基于英汉双语语料库的词汇自动对齐实验系统

李 竹
（语言文字应用研究所）

摘 要：本文首先讨论了英汉互译文本中词汇对译的几种特殊情况，分析了《理想国》中英汉词汇对应的几个具体例子，进而介绍了一个基于英汉双语语料库的词汇自动对齐实验，并描述了该实验系统的具体实现过程。

Corpus-based automatic English-Chinese word alignment system

LI ZHU
Section of Computational Linguistics
Institute of Applied Linguistics, State Language Commission

abstract: In this paper we first discussed several special cases about word translation in English and Chinese text, and analyzed some examples in <The Republic>. Finally we introduced an experiment on corpus-based word alignment in detail.

§1.引言

随着计算语言学研究的不断深入，它的研究手段也在不断变化。对语料库的加工利用，已成为今天计算语言学研究的一个重要领域。双语语料库的研究为机器翻译、双语词典、术语库的建立提供了有力的支持。一般说来，为了方便地从语料库中获取知识，就要对语料进行不同层次的加工处理，以汉语为例，对汉语语料的加工包括分词、词性标注、短语标注、句子结构的分析等等。对一个双语语料库而言，除对其中每一种语言的语料进行分级加工以外，还要作双语的对齐，如双语间段落级的对齐、句子级的对齐以及词汇短语级的对齐。

英汉两种语料间的对齐情况比较复杂。我们对柏拉图《理想国》英汉译本第一卷作了手工统计，其中段落间的对齐除了一对一的情况外，还存在着一对二、二对一及多对多的情况，句子间一对一的情况仅占约三分之二，词汇间的对应更加复杂。到目前为止，人们对双语间句子的对齐研究得多一些。双语间句子的对齐主要有两种方法，1)基于统计的方法，这种方法主要依靠两种语言间的句子长度关系；2)基于词汇的方法，这种方法要寻找两种语言间的同源词和关键词，显然它们对于英汉两种语言是不适用的。总之，由于英汉两种语言间较大的差异，以及对此差异研究的缺乏，英汉双语语料库的对齐还存在着许多问题。这里我们要

介绍的是一个在段落和句子已经手工对齐了的基础上寻找双语语料库中对应词汇的实验系统。双语间词汇的对应具有如下的意义： 1)词汇的互译常可代替语义的标注； 2)双语间词汇的对应，是建立机器词典的基础； 3)大规模语料中词汇的对应是词典编纂的依据； 4)为语言间的对比研究提供第一手资料。

§ 2.英汉两种语言间词汇对应的几种特殊情况

§ 2.1 英汉词汇对译中的几种情况

英汉两种语言间词汇的对应关系极其复杂，特别是在具体的语言环境中，词汇的翻译与语境关系密切。与英汉两种语言间词汇对译相关的问题，大致有如下六个方面，这六个方面给词汇的自动对应造成了困难。

1)词性的转换

英汉两种语言间互译时，常发生词性的转换。为了符合翻译中目标语的语言习惯，避免翻译的痕迹，翻译时常要改变词性，如英文中的名词在翻译成中文时，有可能要变为动词，形容词有可能变为名词或副词，介词有可能变为动词。

2)词语的习惯搭配及修辞

由于英汉两种语言的词的搭配习惯不同，因此英汉词语间静态的译词（如词典中的翻译）在译文中往往不能生搬硬套过来，这就造成了在实际的译文中词无定译的情况。如

〖例 2 － 1 〗
to fire questions at him
像连珠炮似的向他问道

3)释义

等意的一个英汉句子对中，有时会有这样的情况发生，如英语的一个词在汉语中找不到一个确切的词与它相对应，这时就需要释义。这往往造成词与短语或更大的成分相对应的情况。

4)同义反译

即表达同一个意思，但在不同的语言中却用了意义相反的词。例:

〖例 2 － 2 〗
Please tender exact fare
恕不找钱

〖例 2 － 3 〗
Keep in lane
不准换线

5)添词

添词在译文中很常见，添词的目的是为了补足语气、承接上下文或为了避免译文意义含混。添词有时候添的不是一两个词，而是一个句子。

6)减词

译文有时没有把原文的每一个词都翻译过来，没有译出来的那些词往往是原文中次要的部分，这样作的目的是为了使原文的主体部分更加鲜明突出，译文的文字也更加流畅。

§ 2.2 柏拉图《理想国》英汉译本中的两个具体例子

柏拉图的《理想国》是世界名著，不同于一般的科技文献，特别是它的英汉译本都是由第三种语言－－希腊语翻译而来的，译文语言精辟质量高，很少有翻译的痕迹，因此更能显示出英汉两种语言间词汇对应研究的难点和重点。为此我们对《理想国》的中文译本作了词汇的切分，又对第一卷作了段落级和语句级的手工对齐工作。下面用两个具体的例子说明在《理想国》中等意的英汉句子对间词汇对应的问题。例中斜体部分为相互对应的等意句子对。

〖例 2 － 4 〗

I might answer them as Themistocles answered the Seriphian who was abusing him and saying that he was famous, not for his own merits but because he was an Athenian: "If you had been a native of my country or I of yours, neither of us would have been famous."

