文章编号：1003-0077（2011）00-0000-00

面向 ELAN 软件的手语汉语平行语料库构建*

吴蕊珠[1]，李晗静[2*]，吕会华[2]

（1. 北京联合大学 北京市信息服务工程重点实验室，北京 100101；

2. 北京联合大学特殊教育学院，北京 100075）

摘要：手语汉语平行语料库建立的目的是用于机器翻译和语言对比研究，并且能够系统地保存手语资源，保护手语和聋人文化。手语汉语平行语料库存储的内容主要包括手语视频、被采集者信息和标注者信息，以及通过多媒体标注软件 ELAN 转写的十四层标注信息，包括手控和非手控信息。本文提出使用基于向量空间的余弦相似性算法实现了手语语料相似度的计算来帮助语料库去重，并取得了较明显的效果；同时用此算法进行专家相似度测试以确保语料库的质量。

关键词：手语；平行语料库；转写

中图分类号：TP391　　**文献标识码：**A

Construction of Parallel Corpus of Chinese and Sign Language for ELAN Software

WU Rui-zhu[1], LI Han-jing[2] ,LV Hui-hua[2]

(1.Beijng Key Laboratory of Information Service Engineering ,Beijing Union University ,Beijing, 100101, China ; 2. Special Education College of Beijing Union University, Beijing, 100075 ,China)

Abstract: The purpose of the parallel corpus of Chinese and sign language construction is to use for machine translation and language comparison studies, and to systematically preserve sign language resources, protect sign language and deaf culture. The contents of the sign language Chinese parallel corpus store mainly include sign language video, information of the collected person and annotator information, as well as 14 layers of labeling information transferred through the multimedia labeling software ELAN, including manual and non-manual information. This paper proposes to use the cosine similarity algorithm based on VSM to realize the calculation of sign language corpus similarity to help corpus deduplication,and the obvious effect is obtained. At the same time, the algorithm is used to test the similarity of the expert to ensure the quality of the corpus.

Key words: sign language; parallel corpus; gloss

1 引言

手语是一种视觉语言，它是通过手的动作、面部表情的变化和身体的运动进行交流的语言。美国学者威廉姆·斯多基于 20 世纪 60 年代初发表了世界上第一本谈手语的著作《手语结构》，明确提出美国手语是一种自然语言。时至今日，手语语言学成为了语言学研究中不可或缺的组成部分，语言学家们开始从不同层面研究手语，而研究范围也从美国手语扩展到其他国家手语[1]。

需要指出的是，我们这里所说的手语均为自然手语，与手势汉语或手势英语等人造语言是不同的。手势汉语是根据汉语的语法规律、人为造出来与聋人交流的工具，其利用了汉语

* **收稿日期：**　　**定稿日期：**

基金项目：国家语委重点项目“手语语言处理的智能化理论和技术研究”（ZDI135-31）；北京教育科学规划重点课题“基于学习通用设计的聋人课程建设研究与实践 ”（ADA14121）；北京联合大学研究生资助项目

作者简介：吴蕊珠（1990——），女，硕士研究生，信息无障碍辅助技术；李晗静（1974——），女，教授，计算手语语言学；吕会华（1967——），女，副教授，手语和聋人汉语习得。

的语序，与自然手语的语法规律存在很大差别，聋人理解起来存在一定困难[1]。本文是面向自然手语进行收集和整理。

本文的工作主要是建立手语汉语平行语料库。平行语料库是指“由原文文本及其平行对应的译语文本构成的双语语料库，其双语对应程度可有词级、句级和段级几种”[2]。所以手语汉语平行语料库一方面是要有严格的手语语料的采集过程。采集设备及场景设置、采集内容、被采集者的选取和采集用到的诱导材料都需要建立标准。另一方面是用多媒体标注软件ELAN对收集到的手语语料进行手控和非手控信息等的标注，其标注者的选取和标注的标准也需要科学的指导。同时，本文建立的手语汉语平行语料库是为日后建立其他通用的手语语料库提供了有效的参考，为保证语料库标注质量提供了技术支持，它也能够为后续的手语机器翻译提供有力的数据基础。

为了有助于语料的去重和手语语料的分类，以及保证标注质量。提出对标注语料使用基于VSM的余弦相似性算法来实现手语语料相似度的计算。

2 相关研究

2．1 语料库

（1） 国内汉英双语平行语料库

北大计算语言学研究所的双语语料库，英汉对齐的句子已有5万多对，并开发了相应的对齐工具和双语语料库管理软件。正在此基础上做汉英对照短语库，预计规模将达数十万条；哈尔滨工业大学的英汉双语语料库于1998年有3万句子对，已经进行了词性标注，正在扩充为40-50万句子对，在句子、短语、词汇三级实现双语对齐；东北大学的英汉双语语段库：在双语语料库基础上，建造双语语段库，1999年构造了10万双语语段库，进行了基于语段的英汉机器翻译实验;中国科学院软件研究所的英汉双语语料库是双语对齐算法研究。现有 15万对英汉双语对齐句子库，已经切分和标注[3]。

