的	0.148584
bottom	t	最低	0.119597	t	最低	0.127479	f	的	0.146328
break_law	f	候选	0.042876	f	冒险	0.048044	f	的	0.146161
breath	t	呼气	0.349902	t	呼气	0.371555	t	呼气	0.309547
briefing	t	简报	0.225892	t	简报	0.255423	t	简报	0.171755
building	t	建筑物	0.105814	t	建筑物	0.10582	f	的	0.148937
business	t	业务	0.178681	t	业务	0.203324	f	的	0.15169
carry	f	政府	0.046383	f	工程	0.11312	t	进行	0.182572
category	t	类别	0.101924	t	类别	0.115763	f	的	0.146164
census	f	政府	0.190883	t	统计	0.17383	f	的	0.146165

图 3. (1) 英中翻译例子

words	model1 (with both stoplists)			model3 (with English stoplist)			model4 (with no stoplists)		
办事处	t	office	0.375868	t	office	0.395586	t	office	0.364849
保护	t	protection	0.343071	t	protection	0.353903	t	protection	0.322372
报告	t	report	0.358592	t	report	0.395991	t	report	0.355239
备	t	prepare	0.189513	t	prepare	0.240706	t	prepare	0.186544
本地	t	local	0.421837	t	local	0.447433	t	local	0.373569
便会	f	follow	0.023685	f	follow	0.034107	t	will	0.322196
标准	t	standard	0.445453	t	standard	0.46319	t	standard	0.418847
并按	f	adult	0.044959	f	adult	0.05053	f	of	0.072355
不足	t	inadequate	0.093012	t	inadequate	0.116634	f	of	0.086799
部分	t	part	0.313676	t	part	0.38086	t	some	0.24974
财经	t	financial	0.16608	t	financial	0.181867	t	financial	0.167941
参观	t	visit	0.309642	t	visit	0.355927	t	visit	0.29386
单营	t	bill	0.401997	t	bill	0.475395	t	bill	0.444629
车辆	t	vehicle	0.467054	t	vehicle	0.472942	t	vehicle	0.43652
储蓄	t	saving	0.176695	t	saving	0.184575	t	saving	0.15405
处理	t	deal	0.120624	t	handle	0.13537	t	with	0.097049
传真	t	fax	0.400488	t	fax	0.416355	t	fax	0.366898
次序	t	order	0.158448	t	order	0.17405	t	order	0.140455
措施	t	measure	0.409033	t	measure	0.448414	t	measure	0.353883
达到	t	achieve	0.193417	t	achieve	0.235742	t	achieve	0.156993
当局	t	administration	0.429906	t	administration	0.492832	t	administration	0.333284
登记	t	registration	0.226315	t	registration	0.292372	t	registration	0.250755
电子	t	electronic	0.339431	t	electronic	0.354305	t	electronic	0.312737
调	t	adjust	0.072376	t	adjust	0.089118	f	of	0.075687
定	t	determine	0.133398	t	determine	0.201319	f	of	0.111528

图 3. (2) 中英翻译例子

对于只从原语言中去除无用词而在目标语言中保留无用词的情况(表 2 中标为 a/n)我们没有具体测试，因为从直观上看这明显是最坏的情况。从图 4 给出一些词的翻译例子，我们可以看到这种情况远不如完全不要无用词表。这一现象很容易解释：将原语言

中的无用词去除以后，目标语言中的无用词只能与有用词相对应。因此，原语言的有用词在许多情况下都被翻译成无用词(因为它们的高频)。

words	model4 (with no stop lists)			model3 (with Englsih stop list)		
a.m.	f	的	0.146161	f	的	0.185531
access	f	的	0.158034	f	的	0.252953
adaptation	t	适应	0.171756	f	的	0.185532
add	t	补充	0.289871	f	的	0.191875
adopt	t	采用	0.21532	f	的	0.239604
agent	t	代理人	0.169783	f	的	0.223364
agree	t	同意	0.32433	t	同意	0.230816
airline	t	航空公司	0.30955	t	航空公司	0.263289
amendment	t	修订	0.284755	t	修订	0.247788
appliance	f	的	0.146441	f	的	0.191568
apply	t	适用	0.167856	f	的	0.234683
attendance	t	到席	0.158967	f	的	0.185532
auditor	f	的	0.146167	f	的	0.185813
average	t	平均	0.430765	t	平均	0.350076
base_on	t	根据	0.187724	f	的	0.270669
block	f	的	0.148584	f	的	0.214751
bottom	f	的	0.146328	f	的	0.190691
break_law	f	的	0.146161	f	的	0.185532
breath	t	呼气	0.309547	t	呼气	0.256398
briefing	t	简报	0.171755	f	的	0.185952
building	f	的	0.148937	f	的	0.227916
business	f	的	0.15169	f	的	0.220534
carry	t	进行	0.182572	f	的	0.262795
category	f	的	0.146164	f	的	0.185829
census	f	的	0.146165	f	的	0.185669

