面向短语的词语对齐方法

田亮　黄辉　周沁

澳门大学　自然语言处理与中葡机器翻译实验室　澳门

E-mail：{ma96572, derekfw, lidiasc}@umac.mo

摘 要： 自动词语对齐技术在统计机器翻译领域中起了很大的作用。然而，GIZA++训练得出的对齐结果并不是很令人满意。本文提出了一种基于最大匹配法（MMM）和 GIZA++的词语对齐方法。首先，我们使用了最大匹配法分别把平行的英文和中文句子划分成单词和短语，然后通过词典和 GIZA++的共同限制来产生对齐结果。实验表明，尤其是当平行句子中包含短语的时候，我们提出的对齐方法得出的对齐结果要比 GIZA++产生的结果好的多。

关键字： 自动词语对齐、统计机器翻译、最大匹配法、词典、GIZA++

Phrase Oriented Word Alignment Method

Liang Tian　Fai Wong　Sam Chao

Natural Language Processing

Portuguese-Chinese Machine Translation Lab, University of Macau, Macau

E-mail：{ ma96572, derekfw, lidiasc}@umac.mo

Abstract: *Automatic word alignment plays a very important role in statistical machine translation research area. However, the alignment result generated by GIZA++ is not satisfied. In this paper, an alignment method based on Maximum Matching Method (MMM) and GIZA++ is proposed. Firstly, the words and phrases of parallel English and Chinese sentences are detected based on Maximum Matching Method (MMM), and then candidate alignment results are gotten by the constraint of both a dictionary and GIZA++ result. Empirical study demonstrates that the proposed method gives a better alignment result than that of the GIZA++, especially for parallel sentences that have phrases.*

Keywords: *Automatic word alignment, statistical machine translation, maximum matching method, dictionary, GIZA++*

1　引言

词语对齐是自然语言处理领域的一个基本的问题，许多基于双语语料库的应用（如统计机器翻译（SMT）、基于实例的机器翻译（EBMT）、词义消歧（WSD）、词典编撰等）都需要词汇级别的对齐。一般来讲，对齐有篇章(section)、段落(paragraph)、句子(sentence)、短语(phrase)、词语(word)等不同级别的对齐，其目的就是从双语互译的文本中找出互译的片段[邓丹，2004]。其中篇章、段落、句子的对齐技术主要用于语料库的整理，而短语和词语对齐，就是要找出相互翻译的文本中对应的词与词、词与短语、短语和短语之间的相互翻译对。现今的基于短语的统计机器翻译系统中，很大一部分程度依赖于词语对齐（word alignment）[Och et al.，2000；Yarowsky et al.，2000]，词语对齐对统计机器翻译中的短语抽取起到了很大的作用。现在使用最多的词语对齐方法就是使用双语语料库来抽取词语对齐

[Smadja et al., 1996; Melamed, 2000]，其中典型的对齐软件就是 GIZA++ [Och, 2000; Och et al., 2003]。GIZA++实现了 IBM 公司提出的 5 个模型[Brown et al., 1993]和隐马尔科夫模型（HMM）[Och et al., 2003]，其主要思想是利用 EM 算法对双语语料库进行迭代训练，由句子对齐得到词语对齐。表 1 是从 GIZA++对齐文件中取出的一个稍加改进的例子。其中 x 是目标语言句子、y 是源语言句子、a 是对齐结果，比如“3-2”的意思就是说中文句子的第二个单词“在”对齐到英文的第四个单词“in”（英文句子从 0 开始标注）。

表 1. 词语对齐示例

x	I_0 am_1 $studying_2$ in_3 the_4 $university_5$ of_6 $Macau_7$ $._8$
y	NULL({4}) 我 ({0}) 在({3}) 澳门大学 ({5 6 7}) 读书 ({1 2}) 。({8})
a	4-0 0-1 3-2 5-3 6-3 7-3 1-4 2-4 8-5

自从 IBM 公司提出了五个模型后[Brown et al., 1993]，许多科研工作者都对如何提高词语对齐的工作做了深入的研究。Gale 使用了互信息和 χ^2 检验来进行双语对齐[Gale, 1991]。Fung 引进了同义词辞典，使用了 K-vec 方法来获取词语对齐，获得了较高的覆盖率和准确率[Fung, 1994]。Och 引进了对数模型[Och et al., 2002]，并把 IBM 的五个统计模型和隐马尔科夫模型（HMM）进行了结合[Och et al., 2003]，得出了比 IBM 模型高的对齐质量。刘洋等也是用对数线性模型并加入了一些句法信息，使得对齐的效果有了明显的提高[Liu et al., 2005]。其他的一些研究人员也是采用各种方法来提高对齐的准确率，比如 Tiedemann 引进了线索(clue)进行对齐[Tiedemann, 2003]、Fraser 提出了一个繁殖模型(generative model)允许多对多的对齐[Fraser et al., 2007]。我们的方法来源于基于短语的统计机器翻译[Koehn et al., 2003]，该方法把要翻译的句子分成一个个短语，然后把这些短语依次进行翻译，最后根据重组模型得出最终的结果。图 1[1]是一个基于短语的统计机器翻译的示例。当然，这里的短语指的是连续的语言片段，而我们将要使用的短语指的是语言学意义上的短语，比如“am good at”、“a lot of”、“澳门”、“中国人”等这些词语。