（2） 澳大利亚手语语料库

目前最为成熟的手语语料库当属由Johnston等人创建的澳大利亚手语语料库[5]。该库的建设目的从早期的社会语言学描写研究，逐渐转移到手语的传承保护和词典编纂。该语料库的标注包括49层，其中用来对双手手形的意义、运动、位置等手控信息的标注层有37层之多；9层是对眼睛、眉毛、身体、头部等非手控信息的标注；2层是对于手语意义的标注，分别是句子翻译和词语转写翻译；最后一层是注释。澳大利亚手语语料库虽然是最为成熟的手语语料库，但其大部分标注层主要集中在双手空间信息的描述上、标注层过多、耗时耗力。所以该语料库很难复制或者推广。

（3） 德国天气预报手语平行语料库

该语料库的建立是为了手语的翻译和识别，将德语翻译成德国手语是该系统的目标[6]。基于统计的机器翻译要依赖海量的数据，该语料库收集了自1999年以来6年内德国天气预报的手语视频数据，包括2190个手语视频，德语手语句子对有72724对，词语数量872117个，词汇(去掉重复词）有12320个，而且其收集的是国家级天气预报，手语视频质量比较高。包括很多相同的句子句式，比如天气预报中的德语句子“Und nun die Wettervorhersage f¨ur morgen, Donnerstag, den zw¨olften Mai.”，德国手语句子标注为“JETZT WETTER+VORAUS+SAGEN MORGEN DONNERSTAG ZWO¨ LF MAI.”，表达的意思是“And now the weather forecast for tomorrow, the 12th of May.”，语料库中很多手语视频中都会有这样的句式，是有利于基于统计的机器翻译。其由三部分组层：一部分是手语视频数据的标注语料(The Video Corpus)，其中标注层有6层，分别是转写、词语类型、手语句子边界、相应的德语句子边界、德语使用者标注的德语句子翻译。另一部分是德语手语的文本语料(The Bilingual Text-based Corpus)，是将ELAN软件中的标注信息进行导出。还有一部分是天气预报的德语文本语料(The Monolingual Text-based Corpus)。其中标注信息中没有主手、

辅手、非手控信息的描述，对于手语这种空间性的语言，其记录的手语信息不够完整的，且该语料库采集的是规约手语，不是自然手语。

（4） 中国手语语料库

中国的手语语料库建设还处于初始阶段且手语研究逐渐丰富，以制定通用手语的北京师范大学邓猛教授领头的国家语委、中国残联“十二五”科研规划2013年重大课题的“国家通用手语等级标准研制”项目；复旦大学龚群虎的通用手语语料库研究项目“基于汉语和部分少数民族语言的手语语料库建设研究”；由南京特殊教育师范学院承担的国家语委重点科研项目“国家手语词汇语料库建设”是中国第一个手语词汇语料库，采集了九个地区共六万多个手语词视频，语料具有较强的代表性[7]，但是只限于词语级别的；黄晓晓建立的基于情景的手语语料库[8]，包含个人在家庭学校等场合的日常交流，其手语视频转写采用的是word文本文档作为转写文档，转写的格式没有统一的标准，这使文本语料很难成为格式化的可机读文件。除了政府或残联组织投资建设的语料库外，一些研究者为了研究的需要，也建立了或大或小的手语语料库。

目前手语汉语平行语料库建设的规范性差、缺少系统的理论指导、缺乏具体的评测标准，使得手语语料库建设的质量不一，应用性不好，很难满足语料库语言学发展的需要，很难为语言学研究提供及时、全面、权威性高的语料素材，很难为语言学建设提供强有力的数据支持。

2．2 采集内容

为了采集到高质量的手语语料，本节整理了国内外手语语料库的采集内容（如表 1 所示）、被采集者的选取规则以及采集场景的不同设置的材料，为落实本文的采集内容、被采集者的选取和采集场景设置提供参考。

表 1 手语语料库手语类型、题材、形式

	题材	语料库	体裁
自自然手语	自我介绍	1.澳大利亚手语语料库	叙事
	人生经历讲述	1.澳大利亚手语语料库 2.美国手语语料库	叙事
	学术和社会问题	1.法国手语科林语料库 2.牙买加手语手语采集	叙述、说明
	故事、电影讲述、话题对话讨论	1.澳大利亚手语语料库 2.法国手语科林语料库 3.新西兰手语语料库 4.英国手语语料库 5.德国手语语料库	叙事 说明， 对话，多人对话
	图片故事	1.美国手语语料库 2.新西兰手语语料库	讲述 叙事
	有关聋人和手语的问题	1.美国手语语料库 2.新西兰手语语料库	问答 讲述
	学校手语	1.法国手语传播工程 2.牙买加手语手语采集	叙述，解释，议论文，元语言，描述、对话
	日常生活用语	1.中国黄晓晓基于情景的手语语料库	对话
规规约手语	天气预报相关内容	天气预报德国手语语料库	叙述、说明
	航空系统相关内容	航空信息系统手语语料库	叙述、说明

