构建大规模汉越双语平行语料库[1]

罗林[1] 郭剑毅[1,2] 余正涛[1,2] 毛存礼[1,2] 莫媛媛[1]

1.昆明理工大学信息工程与自动化学院，昆明 650051
2.昆明理工大学智能信息处理重点实验室，昆明 650051

摘要 双语平行语料库作为一项重要的基础语言资源，在语言研究和机器翻译研究领域中的作用越来越重要。本文通过研究汉越双语的语言特点，详细介绍了一个大规模汉越双语平行语料库的构建过程，包括汉越双语语料的收集、整理、存储，以及在此基础上实现了汉越双语语料的标注、加工、处理，从而实现了汉越双语平行语料库的构建，该工作的深入和开展将促进相关理论的研究和应用技术不断发展。

关键字：越南语-汉语；语料库构建；双语平行语料库；对齐系统

Construction of large-scale Chinese and Vietnamese bilingual parallel corpus

Luo Lin[1] Guo Jian-yi[1,2] Yu Zheng-tao[1,2] Mao Cun-Li[1,2] Mo Yuan-yuan[1]

1. The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051

2. The Institute of Intelligent Information Processing, Computer Technology Application Key Laboratory of Yunnan Province, Kunming 650051

Abstract

Bilingual parallel corpus as an important basis for language resources, increasing role in the study of language and machine translation research in the field of more important. Through the study of the language features of Chinese bilingual text, introduces in detail the construction of a large-scale Sino Vietnamese bilingual parallel corpus process, including collection, sorting, storage of Chinese and Vietnamese language, and on the basis of the annotation, processing, processing more Chinese bilingual corpus, so as to realize the construction of Chinese and Vietnamese bilingual parallel corpus in this work, the thorough and the development will promote the research and application of technology development of the relevant theories.

Key words Vietnamese - Chinese; Corpus building; bilingual parallel corpus; Alignment System

---

1. **基金项目：本文得到国家自然科学基金(6126041)的资助。作者简介：罗林，男，1980年10月，在读研究生，研究方向：自然语言处理，信息抽取。通讯作者：郭剑毅，教授，电子邮箱：gjade86@hotmail.com**

1. 引言

伴随着计算机技术的迅速发展，语料库资源在自然语言处理研究中的巨大价值已经得到越来越多的认可，特别是双语平行语料库的建设和应用得到了广泛的重视，并取得了一定成果。目前在国内，相关的研究和介绍主要侧重于汉英、日汉、德汉等双语语料库的建设以及对齐加工和标准[1-2]，多级自动对齐技术以及双语平行语料库在机器翻译和翻译知识获取等方面的应用技术，并取得了可观的进展。如哈尔滨工业大学的汉英平行语料库已经直接用来开发汉英双向机器翻译系统，这个语料库约有 6 万个汉语和英语句子，使用多级对齐加工技术，分别按照句子、短语结构和词一一对齐[3-4]。中国科学院计算机研究所和中国科学院自动化研究所也在 973 课题的支持下联合开发了 20 万句子对齐的汉英双语语料库[5-6]。除此之外，中国科学院软件研究所、清华大学、东北大学、南京师范大学、国家语委等单位也相继建立了一定规模的汉英双语语料库[7-9]。但针对全世界被广泛使用处于第 14 位的越南语的研究还很少。目前越南语信息化方面的研究已经取得了一定的成果，如解放军外国语学院周云构建了 66633 个词条的汉-越双语词典，并开发了“汉语越南语机器翻译实验系统 HanViet0.1”[10]，华中师范大学收集并构建了越南语和汉语时政文本的越文译语的双语语料[11]。但就深入开展越南语-汉越双语信息化处理研究来说，还存在诸多问题，一定规模的汉语-越南语词语级对齐语料库的缺乏就是面临的主要问题之一。目前，本课题组已经收集了大量的汉语-越南语对照的书籍和网站（如越南日报网）信息，但还没有现成已经标记好的汉-越双语词对齐语料库，并且语料标记还存在规范性以及其他语料之间的共享方面问题，采用人工方式标记大规模汉-越词语对齐语料、段落对齐语料、句子对齐语料，还面临标记工作量大，效率低等问题，因此本文研究汉语-越南语对齐语料库的构建以及关键技术难点问题，对于推进机器翻译，跨语言信息检索等研究的实用化具有重要意义。

