自然灾害报道语料库的构建及在汉语教学中的应用

沈睿[1] 砂冈和子[2]
[1]日本早稻田大学人间科学学术院
[2]日本早稻田大学政治经济学术院
[1]raymondshenrui@gmail.com [2]ksunaoka@gmail.com

摘要：本文介绍了笔者从约 8 万汉字的 2008 年中国四川大地震的相关报道文章中抽取出特征动词并对其进行了词频统计及词义分类的过程。分析结果表明，自然灾害报道文章中动词的词义分布与一般受灾群众内心压力变化及“创伤后压力恢复模型”相类似。笔者认为，在汉语教学过程中积极地导入此类跨语言跨文化共识，有助于促进不同民族汉语学习者的情感共鸣，从而提高对外汉语教学的质量。
关键词：自然灾害 文本语料库 情感共鸣 特征词抽取 词义分类

Construction and Application of a Text Corpus of Newspaper Articles about Disasters in Chinese Education

Raymond SHEN[1] Kazuko SUNAOKA[2]
[1]Faculty of Human Sciences, Waseda University
[2]Faculty of Political Science and Economics, Waseda University

Abstract: In this paper, we introduce a procedure of feature words extraction from 80000 word-news reports about 2008 China Sichuan earthquake, words frequency analysis and meaning classification. It turns out that the distributions of verbs in disasters news reports are relevant to stress changing process of disaster victims and the “disaster-related stress recovery model”. We consider that shared cognition among multi-linguistic and multi-cultural backgrounds like the “disaster-related stress recovery model” helps arousing emotional resonance, and enhance the motivation of Chinese learning.
Key words: natural disasters, text corpus, emotional resonance, features extraction, meaning classification

1 引言

近年来，随着全球规模的自然灾害频频发生，自然灾害发生时的信息处理受到了各方面专家的重视。在信息全球化的今天，如何高效地运用先进的信息处理技术来解决人类在遭遇自然灾害之后所面临的诸如灾后复兴之类的各种问题，也已成为令人注目的课题。日本语言处理学会在 2012 年的年会上还特设了灾害发生时的语言信息处理主题。我们知道，对于大规模自然灾害的灾后复兴来说，除了短期的物资援助外，诸如防止流言危害之类的长期社会支援也是相当重要的。保坂隆还提出受灾应该分成“直接受灾”和“因流言受灾”(保坂隆，2001)，其重要性可见一斑。

在传统的语言教学中，由于语言教学和民族文化的密切关系，融时事于教学的方法一直被广泛运用。在笔者担任的对外汉语教学课程中，积极向学生们介绍各种中国当代社会的时事也是教学的重要内容之一。2008 年四川大地震和 2011 年日本东北大地震以来，在对受灾群众所经受的苦难感同身受的同时，日本的大学课堂上师生们也对于灾后复兴的支援表示出越来越多的关心。笔者所担任的汉语教学课上也进行了相关的探讨。笔者发现，并不是受关注越多的灾害报道就越能引起读者与受灾群众间的情感共鸣。譬如，在阅读了

关于由福岛第一原子发电站事故引发出的在中国发生的抢盐风波的报道[1]，日本学生们都露出了讶异的表情。即使有关亲情的羁绊这样能够引起共鸣的话题，大家的感想也都各不相同。譬如，对于由中国学生推荐的有关在四川大地震中母亲牺牲生命保护自己孩子的报道[2]，日本学生大都认为有媒体炒作的成分在内。又如对于关于为在四川大地震中失去孩子的父母特设的“再生育政策”的报道等(砂冈和子，2012)。

可见，在语言教学中，如何正确引导学生接受相对客观的自然灾害信息并激发他们的情感共鸣，是一个十分重要的课题。因此，在本文中笔者尝试采集了 2008 四川大地震以来 3 年内的相关新闻报道并构建成一个小型样本的文本语料库，并通过文本挖掘的手法对报道文章进行了词义分类、特征词抽取等一系列的分析，最终与“创伤后压力恢复模型”进行对照，得出了两者基本一致的结论。相信通过对不同的有关自然灾害的文本信息的共性研究，并有效地在汉语教学中利用这类语料库，不但能够加强汉语学习者对汉语本身的把握和理解，还能够促进与汉语学习者在诸如灾后复兴之类问题上的跨地域跨文化的情感共鸣，从而激发汉语学习者对汉语学习的积极性及对汉语文化的认同。

