# 万维网上的双语文本的获取及对齐

许玉祥

yuxiang.xu@uni-konstanz.de

CiTaL, FH-Konstanz Germany及同济大学计算机系

**摘 要:** 对齐的双语文本在计算语言学中有着广泛的应用. 本文介绍一种在万维网上自动获取双语文本并将获得的结果进行对齐的方法.

# Fetching and Aligning Parallel Texts from the Web

Xu Yuxiang

CiTal, FH-Konstanz, Germany

and

Department of Computer Science & Engineering

Tongji University

**ABSTRACT** : In many research areas of computational linguistic aligned parallel text is a necessary resource. In this paper we address a method to fetch and align parallel Texts from the World Wide Web.

## 一、引言

对齐的双语文本（Aligned Parallel Texts）在计算语言学中有着广泛的应用. 一个熟知的例子是基于类比的机器翻译（Machine Translation by Analogy Principle）〔Sato 1990〕. 这种机器翻译方法要求维护一个含有对齐的双语文本的语料库. 另一个例子是 Brown 等〔Brown 1990〕的完全自动的基于统计的机器翻译. 他们的方法以一个大规模的原文与译文对齐的双语语料库作为出发点. 对齐的双语语料库也可以用于技术和法律文本的译文的一致性检验, 以及用于构造双语词典系统.

自从1988年 Kay, M 和 Röscheisen, M〔Kay 1988〕提出双语文本自动对齐的课题以来, 已有许多学者在这一领域里进行了工作. 在印欧语系的语言对之间的工作有〔Gale 1993〕〔Kay 1993〕; 在亚洲语言对之间的研究有〔Tan 1995〕; 汉语和英语之间的对齐研究有〔Liu 1995〕〔Xu 1996〕.

双语文本对齐可以在段落一级(Paragraph Level)上进行，也可以在语句一级上(Sentence level)上进行，进一步地还可以是子句级(Clause level)、短语级(Phrase Level)和词语级(Word Level)的对齐. 段落对齐是语句对齐的基础；语句对齐又是进一步的子句、短语及词语级对齐的基础. 本文所介绍的方法只涉及段落和语句级的对齐.

## 二、双语文本的获取方法

文本自动对齐的前提是程序能获得机器可读的双语文本. 如果双语文本仅仅是印刷品，则必须通过手工或扫描仪输入. 显然用这种方法要获取大规模的双语文本是非常费时的. 目前也有一些机构提供含有双语文本的CD-ROM，如欧洲计算语言学协会EACL(http://www.elsnet.org/resources/eciCorpus.html)就提供大规模的单语、双语及多种语言的语料库.

双语文本甚至多语言文本的另一个有效来源是Internet上的万维网(World Wide Web). 非英语国家的万维网用户在网上发布信息时，为方便其他国家的用户阅读，除了提供一个用自己的母语写的页面外，还常常提供一个英语版的页面. 软件公司也常常将他们的产品介绍及文档资料在万维网上用双语提供给用户. 英语版的内容与母语版的内容有很好的对应关系.

在任何一个万维网搜索软件(Web Search Engine)下用关键词"English Version"或"English"或"English Interface"进行搜索便可得到许多双语万维网页面的指针(Link). 根据这些Link便可阅读双语的万维网页面. 在大多数的万维网阅览器上用View选项中的Source或Document Source选项可以看到万维网页面的原始HTML(HyperText Mark-up Language)文本. 用Netscape〔Ernst 1995〕的保存文件功能可将这些页面的HTML文档存储到盘上. 其他的万维网阅读器(Browser)如Hotjava〔December 1995〕也有类似的功能. 在将原始HTML文本存储到自己的盘上前当然应该征得作者同意.