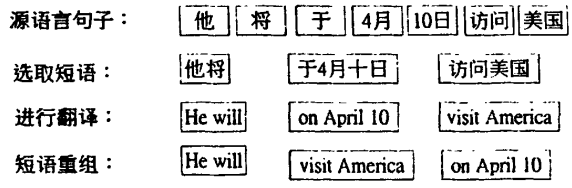

图 1. 基于短语的统计机器翻译示例

词语对齐工具 GIZA++由于不依赖于具体的语言对而在统计机器翻译领域中得到了广泛使用。为了便于在 Windows 下使用 GIZA++，我们曾经使用 Cygwin 把 GIZA++编译成可执行文件，然后通过 Visual Studio 2008 调用的方法把其移植到了 Windows 环境中使用[Tian et al., 2011]。通过实验的观察，GIZA++对于短语的处理，比如“be able to”、“in addition to”、“plenty of”等，效果就不是特别好。考虑到 GIZA++有一定的对齐准确度，我们决定在其基础上选择一些新的算法来达到更高质量的对齐结果。根据算法，首先我们使用了最大匹配法把中文和英文的互译对划分成单词和短语，然后使用词典来进行互译对的初步对齐，其次对剩下没有对齐的单词或者短语根据相似度再次根据词典进行对齐匹配，最后把剩下的单词或者词组根据事先使用 GIZA++训练好的单词互译表（GIZA++对齐后生成的一个

[1] 出自中科院、厦门大学、哈尔滨工业大学的“基于短语的统计机器翻译系统“丝路”1.0 版设计与使用说明”

叫做~actual.ti.final 文件）来进行查询，如果经过这一步后仍然不能对齐的单词则对齐到空（NULL）。最终我们选取了 500 个句子，根据一些评测标准来跟 10 万句的中英语料库（包括选取的 500 句）在 GIZA++训练后的结果进行比较，最终得到了比较满意的对齐结果。

本文的如下内容的编排格式如下：在第二部分我们将给出词语对齐和短语抽取之间的关系，通过该部分可以了解词语对齐在统计机器翻译的重要作用；在第三节中我们将给出我们的面向短语的词语对齐方法；在接下来的第四节中我们将给出我们的实验结果和分析；在最后我们将给出结论和以后的工作。

2 词语对齐和短语抽取的关系

在基于短语的统计机器翻译过程中，很大程度依赖于一个叫做短语表（phrase table）的东西，而短语表的构建则需要词语对齐的帮助。在本节中我们将给出词语对齐和统计机器翻译中的短语抽取（phrase extraction）之间的关系，从中我们可以了解到词语对齐对短语表的构建起到了很大的作用，所以我们有必要对词语对齐进行研究。

在这里我们以统计机器翻译系统 Moses[Koehn et al.，2007]、词语对齐工具 GIZA++为例来说明其之间的关系。词语对齐工具 GIZA++通过 EM 算法对给定的双语语料库进行双向对齐，通过交集（intersection）和并集（union）的操作最终得到较好的词语对齐结果。通过以下两个条件，统计机器翻译系统利用短语对齐的结果进行短语的抽取[Galbrun，2009]：

（1）分别从源语言句子 f 中和目标语言句子 e 中抽取连续的单词序列 f' 和 e'，并且单词序列的长度不能超过 k 个单词。

（2）连续的单词序列 f' 和 e' 的对齐信息 a' 要由源语言和目标语言的对齐信息 a 构建而来，其中 a' 至少要在 a 中包含一个连接。

表 2. 利用 GIZA++对齐结果进行短语抽取示例（能够<->am able to）

词语对齐	我 能够 做 好 它 。 I am able to do it well .
短语表	(我 ||| I); (我 能够 ||| I am able to); (我 能够 做 ||| I am able to do); (我 能够 做 好 它 ||| I am able to do it well); (我 能够 做 好 它 。 ||| I am able to do it well .); (能够 ||| am able to); (能够 做 ||| am able to do); (能够 做 好 它 ||| am able to do it well); (能够 做 好 它 。 ||| am able to do it well .); (做 ||| do); (做 好 它 ||| do it well); (做 好 它 。 ||| do it well .); (好 ||| well); (好 它 ||| it well); (好 它 。 ||| it well .); (它 ||| it); (它 。 ||| it .); (。 ||| .)

表 2 给出了一个利用 GIZA++中英对齐结果进行短语抽取的例子。从该例子中我们可以看出，一个好的对齐结果对短语的抽取的质量好坏有很大的影响。如果“能够”仅仅对齐到“able”，其它对齐不变（如表 3 所示），根据短语抽取的两个条件，我们仍然能够得到“(我 能够 ||| I am able to)”，但是却得不到“(能够 ||| am able to)”的对齐结果。换句话来讲，如果我们日后在翻译的过程中，遇到的是“I am able to”，我们能够得到“我 能够”的翻译结果，但是如果只是遇到“am able to”的话，我们就得不到“能够”的翻译结果（如图 2 所示）。

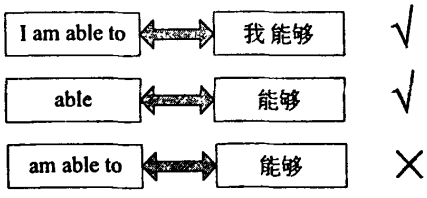

图 2. 单词对齐翻译示例

表 3. 利用 GIZA++对齐结果进行短语抽取示例（能够<->able）

词语对齐	我 能够 做 好 它 。 I am able to do it well .
短语表 （部分结果）	(我 ||| I); **(我 能够 ||| I am able);** (我 能够 ||| I am able to); (我 能够 做 ||| I am able to do); (我 能够 做 好 它 ||| I am able to do it well); (我 能够 做 好 它 。 ||| I am able to do it well .); **(能够 ||| able);** **(能够 做 ||| able to do);** **(能够 做 好 它 ||| able to do it well);** **(能够 做 好 它 。 ||| able to do it well .);** (做 ||| do); (做 好 它 ||| do it well); (做 好 它 。 ||| do it well .); (好 ||| well); (好 它 ||| it well); (好 它 。 ||| it well .); (它 ||| it); (它 。 ||| it .); (。 ||| .)

