doi:10.3969/j.issn.1672 - 4348.2014.05.009

哈希算法与语义映射在语料库对齐中的运用

戴光荣[1,2], 宋玉春[3]

(1. 福建工程学院 人文学院, 福建 福州 350118; 2. 澳门大学 英文系, 中国 澳门;
3. 韶关学院 物理与机电工程学院, 广东 韶关 512005)

摘要: 探讨汉英句级对齐软件设计中两项主要技术,即哈希算法与词典语义映射在对齐中的运用。哈希算法能帮助软件从词典大量的英汉词条语义信息中快速提取所需的对应义,结合语义映射,将需要对齐的句子关键词信息进行语义识别,从而有效提高汉英句子对齐效果。

关键词: 哈希算法; 词典语义映射; 对齐技术; 平行语料库

中图分类号:TP391　　**文献标志码**: A　　**文章编号**: 1672 - 4348(2014)05 - 0454 - 05

Applications of hash algorithms and semantic mapping in C-E sentential alignment

Dai Guangrong[1,2], Song Yuchun[3]

(1. School of Humanities, Fujian University of Technology, Fuzhou 350118, China;
2. English Department, University of Macau, Macau, China;
3. College of Physics and Electromechanical Engineering, Shaoguan University, Shaoguan 512005, China)

Abstract: Automatic sentential alignment of Chinese and English texts is of critical importance to formulate Chinese-English parallel corpora. The alignment quality has direct impact on the reliability of related research such as machine translation, bilingual lexicography and contrastive language studies. This research adopts the two key technologies of alighment software designing, i.e. hash algorithm and semantic mapping to perform sentential alignment of Chinese/English texts. hash algorithms enable retrieving information from the knowledge database in a high speed, and can conduct semantic recognition of the key words/information of the sentences to be aligned via semantic mapping to attain effective E-C corpora alignment. This research can shed new light on Chinese/English sentential alignment.

Keywords: hash algorithm; semantic mapping; alignment technology; parallel corpus

近年来,随着计算机技术的进步以及语料库语言学的发展,平行语料库日渐得到研究者的注意,而且由于其本身所具有的优势,它在众多领域得到了广泛的运用,尤其是在机器翻译和机器辅助翻译中得到越来越多的应用。先前研究表明:基于平行语料库的方法不仅能够提高机器翻译的质量,还可以加强机器辅助翻译中的人机交互。同时双语平行语料库还在专业术语库的创建、多语信息提取、语义消歧等多方面有着广泛的运用。

双语句级层面上的对齐技术是创建及加工双语平行语料库的基础,对齐效果的优劣也将影响后期相关领域工作(如机器翻译、词典编纂、英汉

收稿日期: 2014 - 04 - 24
基金项目: 福建省社科规划一般项目(2013B222);福建省教育科学"十二五"规划重点课题(FJJKCGZ14 - 018)
第一作者简介: 戴光荣(1973 -),男(汉),湖南武冈人,副教授,博士研究生,研究方向:语料库语言学、语料库翻译学。

语言对比研究、译者培训等）的顺利开展。它在理论与实际应用方面，具有不可低估的价值。随着当前中国文化走出去的需求加大，提高本土译者英译的质量，译者需要更多更好的汉英双语平行语料库来做为翻译参考。因此，实现汉英双语句子自动对齐是汉英平行语料库创建者以及相关研究人员的首要任务。

本研究在国内外现有英汉平行语料库技术基础上，将对齐技术作为研究对象，采用哈希算法，通过建立、维护对齐知识库，结合英汉词典语义映射模式，设计出高效对齐软件，实现高质量英汉双语句子层面自动对齐。

一、现有汉英语句对齐技术及其局限性

目前实现汉英句子层面文本对齐的方法大致有如下几种：

（一）基于长度的方法

该方法最早由 Brown 和 Gale 提出，各自侧重点不一。Brown 的算法以词为单位计算句子长度，Gale 则以字符为单位计算句子长度[1-2]。他们认为原文和译文的长度之间存在一定的比例关系，译文的句子长度与原文的句长成一定的正比例关系。对语源相近的语言，如英语与法语，这种方法尤其有效[3]。由于英汉两种语言不存在拼写、语音或语义相似的同源词，所以基于长度算法的同源词对齐方法就不适用于英汉双语对齐。另外，如果单纯地使用基于长度的方法，效果也不佳，因为汉语分词问题还有待完善，利用词的个数作为长度单位也就不可靠，且分词结果也会影响互译信息率的计算[4]。