2．3 被采集者选取

对于被采集者的选取来说，某个语言群体的成员，其语言能力存在差异，手语使用者群

体也不例外。根据Johnston的调查，只有极少数人可以被称作手语的母语使用者。因此最理想的受试是来自第二代聋人家庭的手语使用者。然而在实际生活中，尤其是在较小的聋人群体，研究者往往难以召集到足够数量的理想受试。有鉴于此，Johnston提出了另外一套针对非母语使用者的选取标准，以保证研究的科学性。当中包括：(1)手语的学习年龄不应晚于八岁，以三岁前为最佳；(2)接受聋校教育，以住校生为最佳；(3)每天使用手语；(4)身份上认同聋文化[5]。

2．4 采集场景设置

在场景布置中，如图1所示，荷兰NGT手语视频采集的场景布置中，被采集者和引导者对立而坐，其面前各有一台摄像机，负责采集拍摄对象的手语信息。被采集者和引导者正上方也各自有一台摄像机，负责采集拍摄对象的另一个平面的手语信息。这种场景设置考虑到了手语的空间性，在被拍摄者的头顶和面前都放有摄像机进行拍摄。但是被采集者和引导者的手语采集过程是一个手语对话的过程。如果将其分开，对后面的标注过程是不利的，因为很难理解他们要表达的内容。

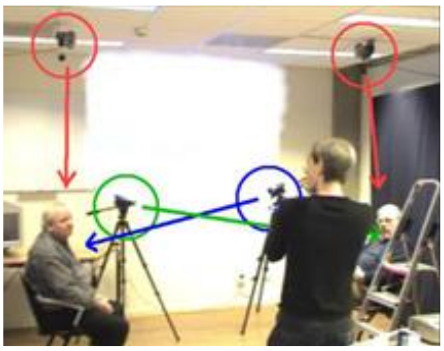

图1 荷兰NGT手语语料库的场景布置[9]

如图2所示，新西兰手语语料库（The Corpus NGT（Nederlands））数据采集时录制对话两人的正面、脸部、由上向下的六个同步视频数据（如图2），新西兰手语语料库的场景设置比荷兰NGT手语语料库的场景设置多了脸部信息的摄取，是手语非手控信息的采集；还合成了被采集者和引导者两人手语对话内容，此类场景布置更适合采集对话形式的手语。其场景布置复杂，要用到六台不同的高清摄像机同时录制，实验室配置花费大。

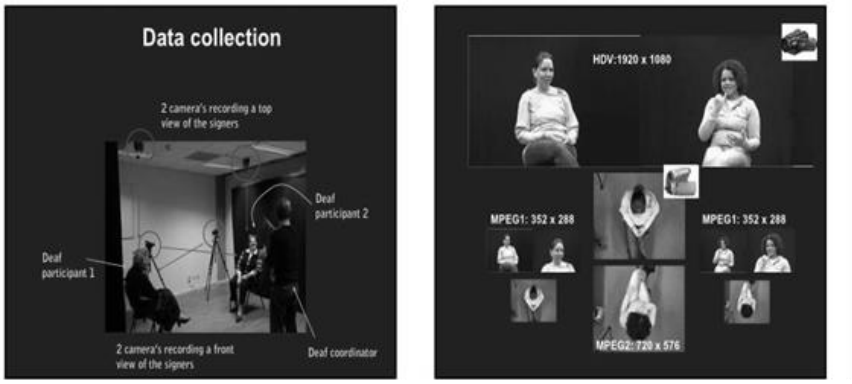

图 2　新西兰手语采集场景设置[10]

2．5 ELAN 工具介绍

ELAN (EUDICO Linguistic Annotator)[11]是荷兰马克斯普朗克研究所为研究心理语言学开发的，目的是为注释和开发多媒体提供良好的技术支持。ELAN 是一个用于对视频或音频文件进行复杂标注的专业工具。使用 ELAN 可以为视频音频添加无限层的标注。标注内容

可以是句子、单词、内容、翻译或者是对视频细节的描述等等。使用 ELAN 对手语视频进行标注可达到事半功倍的效果[12]。

（1）层（Tier）是转写和标注的依托，不同的层可以被赋予不同的标注内容。如注释层、词类层、翻译层等。ELAN中的层可以根据使用者的需求添加。

（2）转写（transcription）指根据音频和视频录入文字或其他符号的操作。以手语为例，是借用汉字和其他字符按照手语顺序记录手语表达的内容和方式，没有翻译加工，记录的是手语表达的信息，并非翻译的汉语句子[13]。