2 相关研究

在传统的汉语教学中，教师们经常会鼓励学生阅读报刊杂志。因为比起普通的教科书来说，报刊杂志中的汉语更接近自然语言，且内容也较生动并贴近生活。通过对这些类别的语料的阅读和理解，学生们能够接触到更自然、实用性更高的汉语。通常来说，在有诸如发生自然灾害之类的非常规事件之时，铺天盖地席卷而来的报道信息往往更容易激发学生们的兴趣。这时作为教师所要做的是：一， 尽可能多地收集相关信息，并根据信息的真实性进行有效的选材。二，鼓励学生大胆地对这些信息进行吸收反馈，并争取能够在文化层面上加深认识。而以上两个问题，在实际操作过程中均有不小的难度。

关于第一个问题，要收集大量的信息数据，光靠手工采集是远远不够的，且效率也相当低下。且在成功采集到所需信息之后，还必须对信息来源加以核实从而过滤掉一些真实性比较低，甚至是流言谣传的内容。随着计算机技术的发展和网络的普及，在计算机工程学及自然语言处理的领域中，已有不少针对这个问题的研究。宫部(宫部，2011)等利用 Twitter[3]的 API[4]自动攫取了在东日本大地震发生后(2011 年 3 月 11 日 16 时 10 分开始)20 天内包含关键词“地震”的微博(总计 1612074 条)，并同时抽取出每条微博的发送时间，发送人、发送地及发送内容等信息。在该研究中，作者通过微博的各种附加信息的时间序列分析，论证了两个假设。1. 发送地域不同，对微博进行回复或引用等的使用倾向也各不相同。2. 从受灾严重的地域发送出的信息逐渐向其他地域移动扩散。除了利用自动攫取收集信息之外，宫部(宫部，2012)等还通过支持向量机(Support Vector Machine, SVM)的学习方式制作了辟谣信息分类器并构建成辟谣信息的云存储。在该研究中，为了提供分类器的学习数据，作者在 2010 年 3 月(东日本大地震发生一年前)的 Twitter 微博和 2011 年 3 月(东日本大地震)包含关键词“地震”的 Twitter 微博中随机抽取出包含关键词“デマ”(日语中流言、谣传的意思)的各 1000 条微博并构建成语料库。根据是否为辟谣信息对每条微博进行分类(是的赋予+1 值，不是的赋予-1 值)，然后作为 SVM 的学习数据(地震前 1000 条中+1 的有 187 条，地震后 1000 条中+1 的有 602 条)。在机器学习的过程中，关键词的前后文、形态素数量、是否含有 URL 链接及是否有引用(RT@)等 4 项被作为特征量。最后，为了验

[1]日本地震，中国抢盐，皇帝不急太监急？《南方报网》2011 年 3 月 17 日

[2]母亲用身体护住婴儿，婴儿被救出时仍熟睡，《网易》2008 年 5 月 17 日

[3]Twitter 主页： https://twitter.com/

[4]Twitter API 主页： https://dev.twitter.com/

证 SVM 的分类精度，作者还同时使用其他算法(贝叶斯分类器： Naïve Bayes, 表决感知器： Voted Perceptron, 决策树： Decision Tree, K 最近邻： K nearest neighbor)，最终发现，还是 SVM 的分类精度最为理想。宫部等在该研究中提倡的机器学习的手法，将大幅度提高辨别并筛选出真实信息的效率。不难发现，利用此手法，除了微博之外，对于其他诸如新闻报道等传递重要信息的文本，也同样行之有效。