## 三、双语万维网页面的的组织形式

双语万维网页面有两种组织形式. 我们把它们分别称为并列双语万维网页面和顺序双语万维网页面. 并列双语万维网页面把同一内容的两种语言的版本同时显示在一个页面上. 并列双语万维网页面只对应一个HTML文本. 顺序双语万维网页面把同一内容的两种语言的版本作为两个页面处理，在一个页面上只显示一种语言，但给出另一个版本的页面的指针. 顺序双语万维网页面对应两个HTML文本. 限于篇幅，本文仅介绍顺序双语万维网页面的对齐. 并列双语万维网页面的处理比顺序双语万维网页面的处理要简单.

| | |
|---|---|
| ADDRESS | H1...H6 |
| BLOCKQUOTE | HEAD |
| BODY | HR |
| BR | LH |
| CAPTION | NOTE |
| CITE | OL |
| DIV | P |
| DL | PRE |
| FN | TABLE |
| FORM | TITLE |
| | UL |

表一　HTML 3.0 的块级元素

```
BEGIN
        Parapraph_is_not_empty ::=FALSE
        WHILE ( NOT EOF of  SHF)
        BEGIN
                Segment ::= Get_a_Segment
                IF ( Segment NOT IN Block-level Elements )
                THEN    BEGIN
                                Put  Segment IN PF;
                                Parapraph_is_not_empty ::=TRUE
                        END
                ELSE    IF Parapraph_is_not_empty
                        BEGIN
                                Put PDL in PF
                                Parapraph_is_not_empty ::=FALSE
                        END
        END
        IF Parapraph_is_not_empty
        THEN
                Put PDL in PF
END
```

图一　段落对齐算法

# 四、HTML文本的特点

如上所述，可在Internet上的万维网中获取双语HTML文本．由于HTML本身所具有的特点，用这种方式获得的文本比用其他方式获得的文本更适合于作自动文本对齐的原始语料．

HTML是一种显示格式控制语言〔Raggett 1996〕．在万维网上可以找到HTML的联机资料（http://union.ncsa.uiuc.edu/HyperNews/get/www/html/guides.html）．

HTML文本本质上是一个普通的ASCII文件．其中除了正文外还有用于控制文本显示的标记（Tags）和指向其他万维网页面或指向同一万维网页面中其他位置的指针（Link）．标记和指针都放在尖括号中．标记通常成对出现．以下特点使得双语HTML文本比其他双语文本更易于作文本对齐的处理：

1. HTML文本是一种自动文本流（Automatic Textstream）．回车换行（Return）对HTML文本的显示不起作用．万维网浏览器会把回车换行用一个空格（Blank）替换掉．连续的多个空格相当于一个空格．就结构而言，可以把HTML文本看成由两种元素构成：块级元素（Block-level Elements）和字符级元素（Character-level Elements）．把块级元素插入到文本中导致当前段落的结束．字符级元素则对段落不产生影响．正文本身属于字符级元素．粗略地说，两个块级元素之间如果存在字符级元素的话可以把它们看成为一个段落．本文以下所说的段落如不加说明均指两个块级元素之间的内容．表一给出了HTML 3.0的块级元素．

2. HTML文本中含有匹配支点（Matching Ankers）．所谓匹配支点是指一个句子中翻译方法比较固定的部分，例如地名和数字的译文．借助于匹配支点可以很好地提高文本对齐的正确率〔XU 1996〕．一个段落中除正文外还可以有其他用于控制字体、字号等的字符级元素．这些字符级元素在两个不同语言的版本中出现的顺序和位置通常是严格对应的，因而可以把它们看作匹配支点．

# 五、段落对齐

由于HTML文本本身的特点，将两个语言版本的HTML原始文件中的段落依次抽取出来便可以完成段落的对齐．对两个原始HTML文本文件依次执行下面的算法便可以得到两个段落相互对齐的文本文件．段落对齐算法见图一．算法中用到以下符号：