（3）标注（annotation）是针对音频或视频内容转写的文字、注释、翻译、国际音标等，标注包括转写。在ELAN中，标注也指时间段上的时间线，时间段内可以没有转写任何内容。

2．6 视频相似度计算

手语是一种视觉语言，没有书面形式，更多的是通过视频录制的方式进行记录。视频的相似度研究为手语语料相似度研究提供了参考。国内外在研究视频相似度问题时，一部分是提取视频的文本信息，Crawler系统[14]可以从视频的URL和主页HTML文件中提取视频的文本信息，比如字幕、视频的题目、摘要、类别、主题，以及相关的人物信息等。还有视频经过文字检测、文字分割、字符识别，使用OCR软件识别[15]，完成由数字图像到字符编码的转化，最终可以将视频相似度转化为文本相似度的计算。

另一部分，是将视频作为图像进行处理，即关键帧之间的相似度计算，转化成图像的相似度计算。以两个视频间对应帧的平均距离作为相似度，条件是视频帧序列遵守时间顺序[16]。采用常见的颜色直方图进行计算比较，但不是直接将两幅图像的直方图进行比较，而是先将视频的关键帧进行区域划分[17]。

3 手语汉语平行语料库的建设

3．1 本文采集的内容

本文手语汉语平行语料库采集内容为聋人日常生活、学习、工作中自然产生的语料，还有通过实验诱导的方式获取的语料。语料库中已标注语料约 5.12GB，约 80 分钟，约 2400 个平行句对，我们有丰富的手语视频资源正在等待处理。

3．2 本文被采集者选取

被采集人群为根据 Berent 提出的手语双语者分类方案筛选被试，将被采集者分为五类，第一类：出生于聋人家庭的聋人，父母从小使用自然手语与其沟通，在获得第一语言手语后，口语成为第二语言；第二类：出生于健听家庭的聋人，早期接触手语，之后接触口语；第三类：出生于健听家庭的聋人，晚期接触手语；第四类：出生于聋人家庭的健听人，早期从聋人父母处自然习得手语；第五类：健听家庭的健听人，如聋校教师、手语翻译等，他们大多因工作需要，成年后学习手语。以上语料提供者还需满足经常使用手语这一条件[18]。

3．3 本文采集场景设置

如图 3 所示，在本文的手语视频采集的场景布置中，被采集者和引导者的位置如图 3 所示，摄像机 1 的视角是负责拍摄被采集者和引导者的对话，而摄像机 2 的视角是主要负责拍摄被采集者的手语信息。这样做的好处是，即记录了对话内容，也记录了被采集者的信息，在后续对采集的语料进行标注的时候，可以参考对话内容，以保证标注的正确性与可靠性，降低标注者的难度。

图 3 本文手语视频采集场景设置

3．4 标注方法

在本文建立的手语汉语平行语料库中，我们的标注层分为 14 层，包括手控和非手控信息。其中，手控分为主手和辅手，是对主手和辅手的位置、手形、运动信息进行标注，标准参考文献[19]中的内容，如图 4 所示。以及包括词语转写、词语翻译、句子翻译 1、句子翻译 2、句子翻译 3、句子翻译 4，其中词语转写是时间段内手势所要表达的意思，以国家通用手语为准；词语翻译是词语转写层融合非手控信息后的翻译，比如词语转写是“雨”，融合非手控信息就可能翻译成“大雨”或者“暴雨”；句子翻译 1 和句子翻译 2 是由手语使用者来标注，分成两个句子翻译是为了处理句子有歧义的情况。句子翻译 3 和句子翻译 4 是有汉语专家标注，分别对句子翻译 1 和句子翻译 2 进行汉语翻译与校验。非手控包括眉毛、眼睛、嘴巴、身体、头部、眨眼，标准参考文献[18]中的内容。如图 5 所示是使用 ELAN 软件进行标注的示例。