（二）基于词汇的方法

这方面的工作以 Kay 和 Röscheisen 的算法为代表。他们认为源文单词和其译文应该是同现的，其分布具有相关性，因此他们采用松散范例（relaxation paradigm）来进行对齐，他们用少量的英、德句子对这种方法作了示例，但未提供准确率[5]。有研究人员对 Kay 和 Röscheisen 的算法进行了改进，提出利用翻译模型进行双语句子对齐的方法，认为最佳句子对齐序列就是在给定的翻译模型下产生该双语语料概率最大的句子对齐状态[6]。与 Brown 及 Gale 的算法进行比较，基于词汇的算法正确率高，而且在处理大量省略的对齐中能轻易确定省略的位置，且鲁棒性好[7-10]。

（三）混合法，即将长度与词汇线索相结合的方法

吴德恺用此方法对齐了相当部分汉英双语的香港汉莎语料库[11-12]。Melamed 的研究也表明，将句子长度与词汇信息相结合进行句子对齐，效果要比单一方法好[13]。也有研究者提出一种以基于长度的统计对齐方法为主、双语词汇信息及标点符号为辅的对齐法。这种扩展方法一定程度上避免了复杂的中文处理（如汉语分词和词性标注），而且在统计方法中引入了关键词信息，以提高句子对齐的正确率[14]。其他学者还提出基于长度和位置信息的混合对齐方法[15]。

从目前句级对齐技术来看，现有的技术存在以下不足：

基于句子长度的方法适应范围大多局限在语源相近、语系相同的两种语言之间（如英语与法语）。对英汉两种语言来说，它们之间存在巨大差异，在实现这两种语言的语句层面上的对齐时，如果采用单纯的基于长度的方法，效果并不会十分理想[16]。

基于词汇信息的方法所遇到的最大问题就是搜索空间比较大，花费的时间太长，获得词汇对等信息的代价比较高。再加之一词多义现象的存在，使得对应信息的搜索变得更加复杂而最终效果不佳。

基于长度与词汇混合的方法也大多局限在语源相近，对于英汉语言来说，也需要很长的时间来实现翻译词义的匹配。

从众多的研究成果看出，汉英两种语言之间差异大，要实现汉英句级之间的自动对齐，要考虑的因素比较多，难度与挑战都相当大。

二、词典语义映射与哈希算法在句对齐中的运用

在本研究中，我们将标点符号、句子长度以及词汇搭配等关键性因素加以综合，提出一种以词典语义映射的方法来实现汉英句子层面的对齐。

所谓基于词典语义映射，即把源语言文本看成单词的序列作为横轴，横轴上的每个点对应一个单词；同样以目标语言文本作为纵轴。用平面上的一个点来表示源语言文本中某个词和目标语言文本中的某个（些）词对译。要准确找到平面

上这个点,可采用基于双语词典信息的方法,为句对齐提供准确、翔实的双语词汇信息,从而为实现高质量的对齐效果提供基本保障。

我们认为,对齐软件必须基于累计的知识,包括汉英词典、中文同义词典、中文反义词词典、常见句子数据库(中文和英文)、习惯用语数据库等。这个知识库有海量的数据,要实现词语意义的一一映射,即在不同语境之下,源语文本(假设以英语为源语,汉语为译语)的某一个词对应于相应的译语词,需要付出很长的时间。为此我们采用 hash 算法来管理和使用这些知识数据。

为完成高质量的对齐,我们应用评价函数对结果进行反馈,形成迭代算法,而迭代算法会大大降低对齐速度。利用 hash 算法的添加、删除、查询知识的快速性,降低每次循环必须做的查找环节的时间消耗,使整个运算过程在普通电脑配置时达到人们可接受的时间预期。所以必须寻找合适的 hash 算法,研究它的性能以及对当前应用的适应性。在研究过程中,我们采用了多种 hash 算法,并应用 MATLAB 对不同算法进行图形化分析。下面简单展现了两种关键算法的不同效果。

算法 1:

```
uint calc_hashnr(const byte *key,uint length)
  {
    register uint nr=1, nr2=4;
    while (length--)
    {
      nr^= (((nr & 63)+nr2)*((uint)(uchar) *key++))+(nr<<8);
      nr2+=3;
    }
    return((uint) nr) % MAX_PRIME_LESS_THAN_HASH_LEN;
  }
```

该函数来自于开放源码的关联数据库管理系统 MySQL,是一个具有较高字符串处理效率的算法。此函数输入字符串指针 *key 和字符串长度 length,返回一个无符号整数值,其中 MAX_PRIME_LESS_THAN_HASH_LEN 限制了函数返回值的范围。