SHF（Source HTML File）：是一个原始HTML文本文件，是算法的输入．

PF（Paragraph File）：是算法的输出，它也是一个文件．在这个文件中段落之间存有段落分隔标志PDL（Paragraph Delimter）．

| | |
|---|---|
| TAG: | HTML文本中一对尖括号之间的内容是一个标记. |
| Block-level Elements: | 表三中的标记属于块级元素. |
| Segment: | 标记是一个Segment, 两个标记之间的内容也是一个Segment. |

## 六、语句对齐

Hypertext 是一种非线性文本, 一篇文章通常分散在若干个长度较短的HTML 文本中. 另一方面由于我们定义的段落指得仅仅是两个块级元素之间的内容, 大多数段落包含的语句数目都很少. 在此前提下我们可以用简单的算法进行语句的对齐. 此外, 我们所获取的对齐的双语句子将用于一个试验性的基于实例的机器翻译系统, 必须保证100% 的正确率. 为此我们在语句对齐时采用最简单的只考虑1:1匹配的方法, 通过人工的校对剔除那些错误的语句对.

语句对齐的过程在只考虑1:1匹配的情况下简化为一个在两个不同语言版本的段落中顺序抽取语句的过程. 语句的分隔符是“. ！？”. 但要注意句号“.”可能出现在缩写记号后和用作数字中的小数点. 在抽取语句的过程中还要去除HTML文本中的字符级标记.

## 七、结束语

本文给出了一种在万维网上自动抽取双语文本并将获得的结果进行对齐的方法. 由于语言表达的差异, 把一种语言的文本翻译为另一种语言的文本时严格的一一对应是很难做到的. 要找到一种在任何情况下都能精确对齐的算法是不现实的. 另一方面, 用手工的方法不可能作大规模的收集工作. 构造一个计算机辅助的双语文本对齐系统是一个较好的折中. 本文的方法可以很好地集成到这样的计算机辅助系统中.

本文是在德国Konstanz, CiTaL进修期间完成的. 借此机会向Professor Dr. W. Thomassen, W. Mallow, A. Maurer, D. Auer, Frau Lutz 及其他同事表示衷心的感谢.

## 参考文献

(Brown 1990) Brown, P; Cocke, J; et al., **A statistical approach to machine translation** *Computational Linguistics* 16(1), 22-29

(December 1995) John December **Presenting Java** *SAMS,Macmillan Computer Publishing, USA.*

(Ernst 1995) Warren Ernst. **Using Netscape** *QUE Corporation, USA* 1995 pp58-59

(Gale 1993) William A. Gale; Kenneth W. Church **A Program for Aligning Sentence in Bilingual Corpora** *Computational Linguistics* 19(1), 75-102

(Kay 1988) Martin Kay; Martin Röscheisen. **Text-Translation Alignment** *Technical Report, Xerox Palo Alto Research Center*

(Kay 1993) Martin Kay; Martin Röscheisen. **Text-Translation Alignment** *Computational Linguistics* 19(1), 121-142

(Liu 1995) Xin Liu; Ming Zhou; Changning Huang. **An Experiment to Align Chinese-English Parallel Text Using Length-Based Algorithm** *Advances and Applications on Computational Linguistics*, pp 62-67

(Raggett 1996) Dave Raggett; Jenny Lam; Ian Alexander **HTML 3 Electronic Publishing on the World Wide Web** *Addison-Wesley*

(Sato 1990) Sato, S; Nagao, M. **Toward memory-based translation**. *Proceedings of 15th International Conference on Computational Linguistics(COLING 90)*

(Tan 1995) Tan, C. L; Nagao, M., **Automatic Alignment of Japanese-Chinese Bilingual Texts** *IEICE Transactions On Information and Systems* Vol. E78-D, No.1 Jan. 1995

(Xu 1996) Donghua Xu; Chew Lim Tan. **Automatic Alignment of English-Chinese Bilinguai Texts on CNS News** *International Conference on Chinese Computing '96 Singapore* pp 90-98