位置		手形		运动	
编号		编号		编号	
A1	中性空间	B1	拇指伸出，其余四指握拳	C1	手由后向前水平移动
A2	胸部	B2	手掌伸直，拇指弯曲贴在掌心，其余四指并齐	C2	手由前向后水平移动
A3	口部	B3	拇指弯曲，其余四指兵器弯向拇指成C形	C3	手前后移动
A4	头一侧或两侧空间	B4	手握拳，拇指搭在食指第二节上或者搭在中指第二节上。	C4	手左右水平移动
A5	额头	B5	中、无名、小三指伸直，分开不并紧，拇指和食指弯曲，拇指搭在食指上。	C5	手向身体一侧水平移动
A6	面前	B6	食、中二指伸直并紧,其余三指弯曲,拇指搭在无名指上。	C6	手一顿一顿向身体一侧水平移动
A7	脸颊	B7	食指伸直,其余四指握拳。	C7	手水平靠近另一手
A8	鼻部	B8	食指伸出弯曲,其余四指握拳,拇指搭在中指上。	C8	手一前一后排列
A9	下巴	B9	食指伸直,中指伸直跟食指成直角,拇指跟中指交叉相搭,其余二指弯曲。	C9	手跨越另一手
A10	头部	B10	拇、食二指伸直分开,形成形,其余三指弯向掌心。	C10	手由一侧向身体正中做弧形移动
A11	肩部	B11	无名指、小指弯曲,拇指搭在无名指上,其余二指并齐,向下弯曲临空压在拇指上。	C11	手由一侧向另一侧做弧形移动
A12	头以上空间	B12	食、中、无名、小四指并齐弯曲,拇指跟食指、中指相抵成空拳。	C12	手拇指和食指套在另一手的拇指
A13	眼部	B13	拇指跟食指相抵,成圆圈,其余三指伸直并紧。	C13	手指跨在另一手的手指上
A14	腮部	B14	拇指跟食指、中指相捏,其余二指弯向掌心。	C14	手由下向上做弧形移动
A15	耳部	B15	食、中、无名、小四指并齐弯曲,手指靠近手掌一节跟手掌成直角,拇指伸出。	C15	手由后向前做弧形移动
A16	颈部	B16	拇指跟中指、无名指相抵,成圆圈,食指和小指伸出。	C16	手由前向后做弧形移动
A17	牙齿	B17	手掌伸直,食、中、无名、小四指并齐。	C17	手一顿一顿向上移动
A18	唇部	B18	食指和中指伸直分开,成形,其余三指弯曲,拇指搭在无名指上。	C18	手一上一下不想接触
A19	太阳穴	B19	食、中、无名三指伸直分开,成形,其余二指弯曲相搭。	C19	手由上向斜下方移动
A20	上臂	B20	中指搭在食指上,成交叉形,其余三指弯向掌心,拇指搭在无名指上。	C20	手由下向斜上方移动
A21	小臂	B21	拇指和小指伸直,其余三指弯向掌心。	C21	手置于身体某部位
A22	腋下	B22	食指和小指伸直,其余三指弯曲,拇指搭在中指和无名指上。	C22	手敲打身体某部位
A23	肩	B23	小指伸直,其余四指弯向掌心。	C23	手揪身上穿的衣服
A24	身体一侧或两侧	B24	拇指和食指张开稍曲不接触,其余三指弯向掌心。	C24	手抚摸身体某部位
A25	肘部	B25	五指张开稍曲。	C25	手在另一手上磨动
A26	腰部以下	B26	拇指伸出,食、中、无名和小指并齐与拇指呈）形。	C26	手自上而下移动
A27	上臂+肘部+小臂	B27	五指撮合在一起。	C27	手自下而上移动

图 4 位置、手形和运动的标注标准[19]