我可以回答他们，象色弥斯托克勒回答塞里福斯人一样。塞里福斯人诽谤色弥斯托克勒，说他的成名并不是由于他自己的功绩，而是由于他是雅典人。你知道他是这样回答的："如果我是塞里福斯人，我固然不会成名，但是，要让你是雅典人，你也成不了名。"

上面这一例子中，专有名词“雅典人”与短语“a native of my country”相对应。

〖例 2 － 5 〗

But the company would not let him;they insisted that he should remain and defend his position ;and I myself added my own humble request that he would not leave us .

但是在座的都不答应，要他留下来为他的主张辩护。我自己也恳求他。

上例中名词“request”与动词“恳求”相对应；名词“company”与的字结构“在座的”相对应。

§ 2.3 科技文献中的词语对应

科技文献的词语对应相对而言要容易一些，它有着较为固定的结构和词语的搭配，没有那么多文学性的修辞。术语是科技文献的一大特点。在英汉科技文献中存在着大量的术语，其中的汉语术语大都是从英语翻译而来的，术语的对齐比较容易做到。要对齐术语首先要把无论是英语文本还是汉语文本中相邻的可以构成术语的词（包括符号、字母等）拼合起来，然后再在文本之间寻找对应。由于术语中存在着还不够标准化、规范化的问题，英汉术语中一对多或多对一的情况也时有发生。

科技文献中也常常会有一词多义的现象。例如，在计算机领域中 control 一词即为多义词，而且还是一个名动兼类词，它可以译为汉语的控制、管理、措施、技术或控制程序、控制系统、控制权等等。总之，要把它的不同译法考虑充分，以便给对齐提供有利条件。

§ 3.基于英汉双语语料库的词汇自动对齐实验

§ 3.1 问题的描述

我们用 S_{Ei} 和 S_{Cj} 分别表示英汉双语语料库中的英语句子和汉语句子，其中 i 和 j 表示句子在语料中的序号；用 W_E 和 W_C 分别表示英语词汇和汉语词汇；用“≡”表示“等意于”。

设　$S_E=S_{EI}S_{EI+1}...S_{Em}$,　$m \geqslant i \geqslant 1$

$S_C=S_{Cj}S_{Cj+1}...S_{Cn}$,　$n \geqslant j \geqslant 1$

$\Sigma E = \{W_E \mid W_E$ 为 S_E 中的实词、词组、或短语 $\}$

$\Sigma C=\{W_C \mid W_C$ 为 S_C 中的实词、词组或短语$\}$

当　$S_E \equiv S_C$　时

若　$W_E \in \Sigma E$

$W_C \in \Sigma C$

且　$W_E \equiv W_C$

则我们说 W_E 与 W_C 对齐了。

这里介绍的自动对齐程序就是要在每一个句子对$\{S_E,S_C\}$上寻找相互对应的词汇$\{W_E,W_C\}$。其中的 W_E 和 W_C 都包括词组和短语，是因为有时句子对中的词是和词组或短语相对应的。

§ 3.2 系统中词典的构造

为实现词语的自动对齐，系统要用到两部词典和一个临时词表。词典 1 收录英文词汇及其对应的汉语词语，以及相关信息。词典 1 的结构为：

E_{word}	L_e	L_{c1}	C_1	L_{c2}	C_2

其中 E_{word} 表示英语词汇；L_e 是数字型的，当 $L_e > 1$ 时，表示可能有一个英文的词组或短语是以 E_{word} 开头的；C_{word1} 和 C_{word2} 是汉语的词、词组或短语；L_{c1} 和 L_{c2} 分别表示 C_{word1} 和 C_{word2} 长度。

词典 2 收录英文词组和短语及其对应的汉语词语，以及相关信息。词典 2 的结构为：

E_{phrase}	L_e	L_{c1}	C_1	L_{c2}	C_2

其中 E_{phrase} 表示英语词组或短语；L_e 是数字型的，其个位表示 E_{phrase} 所包含的词数，其十位表示可能匹配不上的词数；其他符号同词典 1 。