2. 语料库的构建

2.1 语料的收集和整理

语料采集是语料库建设的首要任务，在实际语料采集时，所收集到的双语语

料涉及不同的文体、领域、语体、创作时期等，这些因素直接影响着双语语料库的构建和应用。根据越南语和汉语的语料分布、领域等方面实际情况的分析，在双语语料收集时，需从语料分类、语料年限划分、语料比例这几个方面进行选材。在语料分类方面，按照语料所属领域进行多重分类，对所要收集的双语语料进行分类选材，以保证所收集到的双语语料具有广泛性、多样性，具体分类原则见表1。在语料年限划分方面，尽可能选择年代比较近，但又有一定的时间跨度的双语语料。在语料比例方面，依据语料年限、类别分配的比例、数量进行选取，少数语料可根据实际情况进行比例、数量的调整，从而保证语料的合理性，避免语料的单一性，降低语料库的拟合度。

表 1 语料分类表

大类	小类	数量
政治	哲学、政治、宗教、法律	4 千余篇
历史	历史、考古、民族	3200 余篇
社会	社会学、心理、语言文字、教育、文艺理论、新闻、民俗	2700 余篇
经济	工业经济、农业经济、政治经济、财贸经济	3600 余篇
艺术	音乐、美术、舞蹈、戏剧	1000 余篇
军体	军事、体育	2000 余篇
其他	.....	3000 余篇

原始语料来源于不同的收集者，其中大部分都处于杂乱无章的状态，表现为：存放格式各异，没有形成篇章级对齐单位；文本、领域、语体、创作时间各异；含有不利于加工处理的噪音信息；并且有大量的重复语料等。这些因素均妨碍了对原始语料的进一步加工和利用，因此，必须对原始语料进行系统的整理。在语料的整理过程中，首先对语料按照采集分类标准进行粗分类，并在此基础上用文字处理工具（如：UltraEdit 工具）将越-汉两种语言的文本语料分成句子，每个句子占一行。句子的定义为：以句号、问号、感叹号、分号结尾的一串字符串。最终形成统一格式的文本文件，以便于进一步加工处理。具体格式见图 1 所示。

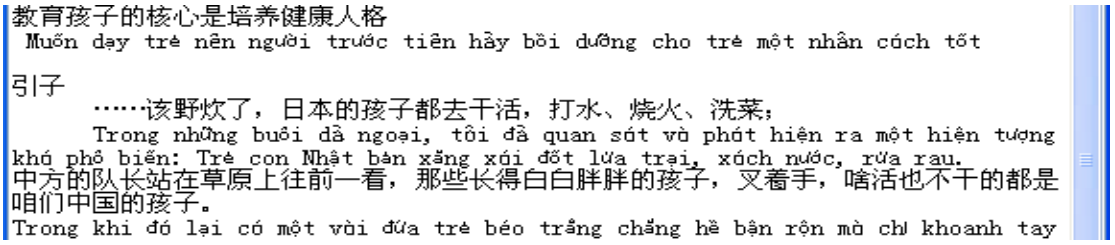

图 1 越-汉语料整理格式

2.2 越南语-汉语双语语料的加工

语料加工是语料库系统性构建的核心任务，一个高质量的语料库并不是任意语料文本的任意集合，需要对收集到的语料进行加工处理，才能形成有应用价值的语料库。汉-越双语语料库采用多级加工处理方法，该方法主要分以下几个步骤，首先对汉-越双语语料文本进行篇章/段落级对齐，提取出篇章/段落级对齐单位的文体、领域、语体等基本属性，并在此基础上，进行句子对齐，最后基于汉-越双语自动分词的技术上进行细粒度的词汇级对齐，最终构建成词汇级的汉-越双语语料库。其原理见图 2 所示。

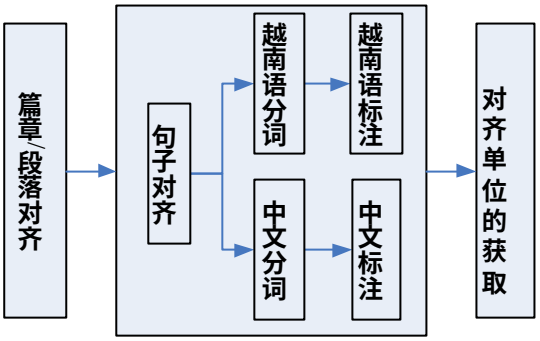

图 2 汉-越双语语料库构建原理

2.2.1 越南语-汉语双语篇章/段落级对齐

篇章/段落对齐是语料对齐的粗粒度的加工处理，同时也常常是进行下一步细粒度对齐的必要的前提。通过汉-越双语语料管理平台对原始语料进行篇章/段落对齐，并将对齐结果存储到语料信息数据库，每个篇章/段落对齐单位在该数据库中都一个记录，包含 id 编号、中文/越南语标题、领域、文体、作者、创作时间等信息。其中 id 编号表示每条记录的序号且唯一，确保中文/越南语标题语料对具有唯一性，例如图 3 所示。