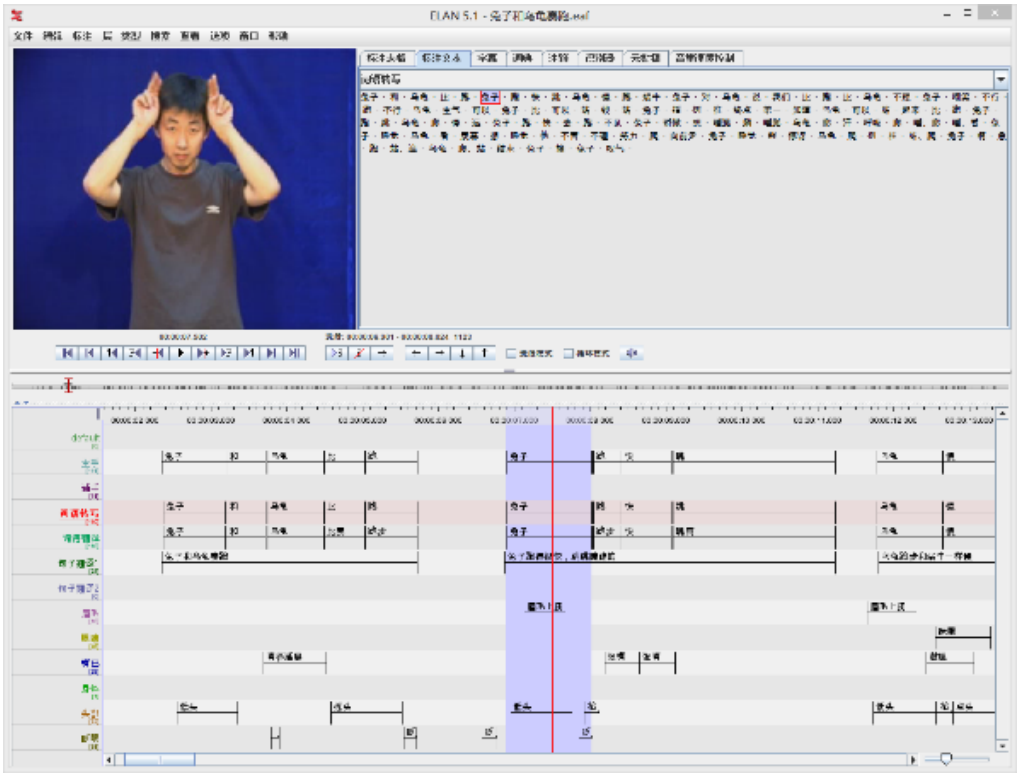

图 5 手语标注示例

3．5 手语语料的预处理

手语语料预处理是整个工作内容的基础，有以下四个方面的内容：手语语料采集、语料的标注、语料专家校验以及语料的更新和存储，手语语料的预处理总体流程如图 6 所示。手语语料的预处理的具体内容如下：

（1）手语语料的采集首先需要确定被采集人和采集内容，接下来按照场景布置要求将拍摄现场搭建好，最后就是对视频的采集与存储。

（2）语料的标注这个过程是由自然手语使用者与汉语专家共同完成的，第一步是将.MP4 文件导入到 ELAN 标注软件中；第二步是按照话题或者固定时间将手语视频进行切分；第三步是建立转写标注层，本文在建立手语汉语平行语料库时，综合了相关研究章节中语料库的优缺点，以及手语汉语平行语料库的用途，增加了翻译部分词级和句子级的平行标注层，减少了空间信息的过多描述，保留了非手控信息的标注，最终确定了十四层的标注层级，接着以手语标注的标准及《国家通用手语》作为参考对手语进行标注。

（3）语料专家校验是首先要校验被采集者及采集内容和手语标注者信息等进行确认，然后根据汉语标注标准和手语标注标准对语料库标注内容进行校验。专家校验就是为了提高语料库的质量，以便使语料库能够更好的建设。

（4）语料的更新与存储将存在的问题进行反馈，由手语使用者和汉语专家将标注转写的语料内容中的任何漏标、误标、多标、标注不统一等情况进行修正更新，形成一套符合标注标准的手语汉语平行语料库。最终，将手语视频的.MP4 文件以及手语语料标注转写语料.EAF 文件进行存储。

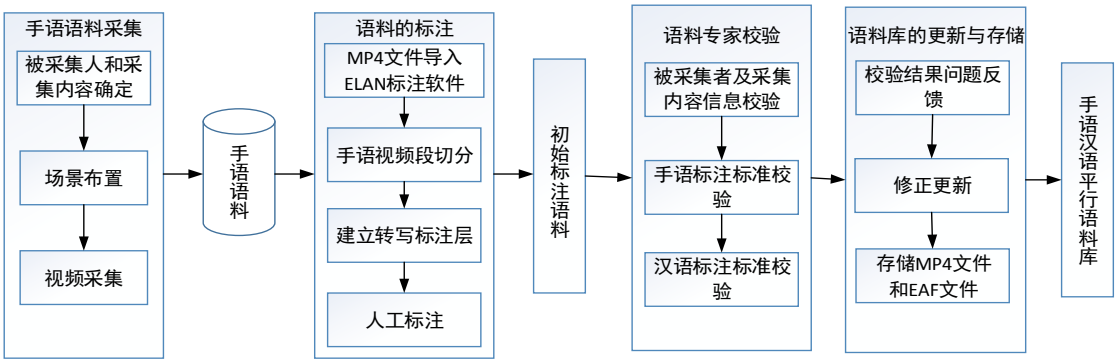

图 6 手语语料的预处理总体流程

4 手语汉语相似度的计算

手语语料相似度的计算有助于语料的去重和手语视频的分类，以及保证标注质量问题等。本文建立的手语汉语平行语料库中词语转写层（词语级别，不涉及语法信息）是对手语视频内容的转写，可以将手语视频转化为文本来处理。我们和相关研究 1.4 中提到的视频的文本信息是不同的，一般视频的文本信息是非常有限的，而且视频语义方面的文本信息很少，所以在视频相似度处理方面是不利的。而本文用到的语料库的词语转写层，是对整段手语视频中语义描述，对其进行手语相似度的计算，为准确性提供了保障。我们使用基于向量空间[20]的余弦相似性来进行手语相似度的计算，还可以用此算法确定标注者的标注质量是否合格。