系统在运行过程中还生成一个临时词表，其结构为：

n_1	n_2	L_{c1}	C_1	L_{c2}	C_2

其中 n_1 和 n_2 表示英文词、词组或短语在英文句子中的起始位置；其他符号同上。

§ 3.3 系统的实现

词汇自动对齐系统通过调用不同的功能模块，实现英汉词语的自动对应。

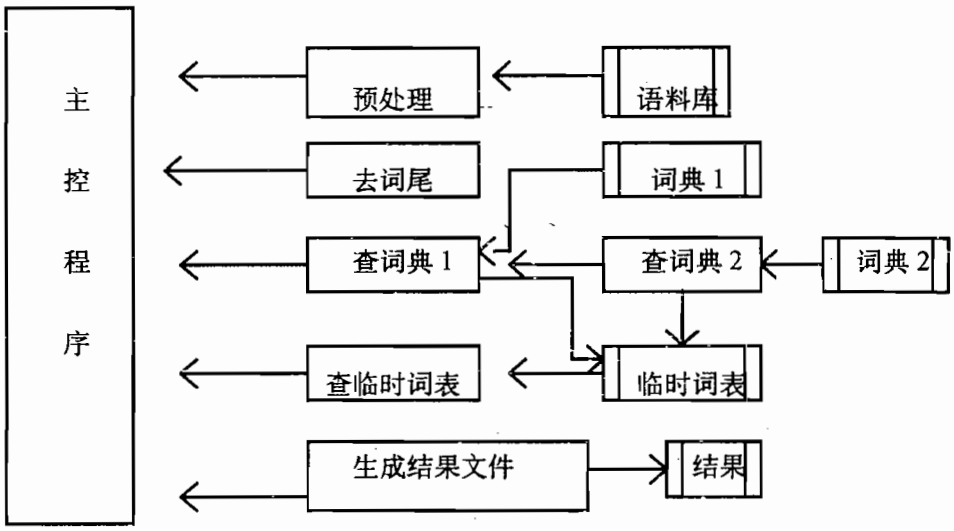
预处理模块，从双语语料库中以词为单位读入一个等意的英汉句子对。

词典 1 的查询模块，依次用从预处理模块中获得的英语词汇，查询词典 1，如查到则生成临时词表。若词典 1 中相应的 $L_e > 1$，则调用词典 2 的查询模块。

词典 2 的查询模块，用英文词 E_{word} 查询词典 2，确定句子对中是否有以 E_{word} 开头的词组或短语，如查到则生成临时词表。

结果文件生成模块，依次用从预处理模块获得的汉语词，查询临时词表，生成英汉词语对照表。

词语自动对齐系统的总体框图如下:



下面给出一个系统运行的具体例子:

〖例 3 - 1〗　预处理程序从语料库中读入句子对(1)(2)后生成结果文件:

(1) 家奴 从 后面 拉住 我 的 披风 说:"玻勒马霍斯 请 您们 稍微 等一下",\
The servant took hold of me by the cloak behind, and said, Polemarchus desires you to wait.\

(2) 我 转过身来 问 他:主人 在 哪儿?\
I turned round, and asked him where his master was.\

结果文件:

servant 家奴
behind 从后面
take hold of 拉住
I 我

cloak 披风
say 说
Polemarchus 玻勒马霍斯
desire 请
you 您们
wait 等一下
I 我
ask 问
he 他
he 他
master 主人
where 在哪儿

§ 4.结语

目前双语语料库中词语级的对齐工作，还是一个难题，没有什么理论、著作、论文可以依据。本文介绍的自动对齐系统只是一个小的实验系统，由于时间等原因，还有许多工作没有作。系统的实现效率也有待进一步提高。

我们以已对齐的句子为基础，作对齐实验，是为了使问题简单化，再者，如能统计出，在某个领域内一个已对齐的句子对中，实词相互对应的百分比，则可以反过来为句子的对齐服务。总之双语语料库中词汇的对齐在理论和应用方面都有很大的价值，今后我们要在这一方向上继续努力。

参考文献

[1] 陈力为、袁琦，计算语言学进展与应用，清华大学出版社出版，1995
[2] 冯志伟，自然语言机器翻译新论，语文出版社，1994
[3] 黄昌宁、夏莹，语言信息处理专论，清华大学出版社，1996
[4] 黄邦杰、译艺谭，中国对外翻译出版公司，1995
[5] 俞士汶、朱学锋，计算语言学文集，1996
[6] 连淑能，英汉对比研究，高等教育出版社，1993
[7] 罗振声、袁毓林，计算机时代的汉语和汉字，清华大学出版社，1996
[8] 红牛，译名研究，计算机世界，1997