	XH	AREA	LY	CNBT	VNBT	CZSJ	USERNAME
1	1	教育	书刊	习惯决定孩子命运	Thói quen quyết định số phận của	2013-7-11 14:36:57	000000
2	0	新闻	新闻	交流与沟通	cha mẹ cần phải chú	1970-1-1	1

图 3 篇章/段落对齐记录

2.2.2 越南语-汉语双语句子、词汇级对齐

句子对齐，即找出源文句子在译文中对应的翻译句子。由于句子对齐的粒度

小于段落的粒度，因此对齐的句子能比对齐的段落提供更详细的对译信息。汉-越双语句子对齐是在段落对齐的基础上进行更细粒度的对齐，利用汉-越双语语料管理平台的句子对功能对段落中的句子对进一步加工，将句子对齐结果集保存到信息数据库，每个句子对齐单位包含对齐 id 编号、段落编号、中文句子、越南语句子等信息。其中段落编号对应于段落对齐记录中的 id 编号，段落和句子之间的是一对一、一对多、多对多对应关系。句子对齐记录集中的 id 编号具有唯一性，保证汉-越句子对不存在重复记录。

词汇对齐相对于段落和句子对齐而言，更复杂，难度也更大。汉-越词汇对齐是在研究源文和译文词汇之间的对应关系的基础上，利用现有的中文分词技术、越南语分词技术，通过汉-越双语语料管理平台和分词辅助工具对对齐的句子进行词汇级对齐，并将结果集保存到信息数据库，每个记录包含汉语分词结果、越南语分词结果、带序号的汉语分词句子、带序号的越南语分词句子、词汇对应关系等信息。带序号的汉语分词句子表示每个中文词汇在汉语句子中的位置编号，带序号的越南语分词句子表示每个越南语词汇在越南语句子中的位置编号，词汇对应关系表示源文中的词汇与越南语中的词汇对应位置编号序列对。具体示例如图 4-5。

HYFCJZ	YYJZ	DXHHYJZ	DXHYYJZ
也/d 是/v 他/r 心理/n 素质/n	đồng thời cũng là dấu hiệu	1-也 2-是 3-他 4-心理 5-素质	1-đồng_thời 2-cũng 3-
对/p 孩子/n 来说/u ，/w 中国/	Đối với trẻ mà nói, rất nhi	1-对 2-孩子 3-来说 4-， 5-中国	1-Đối_với 2-trẻ 3-mà_
有人/r 说/v ：/w “/w 这个/	Có người nói rằng: “Những	1-有人 2-说 3-： 4- 5-“ 6-这	1-Có_người 2-nói 3-rằ
青春期/t 的/b 女孩/n 的/b 隐私	Khi nhận thức của con cái v	1-青春期 2-的 3-女孩 4-的 5-隐	1-Khi 2-nhận_thức 3-c
父母/n 不/d 应/v 害怕/v 而/c	cha mẹ không nên lo sợ mà l	1-父母 2-不 3-应 4-害怕 5-而	1-cha_mẹ 2-không 3-nê
因为/p 自己/r 的/u 女孩/n 已经	vì như thế chứng tỏ một điề	1-因为 2-自己 3-的 4-女孩 5-已	1-vì 2-như_thế 3-chứn
这时候/r ，/w 做/v 父母/n 的/	Trong khoảng thời gian này,	1-这时候 2-， 3-做 4-父母 5-的	1-Trong 2-khoảng 3-th
给/p 女孩/n 一点/t 自己/r 的/	dành cho con một khoảng khô	1-给 2-女孩 3-一点 4-自己 5-的	1-dành_cho 2-con 3-mộ
保护/v 个人/n 隐私/n 是/v 适应	Bảo vệ quyền riêng tư cá nh	1-保护 2-个人 3-隐私 4-是 5-适	1-Bảo_vệ 2-quyền_riên
保护/v 隐私/n 就是/d 保护/v 自	bởi đó chính là cách chúng	1-保护 2-隐私 3-就是 4-保护 5-	1-bởi_đó 2-chính_là 3
让/v 她们/k 在/p 自己/r 的/u	Để con tự do vùng vẫy và v	1-让 2-她们 3-在 4-自己 5-的	1-Để 2-con 3-tự_do 4-
给/p 父母/n 的/u 建议/n 一/m	Lời khuyên thứ nhất: Không	1-给 2-父母 3-的 4-建议 5-一	1-Lời_khuyên 2-thứ_nh
这里/r 要/v 说/v 的/u 是/v ，	Việc tôn trọng quyền riêng	1-这里 2-要 3-说 4-的 5-是 6-，	1-Việc 2-tôn_trọng 3-
用/p 父母/n 的/u 好/a 习惯/n	Dùng những thói quen tốt c	1-用 2-父母 3-的 4-好 5-习惯	1-Dùng 2-những 3-thói