当我们把一本英汉字典的所有单词依次输入时,就得到了多个分布的无符号整型值,这些值重复率越小说明函数映射性能越好。

算法 2:

```
unsigned int hash_ipv(char *str, int len)
  {
    register unsigned int sum = 0;
    register unsigned int h = 0;
    register char *p = str;
    while(p - str < len)
    {
      register unsigned short a = *(p++);
      sum^= a * (p - str);
      h^= a/(p - str);
    }
    return ((sum<<16)|h) % MAX_PRIME_LESS_THAN_HASH_LEN;
  }
```

该函数是采用英文单词高低位组合方法进行设计的 hash 函数(有关该函数的具体操作,请参阅"基于英文单词的快速 hash 索引算法"①),函数的输入和输出数据类型同算法 1。

我们对《牛津高阶英汉双解词典》(2009 年第 7 版)进行整理,收集到 10 万英文词条及其相应的汉语译文词条(如果一个英文单词有多个汉语意义,也一并收入,但排除英文例句及其汉译)。在对齐过程中,从英语词条查找汉语单词,通过 hash 函数把 10 万单词映射到 0 ~ 199,999 的地址区间。

为了图形分析 hash 算法应用于此词典的性能,我们自主研发一种基于 MATLAB 的图形化的分析方法,对上述两种 hash 算法进行比较分析,研究它们的性能以及对当前应用的适应性。我们把 hash 映射的结果转化为在二维平面 $X \times Y$ 上分布的数据,其中区间 X 范围为[0,399],区间 Y 范围为[0,499],$X \times Y$ 平面上共有 400×500 个采样点,对应地址 0 ~ 199,999,每个地址上存贮单词的个数作为对应采样点的 Z 坐标(为了便于观察,所有纵坐标向 Z 轴正向平移 30),在 MATLAB 中用 meshgrid 产生 $X \times Y$ 平面网格数据,然后根据 Z

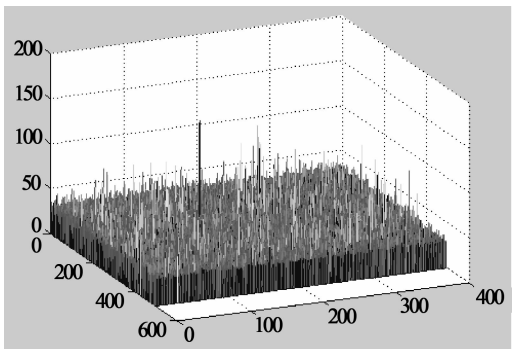
① 该文网址:http://fullfocus.iteye.com/blog/133223,访问日期为 2014-01-20。

坐标用 mesh 函数生成算法 1 和算法 2 的三维分布图，见图 1。

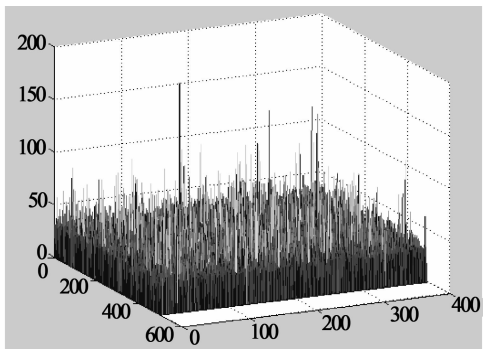
从图 1 可以看出，hash 算法 1 和算法 2 都能均匀地把 10 万单词映射到 0 ~ 199 999 的地址区间，但算法 2 的 Z 坐标高于 30 的线长势更茂盛，说明不同单词映射到同一地址的情况数量上更多，这样就会影响给定的任意单词的搜索语义的效率。

为了更清楚地看出不同单词映射到同一地址空间数量的等级，我们将 hash 算法 1 和算法 2 的正面三维分布侧面图展示如下，见图 2：

从图二可以看出不同单词映射到同一地址空间数量的等级。图 2(a)中最大值为 150，图 2(b)中最大值是 180；图 2(a)中高于 30 的值基本分布在[30,80]区间，图 2(b)高于 30 的值基本分布在[30,100]区间，而且这些值的数量图 2(a)远少于图 2(b)，这说明在不同单词映射到同一地址的程度上算法 2 的更严重。