图 4. 不使用无用词表与只在原语言上使用无用词表的比较

6.3. 错误分析

模型的翻译错误出于多种因素。首先训练语料的质量对翻译模型有很大影响。除此之外还有以下几个因素：

1) 有些原语言的词不是翻译成目标语言中的一个词，而是几个词或一个词组。例如：

新版→ new version

由于我们测试方法的限定，如果这一组翻译词被其它词分开则被认为不正确。但实际上，翻译模型所给的翻译仍有一部分是正确的。这一点没有在试验数据中反映出来。

对于这一问题最好的解决方法是将词组合在一起考虑。如将“ new version”看作是一个元素。这一点我们已经在预处理中通过对中英文词组的组合在一定程度上作了一些处理，但处理的词组很有限。我们今后的工作将对这一问题进一步处理。

2) 有些词的正确翻译被放入了无用词表。如我们将“ 年”作为无用词。这样，“ year”的翻译就成为“ 今年”，“ 每年”，等。

3) 翻译模型受训练语料的领域限制。在我们的语料中有不少是香港议会的发言。因而，模型给出下列翻译：

mr →议员

miss →议员

house →内务

6.4. 用于跨语言信息检索

翻译模型可以与一个单语信息检索系统相结合实现跨语言信息检索。在这里的试验中我们使用稍加改进的 SMART 系统 [1]。这是一个使用矢量检索模型的系统。我们将翻译模型所提供的前 N 个翻译词输给 SMART 以建立问句矢量。

为了确定 N，我们做了一组试验，发现在 N 为原问句词数的 1.5 倍时效果较好。在翻译概率的使用与否上，我们也观测到一些区别：不使用概率时效果稍好。另外，翻译模型所提供的翻译词也可以和某一翻译词典提供的词相结合，以更好地覆盖原词的翻译。为此，我们使用了网上可以得到的一个小词典 [8]。

在下面的表 3 中我们列出所得到的结果。这一组结果是在 TREC [10] 试验的英语和中文测试集上得到的。表 4 给出一些测试集的统计数据。

	中英跨语言检索	英中跨语言检索
单语检索	0.3861	0.3976
翻译模型	0.1504 (39.0%)	0.1841 (46.3%)
词典	0.1530 (39.6%)	0.1427 (35.9%)
翻译模型 + 词典	0.2583 (66.9%)	0.2232 (56.1%)

表 3. 跨语言信息检索结果

	中文测试集	英语测试集
文件数	164 789	113 005
体积	170 M	550 M
问句数	54	21

表 4. 测试集统计数据

我们所得到的最好的结果是中英跨语言检索是英语单语检索的 66.9%，而英中跨语言检索是中文单语检索的 56.1%。这一结果比我们以前在英法跨语言检索上得到的结果 [16] 差很多 - 英法跨语言检索能达到单语检索的 80-90%。但这一差别主要是语言之间的差别造成的。使用中英机器翻译系统也同样与英法有很大的差别。我们使用了万维网上能得到的一个机器翻译系统 [17]，所得到的英中跨语言检索也只有单语检索的 50.3%。尽管我们没有使用中英机器翻译系统进行比较，在目前所能得到的报告中，使用机器翻译系统在同样的试验集上最高也只能达到单语检索的 70%[13]。这些结果基本

相对于使用翻译模型与翻译词典相结合的方法。因此我们仍然能说以翻译模型为基础的方法能大致达到与机器翻译系统相当的水平。

7. 总结

这篇文章介绍了我们从网上自动获取平行网页、以此为基础训练统计翻译模型及将翻译模型用于跨语言信息检索的尝试。这一尝试表明在现有的万维网中有大量的平行语料，可以通过自动搜索获取。获取的平行网页可以用来训练统计翻译模型。而我们在跨语言信息检索上的试验说明这样的模型是非常有用的。对于一些还没有机器翻译系统的语言，这一结果尤为重要。我们可以由此实现有效的跨语言信息检索系统。对于已有机器翻译系统的语言，统计翻译模型仍可以作为一种补充。