图 4 句子对齐记录

XH	CYXH	HYCH	CX	YYCH	YYCHXH
347	1	现在	t	bây_giờ	2
347	2	总像	d	gần_như	4
347	3	是	v	là	5
347	4	有	v		
347	5	什么	r	gì	7
347	6	事	n	chuyện	6

图 5 词汇对齐记录

2.3 汉-越双语语料管理平台的开发

汉-越双语语料管理平台是一个相当重要的双语加工工具，使用 Java 技术、Oracle 数据库存储等技术实现了汉-越双语语料管理系统，该系统具备以下功能：

1） 篇章/段落对齐管理；

篇章/段落对齐主要通过两种方式输入到语料库中，第一种方式是对采集到没有噪声的篇章或段落通过 Excel 文件导入的方式直接导入到语料数据库中；第二种方式是通过人工录入的方式对采集到汉-越双语篇章或段落进行输入；具体功能如图 6。

图 6 篇章/段落对齐

2） 句子对齐管理；

句子对齐是在篇章/段落对齐的基础上，先通过辅助工具进行句子对齐，再通过人工校对、调整后，将结果录入到系统中进行保存。具体功能如图 7。

图 7 句子对齐

3） **词对齐管理；**

完成句子对齐后，系统集成了 ICTCLAS **中文分词、越南语分词工具，分别对中文、越南语句子进行自动分词标注，人工校对无误后，保存数据。具体功能如图** 8。

序号	2			
汉语句子	总之,良好的家庭环境，对于女孩来说，胜过父母给她百万金银。			
越语句子	Nói chung, đối với con gái, môi trường gia đình tốt còn quý giá hơn cả nghìn vàng.			
汉语分词句子	1-总之 2-, 3-良好 4-的 5-家庭 6-环境 7-， 8-对于 9-女孩 10-来说 11-， 12-胜过 13-父母 14-给 15-她 16-百万 17-金银 18-。			
越语分词句子	1-Nói_chung 2-, 3-đối_với 4-con_gái 5-, 6-môi_trường 7-gia_đình 8-tốt 9-còn 10-quý_giá 11-hơn 12-cả 13-nghìn 14-vàng 15-.			
词语对齐情况	总之-Nói_chung|良好-tốt|家庭-gia_đình|环境-môi_trường|对于-đối_với|金银-vàng			
序号	词语序号	汉语词汇	越语词汇	操作
2	1	总之	Nói_chung	修改
2	2	，	**	修改
2	3	良好	tốt	修改
2	4	的	**	修改
2	5	家庭	gia_đình	修改
2	6	环境	môi_trường	修改

图 8 **词对齐**

3. 结束语

本文通过研究现有的中文、越南语分词技术和语料库构建方法，首先对汉-越语料进行篇章/段落对齐，然后再进行句子对齐，并在此基础上，对汉-越句子对分别进行分词和词性标注，通过对实际文本对照分析，建立双语词汇级对齐语料。此外，基于上述方法实现了汉-越双语语料管理平台。

4. 参考文献

[1] **王占军，姚卫东.一种汉英双语句子自动对齐算法**[J].**计算机仿真，**2009,26(2):329-3.

[2] **钱丽萍，赵铁军，杨沫昀，等.基于译文的英汉双语句子自动对齐**[J].**小型微型计算机系统**,2001，22(1):123-125.

[3] **原双庆，李芳，盛焕烨．多语种翻译词汇的在线自动抽取［J］．计算机研究与发展，**2004, (2):843-847.

[4] Philip Resnik，Noah A．Smith．The Web as a Paral-lel Corpus[J]．Computatinal

Linguistics，2003, 29(3):349-380.

[5] Pu-Jen Cheng，Wen-Hsiang Lu，Jer-Wen Teng，et al．Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpor[C]//Annual Meeting of the Association for Computational Linguistics (ACL-2004).

[6] Pu-Jen Cheng，Jei-Wen Teng，Ruei-Cheng Chen，et al．Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval[C]//The Proceedings of the SIGIR-2004.

[7] W.Kraaij，J.-Y.Nie，M.Simard．Tmbedding Web-based Statistical Translation Models in Cross-Language Inforrnation Retrieval[J] ． Computational Linguistics，2003, 29(3):381-419.

[8] 刘非凡，赵军，徐波．大规模非限定领域汉英双语语料库建设及句子对齐研究［C］//全国第７届计算语言学联合学术会议，2003:339-345.

[9] 孙茂松，陈群秀．语言计算与基于内容的文本处理［M］，清华大学出版社，2003, 7,97-102.

[10]周云.汉语越南语机器翻译实验系统[D]. 解放军外国语学院硕士论文,2006.

[11]胡氏贞英.汉语非文学文本越译研究—以汉语时政文本越译为例[D].华中师范大学博士论文,2011.