从以上的比较中我们可以知道，好的词语对齐质量对高质量的短语抽取起到了很大的作用。虽然 Fraser 曾经提出了词语对齐错误率（AER）的降低并不会明显的提高机器翻译的质量[Fraser et al.，2007]，可是我们从观察实验中得知，这与短语抽取的算法有关，这里我们不加以论述。举个例子来说，示例中的短语抽取的 k 取值为 7，我们的短语表就会有“(我

能够 做 好 它 ||| I am able to do it well)" 的对齐结果。那么如果我要翻译 "我能够做好它" 这句话，就一定能得到正确的翻译结果，因为这句话就在短语表中。即使把"好"对齐到"it"，"它"对齐到"well"，根据短语抽取的算法，我们仍然可以得到"(我 能够 做 好 它 ||| I am able to do it well)" 这个短语项。这句话仍然可以正确翻译，但是如要翻译单个单词 "它" 或者 "好" 可就翻译不出正确的结果来了。所以好的对齐质量对统计机器翻译的结果还是有影响的，至少在现有的统计机器翻译的框架下，词语对齐起了很大的作用，所以我们觉得有必要对词语对齐质量加以改进和提高。

3 基于短语的词语对齐

通过上一节，我们了解到词语对齐对统计机器翻译的短语抽取起到了很大的作用。在这一节中，在给出我们使用的词语对齐方法之前，首先给出现今的几种对齐的方法，然后简单的介绍一下我们用来中英文分组的最大匹配法，最后给出我们的对齐方法。

3.1 词语对齐的方法

对于词语对齐，已有的方法主要有四种：基于字符的方法、基于统计的方法、基于语言学的方法、多种方法的混合[邓丹，2004]。下面简单的介绍一下几种方法的主要思想和优缺点。

1. 基于字符的方法。该方法以两种语言含有的同源词在词形上面的共同之处进行词语对齐[Church，1993]。因为同源词的字符串有很多相似的字符组成，通过对互译文本进行字符上的对齐路径搜索，就得到了匹配的词语。比如英语句子中含有一个单词 "government"，然后通过 chart-align 方法从互译对法语中找到"gouvernement"这个互译的单词作为对齐结果，如图 3 所示。其中按照 Church 的说法,这里要使用一个误差(residual)作为判断的标准，其公式如下:

$$f(x)-cx \qquad (1)$$

其中，x 是英文单词在英文中的位置，$f(x)$是对应的法语单词在法语句子中的位置，c 是两个英、法文件长度的比率，Church 使用了 0.91。

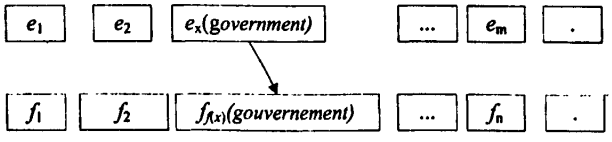

图 3. 基于字符的对齐方法示例

基于字符的对齐方法主要适用于含有许多同源词的欧洲语言，而对于使用不同字符编码的中英、日英等语言对则效果不佳。

2. 基于统计的方法。统计方法通过对大规模双语语料库的统计训练，获得双语对译词的同现概率，以此作为对齐的基础，图 4 展示了基于统计方法的一般过程。Brown 等实现了基于统计机器翻译模型的词对齐[Brown et al.， 1993]，Gale 使用互信息和 χ^2 检验对齐双语词汇[Gale，1991]。Fung 采用 K-vec 方法进行词语对齐[Fung，1994]。这种方法是现今对齐的主流，可以不用依赖具体的语言对，通过大规模的语料库就可以自动对齐。但是这种方法对于低频词的把握不好，容易出现数据稀疏问题，另外对齐的训练过程较长，比较消耗内存。

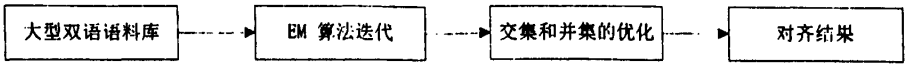

图 4. 基于统计方法的词语对齐一般过程

3. 基于语言学知识的方法。该方法以双语词典和同义词词典等语言学知识作为词对齐的基础。Ker 和王斌等根据语义类实现词对齐[Ker, 1997；王斌，1999]。Ker 的方法据说是克服了基于词的词语对齐方法覆盖率不高的问题，在 416 句训练库之外的测试集上，取得了准确率 90.0%下，覆盖率 88.2%的英汉词语对齐结果。Huang 根据 Ker 的方法进行了基于类的中一朝词语对齐试验，但试验结果远没有 Ker 的好[Huang, 2000]。我们的方法也引进了词典，使用了相似度计算，公式如下：

$$similarity(s_1, s_2) = \frac{2s}{|s_1| + |s_2|} \qquad (2)$$

其中，s 是字符串 s_1 和 s_2 的相同单词的个数，$|s_1|$ 和 $|s_2|$ 分别是其长度大小。

利用双语词典的方法，可以快速的找到精确的词义对齐，但是双语词典覆盖面有限，无法应付真实文本中灵活的翻译现象。另外双语词典中的多个义项会造成跟译文句子中多个词匹配的歧异，如果我们想利用这种方法就得适当的改进这种不足。