4．1 算法介绍

向量空间模型的概念最早在上世纪 60 年代被 Salton 等人提出，并很快在文本分类、信息检索等领域得到广泛应用。其定义为，对于待检查手语 B 中的每一个词语，使用 B_i 代表此段手语中第 i 个词语的权重，同样使用 A_i 表示已有手语 A 中的第 i 个词语的权重，从而可以使用 A_i=（a1, a2, …, an）和 B_i=(b1, b2, …, bn) 表示待检查手语 B 和已有手语 A 的词语权重向量。在得到手语的词语权重向量之后，通过余弦相似性算法计算 A_i 和 B_i 两个向量的余弦相似度从而判断待检查手语 B 和已有手语 A 之间的相似度。余弦相似度的计算公式如下：

$$similarity(\text{相似性}) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (1)$$

其中，$\|A\|$ 和 $\|B\|$ 表示向量的模。由于 A_i 和 B_i 均大于等于 0，所以(1)式的值是一个 0 到 1 的值，0 表示两段手语语料的相似度为零，1 表示两段手语语料完全相似。

4．2 算法流程

首先是语料预处理过程，将转写层语料导出成文本，并去除相应的停用词。然后按照余弦相似性算法的步骤，第一对手语视频 A 的转写层和手语视频 B 的转写层的所有词进行列举；第二是计算各自的词频；第三确定各自的词频向量；第四计算两个词频向量对应的夹角。最后就是确定手语视频 A 和手语视频 B 的相似度。

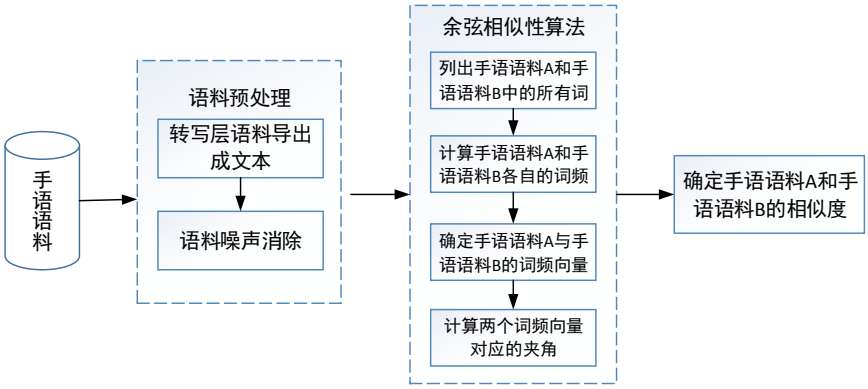

图 7 基于余弦相似性算法的手语相似度计算流程图

4．3 实验结果

本文选取了语料库中的手语语料进行了实验，如图 8 所示，其中手语视频①和手语视频 1 是同一内容、不同被采集者进行手语讲述、经同一转写者处理（手语视频对 2 和②，3 和③，4 和④，5 和⑤处理条件相同，表 2 中的实验结果为图 8 中的结果 2）。通过本文提出的相似度计算方法得到的结果为 0.5066。而手语视频①和其他手语视频进行计算得到的相似度为 0.2376、0.0614、0.2818 和 0.1436，相比较而言，相似度降低。说明同一内容的手语视频比不同内容的相似度高，证明了算法的有效性。实验也将相似度高于 0.5 的手语视频进行专家校验，得到了同样的结果。

表 2 手语视频相似度实验结果

相似度	手语视频 1	手语视频 2	手语视频 3	手语视频 4	手语视频 5
手语视频①	0.5066	0.2376	0.0614	0.2818	0.1436
手语视频②	0.1220	0.5542	0.1709	0.0802	0.1124
手语视频③	0.1630	0.3594	0.6483	0.1619	0.2031
手语视频④	0.2531	0.2811	0.1055	0.6869	0.1193
手语视频⑤	0.2507	0.3463	0.2321	0.2177	0.5259

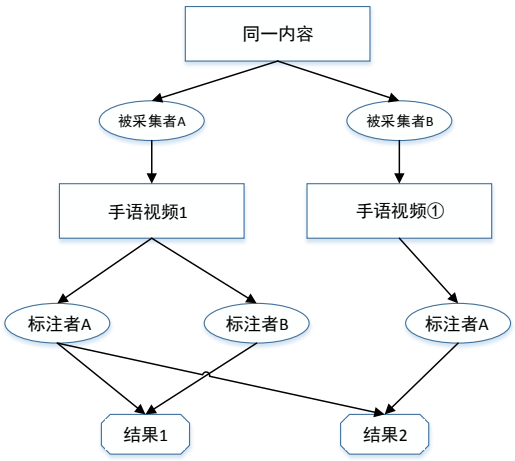

图 8 手语视频相似度计算对比实验

另外，如图 8 所示，标注者 A 和标注者 B 对手语视频 1 进行了转写。我们用上述方法进行计算，计算相似度结果 1 为 0.8958，也就是说明同一手语视频在不同标注者进行转写的情况下，其相似度计算结果有一定的可信度；为了保证语料库的标注质量，此算法可以用作专家相似度测试。标注者 A 为专家，标注者 B 是一般标注者，如果通过标注培训，标注者与专家的标注语料相似度达到相似度阈值，经过专家讨论研究，此阈值设为 0.85，如果计算