本文中用到的翻译模型是 IBM 模型 1。在大部分情况下这对跨语言信息检索是合理的，但在翻译歧义性很大的问句时就有不足。在这种情况下，我们希望翻译模型所给的翻译词与上下文有关。以此在以后的工作中，我们将尝试使用 IBM 模型 2 和 3。

另外，在统计模型中加入语法特征也有可能对翻译质量有所改进。如，parallel 在作为形容词和名词时可以分别翻译成“平行”和“平行线”。如果翻译模型根据词性来确定她的翻译，有可能使翻译质量更好。

对于讨论中提到的一些问题，如词组的问题，我们也将进一步加以改进。值得一提的是中英文之间人名地名等专有名词和未登陆词的翻译。这在相似语言间不是很大的问题，它们的翻译很多时候是它们本身。但在中英文这样不相似的语言之间这是一个重要的而且困难的问题。一种可能的方法是从平行语料中识别专有名词，然后通过统计将它们对应。但这一方法只对被训练语料充份覆盖的专有名词有效。对于其它专有名词或未登陆词还需要找出别的方法。

最后，对于中文的信息检索还有值得探讨的地方。例如应以什么方法切词？在跨语言信息检索中是否应该同时用 n-gram？然后将 n-gram 与词相结合？等等。我们希望这篇文章能引出对这些问题更深入的研究。

参考资料

[1]Buckley, C. (1985). *Implementation of the SMART information retrieval system*, Technical report, #85-686, Cornell University.

[2] P.F. Brown, J.C. Lai, and R.L. Mercer. Aligning sentences in parallel corpora. In 29th Annual Meeting of the Association for Computational Linguistics, pages 89-94, Berkeley, Calif., 1991.

[3] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of machine translation: Parameter estimation. Computational Linguistics, 19:263-311, 1993.

[4] J. S. Chang and al. Chinese word segmentation through constraint satisfaction and statistical optimization. In ROCLING 4, pages 147-165, 1991.

[5] K. J. Chen and S. H. Kiu. Word identification for Mandarin Chinese sentences. In 5th International Conference on Computational Linguistics, pages 101-107, 1992.

[6] S. F. Chen. Aligning sentences in bilingual corpora using lexical information. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pages 9-16, Columbus, Ohio, 1993.

[7] Jiang Chen, Jian-Yun Nie. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. Proc. of the 6th Applied Natural Language Processing Conference, pp. 21-28, Seattle, 2000.

[8] Paul Denisowski. Cedict (Chinese-English dictionary) project. http://www.mindspring.com/~paul_denisowski/cedict.html, 1999.

[9] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pages 177-184, Berkeley, Calif., 1991.

[10] D. K. Harman and E. M. Voorhees (eds.), *Text REtrieval Conference (TREC-6).* Gaithersburg, 1997.

[11] P. Isabelle, G. Foster, and P. Plamondon. SILC: a System for Language and coding identification. http://www-rali.iro.umontreal.ca/ProjetSILC.en.html, 1997.

[12] M. Kay and M. Röscheisen. Text-translation alignment. Computational Linguistics, 19:121-142, 1993.

[13] Kwok, K. L. English-Chinese cross-language retrieval based on a translation package. *Conference on Research and Development in Information Retrieval, ACM-SIGIR*, 1999.

[14] N. Y. Liang and Y. B. Zhen. A Chinese word segmentation model and a Chinese word segmentation system PC-CWSS. In COLIPS'91, volume 1, pages 51-55, 1991.

[15] Jian-Yun Nie, Wanying Jin, and M. L. Hannan. A hybrid approach to unknown word detection and segmentation of Chinese. In International Conference on Chinese Computing, pages 326-335, Singapore, 1994.

[16] Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining parallel texts from the Web. In ACM SIGIR'99, pages 74-81, August 1999.

[17] http://www.readworld.com/translate.htm, 1999

[18] Michel Simard, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In Proceedings of TMI-92, Montreal, Quebec, 1992.

[19] R. Sproat and C. Shih. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, 4(4):336-351, 1991.

[20] Dekai Wu. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In ACL-94: 32nd Annual Meeting of the Assoc. of Computational Linguistics, pages 80-87, Las Cruces, NM, June 1994.

[22] Dekai Wu. Large-scale automatic extraction of an English-Chinese lexicon. Machine Translation, 9(3-4): 285-313, 1995.