4. 各种方法的综合。Tiedemann 提出了基于线索的词语对齐方法[Tiedemann，2003]。Ker[Ker，1997]和 Huang[Huang，2000]虽然以语言学知识为词语对齐的着手点，但实际上用了统计的方法。Huang 也使用了基于字符的方法。

综合的方法，我们认为更适合实际的情况。人类在学习的过程中，不断的由单词积累成句子，最后到了篇章，对于不认识的单词，虽然我们可以进行适当的推理，但是也不是百分百的正确。我们不能要求计算机完全模仿人类的过程，但是有些东西还是可以去借鉴的。对于机器识别的词典就相当于人类的学习单词的过程，对于未登录词则相当于我们暂时还不认识的单词，对这些词我们可以使用统计等方法来进行“推理”，达到一个初步的对齐学习过程。所以我们觉得综合的方法，只要算法合理就可以达到一个好的对齐效果。

3.2 最大匹配法

所谓最大匹配法就是尽可能的用最长的词来匹配句子中的字符串[陈小荷，1999]。一般来讲最大匹配法分为正向和逆向匹配法两种，正向匹配法采用从句子的左边到右边的处理顺序，逆向匹配法采用从右到左的方式。在实践中，我们把英文的分组采用正向匹配法，而中文的分词则采用逆向匹配法和概率的结合。

表 4. 英文词典示例

英文单词	单词释义
I	我
am able to	能 能够
plenty of	大量 很多 许多
university of macau	澳门大学 澳大

使用最大匹配法需要一个词表，并且规定最大字符串的长度“*MaxWordLength*”。我们的词典的构建来自于《现代汉语词典》和网络上总结的成语词典[1]。其中我们的中文词典只包含两个单词以上的词语或者成语，最长不超过七个单词。表 4 和表 5 给出了一个英文和中文词典的示例。有了词典后，我们可以构建初步的分组算法，以下是最大匹配法的简单算法过程[陈小荷，1999]，图 5 给出了算法流程图[2]。

（1）待切分的字符串 s_1，已经切分分组的字符串 s_2；

（2）如果是 s_1 是空串，则转到（6）；

[1] http://chengyu.itlearner.com/

[2] http://www.52nlp.cn/maximum-matching-method-of-chinese-word-segmentation

（3）从 s_1 的左边复制一个字符串 w，长度不超过 *MaxWordLength*；

（4）如果在词典中找到这个子字符串 w 或者 w 最后是一个单个单词，那么把 w 和一个分隔符（可以是空格或者是自己需要的符号）放到 s_2 中；

（5）去掉 w 中右边的一个单词，继续转到（4）进行处理。

（6）分组结束。

表 5. 中文词典示例

中文单词
我们
跑龙套
量力而行

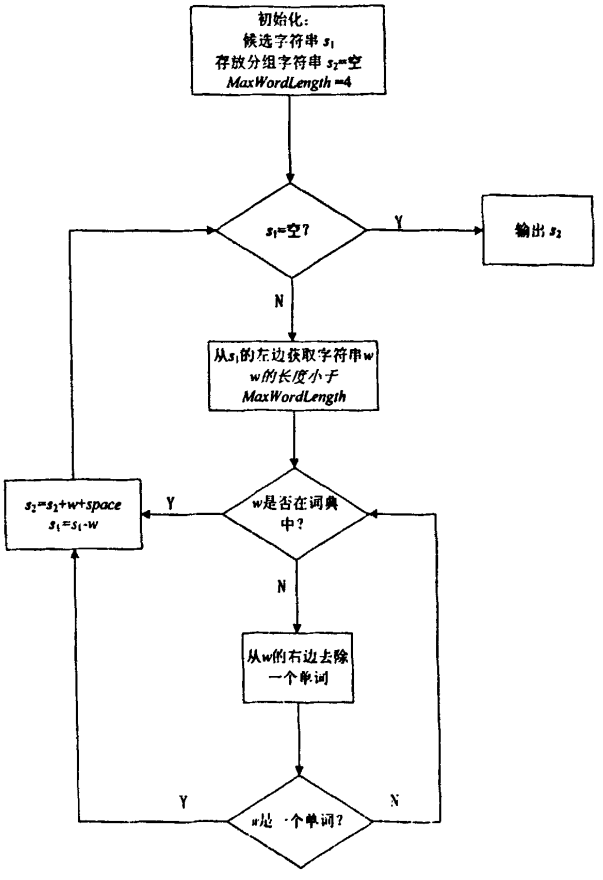

图 5. 最大匹配法算法流程图

举个简单的例子来说明我们将要采用的最大匹配法的用途。比如我们想把“I am able to do it well.”进行分组，那么我们最终想得到的结果就是“I”、“am able to”、“do”、“it”、“well”、“.”这六个单词或者词组，然后就可以进行我们的对齐操作。假设我们仅仅拿“I am able to”为例，表 6 给出了具体的最大匹配法的过程。