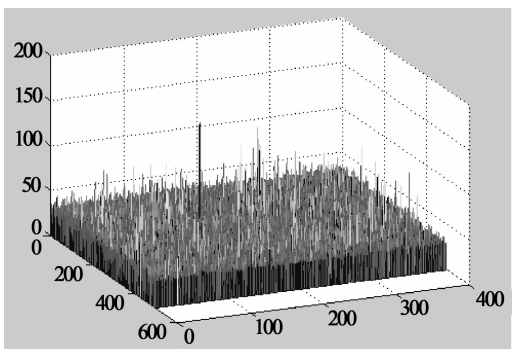
(a)算法 1



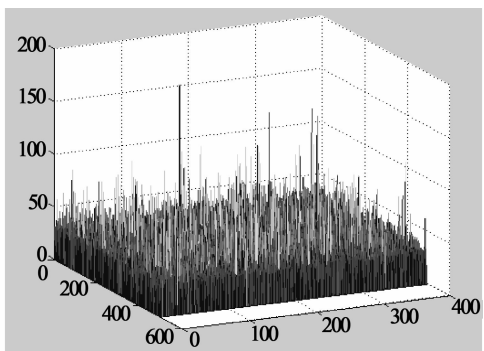
(b)算法 2

图 1　hash 算法 1 和算法 2 的三维分布正面图

Fig. 1　3D front view of hash algorithms 1 and 2



(a)算法 1



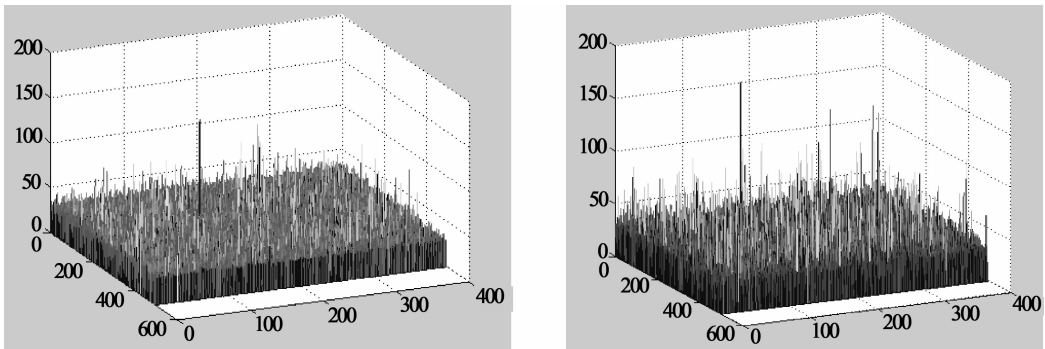
(b)算法 2

图 2　hash 算法 1 和算法 2 的正面三维分布侧面图

Fig. 2　3D lateral view of hash algorithms 1 and 2

图 3 可以反映单词在地址空间上分布的均匀性，仔细观察，图 3(a)颗粒感是点，而图 3(b)的颗粒感是短线状，这说明算法 1 的均匀性能优于算法 2。我们可以根据这个思路研究更多 hash 算法应用到对齐、管理语料库时的性能。

算法 1 已经应用在我们开发的对齐软件中。处理方法是首先把词典的英文单词用此 hash 算法生成地址码，这个地址码同此英文单词及其解释形成一条数据记录，然后把这些记录按照地址码的大小从小到大排序，由于 hash 算法计算的地址码在[0,199 999]区间里有缺失，这些缺失的地址码补充空白的记录，把这 20 万条记录作为程序的数据在对齐软件运行时调入内存。要查找某个英文单词的意义，首先用算法 1 求出地址码，根据这个地址就可以在内存的数据区中查出包含此英文单词的记录，在这个记录中再找到此单词的意义。词典中单词的译文是脱离上下文语境的，在对齐过程中，我们结合该词与前后词搭配情况进



(a)算法 1 (b)算法 2

图 3 hash 算法 1 和算法 2 的三维分布顶视图

Fig. 3 3D top view of hash algorithms 1 and 2

行语义识别,转化为程序可识别的语义知识,进而提供语义数据支持。应用结果表明,此算法可以准确、快速地为对齐算法提供有效的语义数据支撑,从而达到了较好的对齐效果。经对 15 个政府文本(中文单语包含约 10 万字)的先导性测试,汉英句子对齐准确率达到 95.7% 以上。

三、结论与展望

基于语义的英汉双语对齐算法是比较合理的算法,而此方法必须有语义数据库做支撑,这个语义数据非常庞大,对其要准确而快速地使用,必须采用专门的海量数据处理方法,hash 算法就是一种重要的行之有效的方法。hash 算法有很多种具体运算方法,必须采用类似本文的方法根据具体情况对其进行有效的分析和实验。单个 hash 算法可实现数据的查找、删除,对于对齐算法甚至自动翻译领域,需对后台支撑的海量数据不断扩充,这可采用数据库索引等方法进行完善。需要明确的是,在数据扩充量越来越大的时候,数据库索引的建立和检索速度会随着下降,因此还需对搜索引擎增量索引和检索技术进行完善。我们在研究过程中发现,要实现准确且有实用价值的对齐系统,必须采用融合模糊技术、神经网络等人工智能方法,来实现初步模仿人对语句和语句对齐的理解。有关模糊技术与神经网络的运用,我们将有专文探讨。

参考文献:

[1] Brown P F, Lai J C, Mercer R L. Aligning sentences in parallel corpora [C]//Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics[ ]. Berkeley: ACL,1991:169 - 176.