结果小于阈值，则不能通过专家相似度测试，重新进行标注培训；如果计算结果大于阈值，则表明次标注者可以进入语料库的标注工作，保证了语料库的标注质量。

实验结果表明，本文用到的基于向量空间的余弦相似度的计算方法是可以解决手语视频中的相似度计算的，也有非常明显的效果。

5 总结

本文总结了国内外手语平行语料库建立、采集内容、被采集者选取和采集场景设置的优缺点，最终确定了本文建设的手语汉语平行语料库的方案；并对 ELAN 软件进行了详细介绍，以及确定了基于 ELAN 的语料库的标注方法和标准；还对手语语料的预处理过程进行了分析和研究；最后基于向量空间的余弦相似性算法的实验效果较明显，能为手语语料去重提供了有力支持，也提高了研究人员管理和检索手语语料的效率，同时保证了手语语料库的质量。未来我们会基于手语汉语平行语料库对手语的机器翻译和各种自然语言知识的进行挖掘研究。

参考文献

[1] 刘俊飞. 手语是一种自然语言[N]. 中国社会科学报,2012-03-26(B04).

[2] 刘超朋. 平行语料库概述[J]. 燕山大学学报(哲学社会科学版), 2007(s1)

[3] 姚登峰, 江铭虎, 阿布都克力木·阿布力孜,等. 中国手语信息处理述评[J]. 中文信息学报, 2015, 29(5):216-227.

[4] 冯志伟.中国语料库研究的历史与现状[J]. Journal of Chinese Language and Computing 2002(1):43-62

[5] Johnston T. W(h)ither the deaf community Population, genetics, and the future of Australian sign language.[J]. American Annals of the Deaf, 2004, 148(5):358.

[6] Bungeroth J, Stein D, Dreuw P, et al. A german sign language corpus of the domain weather report[J]. International, 2000:29.

[7] 赵晓驰, 任媛媛, 丁勇. 国家手语词汇语料库的建设与使用[J]. 中国特殊教育,2017(1).

[8] 黄晓晓. 基于情景语料库的自然手语构词研究[D]. 南京师范大学, 2012.

[9] Crasborn O, Zwitserlood I. The Corpus NGT: an online corpus for professionals and laymen[C]// The Workshop on Crasborn. 2008:44-49.

[10] Inge Zwitserlood,Onno Crasborn,Johan Ros.The NGT Workshop on Sign Language Corpora:Linguis-c Issues 24 July 2009:44-49

[11] Birgit Hellwig ELAN - Linguistic Annotator version 4.5.0[M] 2013-01-07 The latest version can be downloaded from: http://tla.mpi.nl/tools/tlatools/elan/

[12] 李恒, 吴铃. 手语语料库建设基本方法[J]. 中国特殊教育, 2013(3):38-42.

[13] 龚群虎, 杨军辉. 中国手语的汉语转写方案. 2009-2-13.

[14] Swart W D, Asmussen M L, Mccoskey J S. Video and digital multimedia aggregator remote content crawler: US, US 8285701 B2[P]. 2012.

[15] Rodriguez K J, Bryant M, Blanke T, et al. Comparison of Named Entity Recognition tools for raw OCR text[C]// Konvens. 2012:410-414.

[16] Lienhart R, Effelsberg W, Jain R. VisualGREP: A Systematic Method to Compare and Retrieve Video Sequences[J]. Multimedia Tools & Applications, 2000, 10(1):47-72.

[17] 周生, 胡晓峰, 罗批,等. 视频语义相似度网络研究[J]. 计算机应用, 2010, 30(7):1962-1966.

[18] 吕会华, 刘辉, LVHui-hua,等. 基于ELAN软件的中国手语语料库建设研究与实践[J]. 中国听力语言康复科学杂志, 2014(4):298-301.

[19] 姚登峰, 江铭虎, 阿布都克力木·阿布力孜,等. 基于音系学模型的手语理解[J]. 中文信息学报, 2018, 32(1).

[20] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in

Vector Space[J]. Computer Science, 2013.
[21] 陈二静, 姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现, 2017, 1(6):1-11.

通讯作者联系方式：李晗静　北京市东城区永外蒲黄榆二巷甲 1 号 100075　13436569111
tjthanjing@buu.edu.cn
第一作者联系方式：吴蕊珠　北京市朝阳区北四环东路 97 号 100101 17611546889
504648339@qq.com