中文的分组实际上也是中文的分词问题，我们本来是使用的中科院的中文分词程序，可

是在对齐过程中我们发现如果使用默认的分词程序达不到我们对齐的要求，所以最后我们采用了逆向最大匹配法加上概率统计的知识进行了中文的分组。比如我们要对“我们从马上下来”和“We get down from the horse”这个双语对进行对齐。首先我们根据一般的最大匹配法得到一个初步结果“我们 从 马上 下来 ”，得到这个结果仅仅使用了一个如表 5 所示的词典，但是这个对齐结果就不好进行词语的对齐。接下来我们要进行校正，因为我们使用了中英双语，所以最终我们结合英文句子来进行最终的中文词语划分。我们查询到的“we”的翻译中含有“我们”的释义，所以我们认为中文中的分词“我们”就是正确的，但是到“马上”的时候，就出问题了。英文中“horse”对应的翻译是“马”，所以最终我们认为这个“马上”的组合不正确，要进行拆分。最终经过这种思路我们得到了“我们 从 马 上 下来 ”的分词结果。在双语对齐的时候我们可以采用英文翻译校正的方法，换句话来讲，该分词方法采用了英语译文的信息来得到中文的分词结果。

表 6. 最大匹配法示例

初始值： s_1= “I am able to”，s_2= “ ” *MaxWordLength*=4
（1）w= “I am able to”，查表无 w，w 变为 “I am able”； （2）w= “I am able”，查表无 w，w 变为 “I am”； （3）w= “I am”，查表无 w，w 变为 “I”； （4）w= “I”，查表存在，则 s_1= “am able to”；s_2= “I ”； （5）w= “am able to”，查表在词典中，则 s_2= “I am able to” （6）结束
结束：这样就把 s_1 分成两组 “I” 和 “am able to”。

通过最大匹配法的处理，我们分别得到了英文和中文中的单词和词组，基本上得到了可能的对齐对，接下来我们就可以使用一些方法对潜在的对齐对进行对齐了。

3.3 词语对齐方法

最大匹配法的引入使得我们得到了中文和英文句子的一个个词语分组，接下来我们要进行这些分组的对齐。在实验中我们观察到中英（将要研究的中葡）的结构特点，很大一部分情况下一个中文词语可以对应多个英文单词。基于这个特点，我们采用了单向对齐的思路，就是把英文单词对齐到中文单词上，我们根据 GIZA++对齐的要求，这里一个或者多个英文单词可以对齐到一个中文词语上去，反过来不行。

我们的对齐思路总的来说是分为三个过程，下面给出了算法的过程：

（1）第一个过程是使用最大匹配法进行中英词语的分组。其中英文的分组过程中要同时查询到每个单词或者短语对应的译文解释，保存成<e_i，t_i>的格式（e_i 是英文单词或者短语，t_i 是对应的译文，i 不超过分组后单词和短语的总个数）。中文的分组仅含有对应的单词或者词组，保存格式形如<c_j>，其中 j 不超过中文分组后的单词和词组的个数总和。

（2）第二个过程简单的查询匹配。这一步是根据建立的词典来查询中文单词 c_j 是否在英文句子中的英文单词的解释义项 t_i 中，如果在的话那就直接找到这个对齐对< e_i，c_j>。

（3）第三个过程是两次相似度计算的过程。其中第一次相似度的计算在构建的词典中进行。对于不在英文单词或短语义项中的中文单词，可以根据相似度的匹配来找到可能的对齐对。比如“我能够做好它”这句话中的“好”可能不在英文单词“well”的解释义项（好的；好地；水井）中，但是其中含有“好地”，我们就可以通过相似度的计算认为这个“好”和“好地”可能对应一个英文单词“well”。第二次相似度的计算来自于 GIZA++训练产生的一个短语对齐概率表（一个名叫~actual.ti.final 的文件）。我们把中文句子中经过第一次相似度计算仍然未对齐单词进行再次相似度的计算。如果经过这一步还有剩下的英文单词，我们就把它