[2] Gale W A, Church K W. A program for aligning sentences in bilingual corpora[J]. Computational Linguistics,1993,19(1):75 - 102.

[3] 王建新. 计算机语料库的建设与应用[M]. 北京:清华大学出版社,2005:121 - 122.

[4] 黄俊红,范云,黄萍. 双语平行语料库对齐技术述评[J]. 外语电化教学,2007(6):21 - 25.

[5] Kay M, Röscheisen M. Text-translation alignment[J]. Computational Linguistics,1993,19(1):121 - 142.

[6] Chen S F. Aligning sentences in bilingual corpora using lexical information[C]//Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics. Columbus: ACL,1993:9 - 16.

[7] McEnery T, Piao S,Xin X. Parallel alignment in English and Chinese [C]//Botley S P, McEnery A M,Wilson A (eds), Multilingual Corpora in Teaching and Research. Amsterdam: Rodopi,2000:177 - 191.

[8] Oakes M, McEnery T. Bilingual text alignment: an overview [C]// Botley S P, McEnery A M, Wilson A(eds). Multilingual Corpora in Teaching and Research. Amsterdam: Rodopi,2000:1 - 37.

[9] Simard M, Foster G, Hannan M L, et al. Bilingual text alignment: where do we draw the line? [C]// Botley S P, McEnery A M,Wilson A (eds). Multilingual Corpora in Teaching and Research. Amsterdam: Rodopi,2000:38 - 64.

[10] 冯敏萱. 汉英平行语料库的平行处理[M]. 北京:世界图书出版公司,2011:21.

(下转第 463 页)

开发以及保护。

参考文献：

[1] 王文章. 非物质文化遗产概论[M]. 北京：教育科学出版社，2013.
[2] 谢建平. 功能语境与专门用途英语语篇翻译研究[M]. 杭州：浙江大学出版社，2008.
[3] 单霁翔. 大运河遗产保护[M]. 天津：天津大学出版社，2013.
[4] 华干林，黄俶成. 非物质文化遗产保护所面临的几个问题[J]. 民族遗产，2009(2)：75－83.
[5] 顾军，苑利. 论扬州文化的传承与弘扬[J]. 扬州大学学报：人文社会科学版，2001(11)：75－78.
[6] 黄友义. 坚持“外宣三贴近”原则，处理好外宣翻译中的难点问题[J]. 中国翻译，2004(6)：29－30.
[7] 贾文波. 应用翻译功能论[M]. 北京：中国对外翻译出版有限公司，2012.
[8] 肖曾艳. 略述非物质文化遗产的旅游开发[J]. 肇庆学院学报 2008(1)：42－45.
[9] 陈芳蓉. 中国非物质文化遗产英译的难点与对策[J]. 中国科技翻译，2011(2)：41－44.

（责任编辑：王明秀）

（上接第 458 页）

[11] Wu D. Aligning a parallel English-Chinese corpus statistically with lexical criteria[C]// The Proceedings of the 32nd ACL. New Mexico State University, Les Cruces, New Mexico, 1994：80－87.
[12] Wu D. An Algorithm for Simultaneously Bracketting Parallel Texts by Aligning Words[C]// Proceedings of the 33rd ACL. MIT, Cambridge, Mass, 1995：244－251.
[13] Melamed I D. Models of translational equivalence among words[J]. Computational Linguistics, 2000, 26(2)：221－250.
[14] 张艳，柏冈秀纪. 基于长度的扩展方法的汉英句子对齐[J]. 中文信息学报，2005，19(5)：31－37.
[15] 李维刚，刘挺，张宇，等. 基于长度和位置信息的双语句子对齐方法[J]. 哈尔滨工业大学学报，2006，38(5)：689－692.
[16] 梁茂成，许家金. 双语语料库建设中元信息的添加和段落与句子的两级对齐[J]. 中国外语，2012，9(6)：37－43.

（责任编辑：王明秀）