（们）对齐到空（NULL）。图 6 给出了整个算法的流程图。

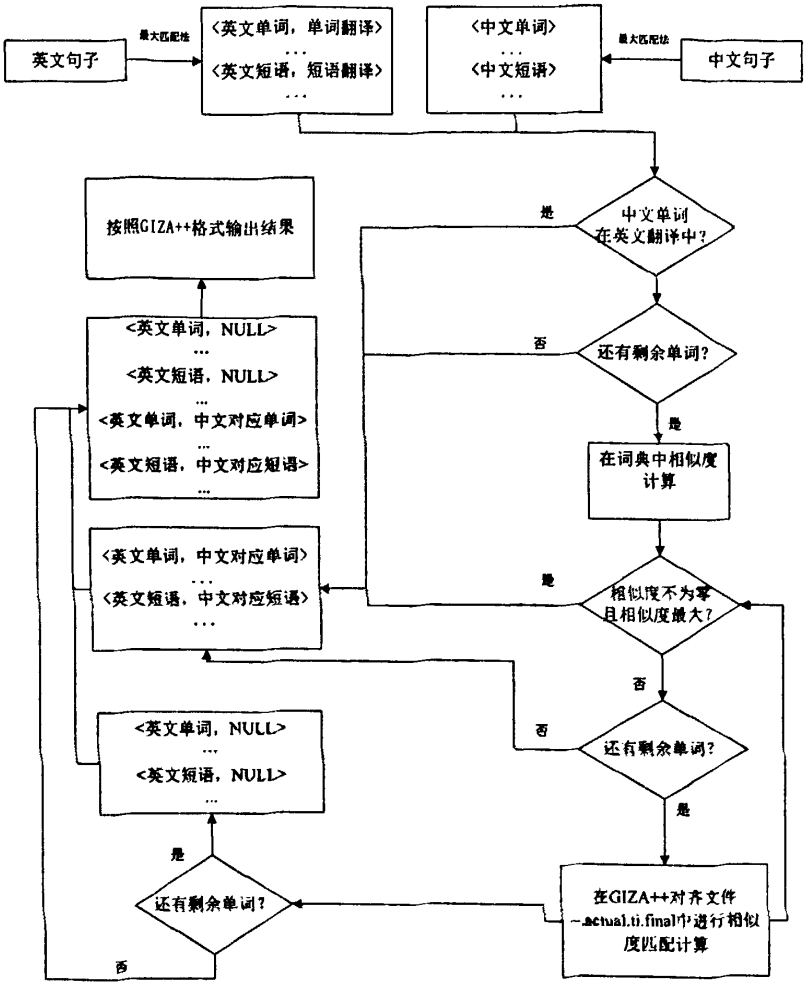

图 6. 基于短语的词语对齐流程图

为了更加明白我们的思路，下面给出一个例子。假设我们存在以下两个词典（表 7 和表 8），其中表 7 是自己构建的词典，表 8 是 GIZA++通过双语语料库训练生成的词语对齐概率表。经过最大匹配法处理后的中英文句子“我 能够 做 好 它 。”和“I am able to do it .”是我们的处理对齐的对象。那么具体的处理过程如图 7 所示。

表 7. 基本词典

英文单词	中文释义
I	我
am able to	能
do	做
it	它
well	水井
.	。

表 8. GIZA++自动生成的词语对齐概率表

英文单词	中文对齐单词	对齐概率
well	好的	0.564526
well	好地	0.996908
well	水井	0.085487

（1）从词典中找出具有中文单词释义的英文单词或者短语，优先对齐。

首先，我们可以根据词典把四个单词进行对齐：
<I，我>
<do，做>
<it，它>
<.，。>

（2）对剩下单词从词典中进行相似度计算，找到最大的一个，最为可能的对齐。

其次剩下的短语是be able to和单词well，把他们的英文释义和中文单词进行相似度计算：
其中得出：similarity（能，能够）=2*1/（1+2）=2/3，其它为0
又得出一条对齐对：<be able to，能够>

（3）从GIZA++单词对齐表中进行相似度计算，找出可能的对齐对。

现在剩下的单词是well，把GIZA++生成的单词表中释义和中文单词计算相似度：
最后经计算得：similarity（好的，好）=2*1/（1+2）=2/3
Similarity（好地，好）=2*1/（1+2）=2/3
Similarity（水井，好）=2*0/（1+2）=0
这里我们可以确定，well可以和中文单词“好”对齐。

（4）如果还有单词，可能是不翻译的单词或者词典中没有此项，把该单词对齐到空。

已经没有单词可以对齐，结束该句子对齐，得到如下GIZA++格式对其结果：
#sentence pair (1) source length (6) target length (8)
I am able to do it well .
NULL ({}) 我 ({ 1 }) 能够 ({ 2 3 4 }) 做 ({ 5 }) 好 ({ 7 }) 它 ({ 6 }) 。 ({ 8 })

图 7. 基于短语的词语对齐的方法例子

3.4 异常句子处理

对于绝大部分句子可以采用上述的方法进行对齐，可是对于一些特殊的句子，我们就需要特殊处理。这里仅给出对于句子中含有两个以上的相同单词或词组的处理情况（如表 9）。对于这种情况，如果单纯的使用词性来处理，我们可以把“例…”区分开来，可是处理不了后面的句子，因为他们好多都是同一词性。最终我们选择了以下的处理方法：

（1）对于含有两个以上的相同数字对齐的情况，一般来讲这些词语按照顺序依次对齐到同一个中文单词上。我们可以简单的把英文单词前面的对齐到中文句子前面的单词，后面的单词对齐到后面的单词。

（2）对于含有重复数字的对齐信息，但是数字不相同的，我们认为是短语对齐错误。比如例三的句子中的（6 5 7 8）和（5 7），这里含有重复的对齐信息（5 7）。根据对齐规律，对齐后的数字要连续递增，那么（6 5 7 8）中的 5 小于前面的数字，去除 5 后将构成连续递增序列，所以符合对齐规律。我们最终就把 5 给去掉，得到了正确的对齐结果。

（3）根据单词前后的结合概率来进行判断。这种方法要求我们通过语料库进行训练得出类似于语言模型的信息，然后根据概率判断结合的概率，选择概率最大的作为结果。

表 9. 句中含有两个以上相同单词的对齐处理（数字代表英文单词的位置）

例一	I book the book . NULL({ }) 我 ({ 1 }) 预订 ({ 2 4 }) 那本 ({ 3 }) 书 ({ 2 4 }) 。 ({ 5 })
例二	He said he would come . NULL({ }) 他 ({ 1 3 }) 说 ({ 2 }) 他 ({ 1 3 }) 要 ({ 4 }) 来 ({ 5 }) 。 ({ 6 })
例三	I am a student of University of Macau . NULL({ }) 我 ({ 1 }) 是 ({ 2 }) 澳门大学 ({ 6 5 7 8 }) 的 ({ 5 7 }) 一名 ({ 3 }) 学生 ({ 4 }) 。 ({ 9 })

实验结果和分析

为了检验我们想法的可行性，我们开发了词语对齐和评测的系统。最终我们从 10 万句的中英语料库[Tian et al., 2010]中抽取了 500 句作为评测数据（英语句子平均长度约 21 个单词，中文句子平均长度约为 23 个单词）。首先，我们把这些句子进行人工对齐，格式采用 GIZA++标准。然后，使用我们的系统得出了 500 句的对齐的结果，其次使用包含这 500 句的 10 万句的中英语料库，通过 GIZA++训练得出了对齐的结果，最后把这两个对齐结果通过我们开发的系统进行评测，得到了如表 10 和表 11 的评测结果。评测过程中，我们使用了 GIZA++格式的文本作为标准（如表 12 所示），并且选择了最后一行具有对齐结果的句子作为评测的对象。在开发评测系统的过程中，我们把精确度（precision）、召回率（recall）、F 权重（F-measure）和词语对齐错误率（AER)作为评测词语对齐质量的标准[Koehn, 2010; Och et al., 2003]。记待评测对齐的结果集合为 A，其中把人工对齐的结果标记为两类集合，确定性对齐集合 S（Sure links）和不确定性对齐 P（Possible links），其中使用的公式如下所示：

$$precision = \frac{|A \cap P|}{|A|} \tag{3}$$

$$recall = \frac{|A \cap S|}{S} \tag{4}$$

$$F - Measure = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \tag{5}$$

$$AER = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{6}$$

表 10. GIZA++500 句对齐结果

Evaluation Result	
Test File	Giza++.txt
Reference File	Reference.txt
Precision	0.44754
Recall	0.45048
F-Measure	0.44900
AER	0.57124

表 11. 基于短语分组的 500 句对齐结果

Evaluation Result	
Test File	ownAlignment.txt
Reference File	Reference.txt
Precision	0.87389
Recall	0.89041
F-Measure	0.88208
AER	0.13933

从结果中我们可以看到，面向短语的词语对齐效果要比 GIZA++的训练结果有了很大提高。表 13 给出了几个句子的对齐结果。在这几个例子中，前两个例子是对于一个句子中含有相同词语的对齐情况。例 3 是含有短语的对齐结果，例 4 是含有介词短语的情况。通过实验结果的分析得知，对于 GIZA++没有能很好的处理的短语对齐，我们的方法得到了显著的改善。

表 12. GIZA++测评格式

行数	句子内容
1	# Sentence pair (263) source length 7 target length 7 alignment score : 7.00841e-06
2	He came here two years ago .
3	NULL({ }) 他 ({ 1 }) 两 ({ 2 }) 年 ({ 5 }) 前 ({ 4 }) 来到 ({ 6 }) 这里 ({ 3 }) 。 ({ 7 })

表 13. 面向短语的对齐结果和 GIZA++结果对比

1	GIZA++	He said he would come . NULL({ }) 他 ({ 1 }) 说 ({ 2 }) 他 ({ }) 要 ({ 4 }) 来 ({ 3 5 }) 。 ({ 6 })
	短语分组	He said he would come . NULL({ }) 他 ({ 1 }) 说 ({ 2 }) 他 ({ 3 }) 要 ({ 4 }) 来 ({ 5 }) 。 ({ 6 })
2	GIZA++	I am a student of university of Macau . NULL ({ 9 }) 我 ({ 1 }) 是 ({ }) 澳大 ({ 6 8 }) 的 ({ }) 一名 ({ 2 }) 学生 ({ 3 4 5 7 }) 。 ({ })
	短语分组	I am a student of university of Macau . NULL ({ }) 我 ({ 1 }) 是 ({ 2 }) 澳大 ({ 6 7 8 }) 的 ({ 5 }) 一名 ({ 3 }) 学生 ({ 4 }) 。 ({ 9 })
3	GIZA++	The customers are accustomed to the disgusting custom . NULL ({ 9 }) 顾客们 ({ 1 }) 习惯了 ({ 2 }) 令人讨厌的 ({ 4 6 8 }) 风俗 ({ 3 5 7 }) 。 ({ })
	短语分组	The customers are accustomed to the disgusting custom . NULL ({ 1 6 }) 顾客们 ({ 2 }) 习惯了 ({ 3 4 5 }) 令人讨厌的 ({ 7 }) 风俗 ({ 8 }) 。 ({ 9 })
4	GIZA++	The dust in the industrial zone frustrated the industrious man . NULL ({ 8 11 }) 工业区里的 ({ 1 }) 灰尘 ({ 2 5 }) 使 ({ 4 }) 勤勉的 ({ 3 6 }) 人 ({ 10 }) 灰心 ({ 7 9 }) 。 ({ })
	短语分组	The dust in the industrial zone frustrated the industrious man . NULL ({ }) 工业区里的 ({ 3 4 5 6 }) 灰尘 ({ 2 }) 使 ({ 7 }) 勤勉的 ({ 9 }) 人 ({ 10 }) 灰心 ({ 7 }) 。 ({ 11 })

这里有两点我们要给予说明，第一点是我们尽可能的根据英文单词的释义来进行中文分词，这样可以保证一个或者多个英文单词对齐到一个中文词组上。这样我们就会把形容词的“的”字，一起划归到中心词上，比如例 3 中的“令人讨厌的”就会作为一个整体直接对齐到英文单词“disgusting”上。此外，英文的短语也会作为整体对齐到中文词组上。第二点是考虑到日后的翻译，我们把介词短语作为一个整体来对齐，比如表 13 中的最后一个例子中的“in the industrial zone”就对齐到了“工业区里的”。

此外根据实验，我们在 Intel Core i5 CPU 2.8GHz，内存是 2G 的普通机器上的 Linux 环境中，在 GIZA++运行 38 万句的中英语料库（香港法律文本加自建的语料库，平均单词 26），总共使用了大约 14 个小时，也就是说每句大约需要运行 0.13 秒。最终我们开发的系统，根据句子长度的不同，每句需要 0.04-0.50 秒，基本上可以满足日常研究的需要。

结论和工作展望

本文我们首先通过展示词语对齐和统计机器翻译中短语抽取之间的关系，揭示了词语对齐在统计机器翻译系统中的重要作用，然后通过几种对齐方法和最大匹配法的介绍引出了我们的面向短语的词语对齐方法。最终我们开发了基于此算法的对齐系统，得出了较 GIZA++好的对齐结果。

本文提出的词语对齐方法，关键在于两个地方，一个就是要使用最大匹配法把中英文句子中的单词和短语进行抽取。另外，我们使用了英文译文来校验中文分词的正确性的方法。另一个就是使用相似度计算的方法给出了可能的对齐结果。在这种方法中我们把处理的目标集中在要处理的单个句子上，而不是整个语料库中的句子。这样既便于处理每个句中的单词，又提高了系统运行的效率。系统中的词典的作用也很明显，首先我们从大型语料库中提取出其中的单词和词组，然后调用谷歌翻译[1]进行翻译，最终得出我们需要的词典格式。

面向短语分组的词语对齐方法，也可以很方便的适用在英-日、葡-汉等类似的语言对中。对于能否适用在英-葡、英-法等这样的语言对中，我们还没有验证。另外，中、英句子千差万别，有很多特殊的句子我们还没有考虑到，可能在实践的过程中会出现新的问题。另外中文的分词问题我们还没有考虑人名的问题，这也给词语的对齐造成了一定的影响。日后的工作中，我们既要继续完善相应的不足，也要开始着手进行把我们的对齐结果应用到统计机器翻译系统中。

致谢

实验室的机器翻译工作得到了澳门大学科学研究委员会（项目标号：UL019/09-Y2/EEE/LYP01/FST）和澳门科学技术发展基金（项目编号：057/2009/A2）的支持，特此表示感谢！

参考文献

Brown, P., Della Pietra S., Della Pietra V., Mercer R. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2), pp. 263–311.

陈小荷. 1999. 现代汉语自动分析--Visual C++实现. 北京语言文化大学出版社.

Cherry, Colin 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.

Church K.W. 1993. Char align: A program for aligning parallel texts at the character level[A]. Proceedings of the

[1] http://translate.google.cn/#

31st Annual Meeting of the Association for Computational Linguistics[C]. Columbus, Ohio,1993. 1-8.

邓丹. 2004. 汉英词语对齐技术研究. 硕士学位论文. 中国科学院研究生院.

Fraser, Alexander and Daniel Marcu.2007. Getting the structure right for word alignment: LEAF. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.55-60, Prague, June.

Fung P., Church K.W. 1994. K-vec: A new approach for aligning parallel texts. Proceedings of the 15th International Conference on Computational Linguistics (Coling-94)[C], Kyoto, 1994. 1096-1102.

Gale, W. and Church, K. 1991. Identifying Word Correspondences in Parallel Texts. Proceedings of the 4th DARPA Speech and Natural Language Workshop[C], Pacific Grove, CA, 1991. 152-157.

Galbrun, Esther. 2009. Phrase table pruning for Statistical Machine Translation. Series of Publications C. Report C-2009-22.

Huang, Jin-Xia, Key-Sun Choi. 2000. Chinese-Korean word alignment based on linguistic comparison. In Annual Meeting of the Association for Computational Linguistics. 392-399.

Ker, Sue J. and Jason S. Chang. Align more words with high precision for small bilingual corpora[J]. Computational Linguistics and Chinese Language Processing, 1997, 2(2):63-96.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

Koehn, Philipp, Franz Josef Och, Daniel Marcu. 2003. Statistical phrase-based translation, In Proc. of HLT-NAACL, pp. 127–133.

Koehn, Philipp. 2010. Statistical machine translation. New York. Cambridge University Press. 2010.

Liu, Yang, Qun Liu and Shouxun Lin. 2005. Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, June 2005. c2005 Association for Computational Linguistics, pages 459–466.

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

Och, Franz J. 2000. Giza++: Training of statistical translation models.

Och, Franz J. Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In: Proceedings of the 18th Int. Conf. on Computational Linguistics. Saarbrucken, Germany, pp. 1086–1090.

Och, Franz J, Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, volume 29, number 1, pp. 19-51.

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

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

Tian, Liang, Fai Wong, Sam Chao. 2010. An Improvement of Translation Quality With Adding key-words in Parallel Corpus. Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao.pages:1273-1278.

Tian, Liang, Fai Wong, Sam Chao. 2011. Word Alignment Using GIZA ++ on Windows. Proceedings of Thirteenth MT Summit, Xiamen, China.

王斌. 1999. 汉英双语语料库自动对齐研究.中国科学院计算技术研究所博士论文. 1999.

Yarowsky, David and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL), pages 207–216, Hong Kong.