关于第二个问题，与只是在技术上需要突破的第一个问题比较来说更为困难。众所周知，文化差异历来都是横亘在语言教学中的障碍。在教授汉语的同时，向汉语学习者传播中国文化也是重要的课题。如何从语料中总结出不同母语背景的人都能理解和掌握的共性规律，正是解决第二个问题的关键所在。已有不少学者致力于类似的研究。Qu(2009)对国内知名网络社区天涯社区[1]在 2008 四川大地震发生之后一星期内有关地震的帖子(最终取样 2266 条)进行了定性和定量的分析。在构建帖子分类表图的基础上，作者把所有相关帖子分成非独立的四类：消息相关(information-related)，意见相关(opinion-related)，行动相关(action-related)和情感相关(emotion-related)。此外，在 Qu(2010)中，作者对 2010 玉树大地震发生后 48 天内新浪微博[2]上关键词“玉树+地震”或“青海+地震”的微博(总计 135918 条)的内容、话题倾向及消息传播过程进行了一系列分析，从而把握并总结了中国网络用户在利用微博对自然灾害做出反应时的特点。长尾（2011)在文章中介绍了地震灾害发生时的新闻报道内容变化的分析方法，从而总结出了防止因媒体报道所产生的流言危害的对策。在该研究中，作者通过形态素解析及关联规则探悉等文本挖掘的手法，对连续两个月中的报道内容的变化进行了定量的分析，从而推测出了造成不良影响的信息淡化的大致时期。佐藤(2007)提出用 TFIDF 的手法对有关灾害、危机的语料进行关键词自动抽取，并构建了一个由 19 种有关灾害、危机的网络新闻报道的语料库用来进行效果验证。其结果是，关键词自动抽取的精度达到了约 8 成。

佐藤(2007)还对有关灾害与危机的语料收集进行了如下的论述[3]：

> 在研究防灾的领域里，除了灾害的“自然属性”之外，诸如受灾方、救灾方及受灾地以外的人的应对以及着眼于复兴之类等问题的“社会属性”也得到了越来越多的关注。针对灾害的自然属性的研究中，我们通过实时监测地震、降雨量、水位等来迅速把握现况，并对累计的庞大观测数据进行解析，从而提高防灾能力。而针对灾害的社会属性，我们需要迅速收集各种相关信息资料，对灾害和危机发生时的现况进行实时且正确的把握，并把过去的有关灾害和危机的信息资料以更容易取用的方式累积起来。
>
> 有关灾害和危机的信息资料多以文本形式被保存下来，常见的有如下几种：受灾和应对的速报，新闻报道，有关部门公开的网页、记录和访谈、发行刊物。在灾害发生时，这些信息资料有助于把握了解受灾情况，而在平时，还可以在分析过去发生的灾害与危机时的应对或危机模拟训练中用来对灾害进行还原和再现。

从以上论述来看，积累有关自然灾害的信息数据的重要意义不言而喻。笔者还发现，以上介绍的已有相关研究大都采用了一个相似的研究流程，即：数据采集(自动攫取，自动存档)，特征词抽取(关键词抽取)，定性定量分析(时间系列分析)。此外这些研究都有一个共通的不足之处，即对语料信息的时间跨度的应对不够充分。宫部(2011)中数据的时间跨度为 20 天，宫部(2012)中为 1 个月，Qu(2009)中为 8 天，Qu(2011)中为 48 天，长尾（2011）

[1] 天涯社区主页: http://www.tianya.cn/
[2] 新浪微博主页: http://www.weibo.com/
[3] 笔者根据日文原文做了翻译。

中为 2 个月，佐藤(2007)中最短的为 1 个月，最长的也不超过 1 年半。要充分有效地进行时间系列分析，就必须收集到时间跨度大且数据量平衡的语料信息。最后，在这些信息工程学领域研究中所使用的词汇中，多以易于反映时间内容的事件名词为主。长尾（2011)指出，以高频词（事件名词）为线索进行抽取的手法容易把握且抽取成功率高，但是由于词义范畴的划分过于简单，导致了无法进行词义分类。

综合参考以上相关研究及其不足之处，笔者为本研究的开展总结出以下四点：

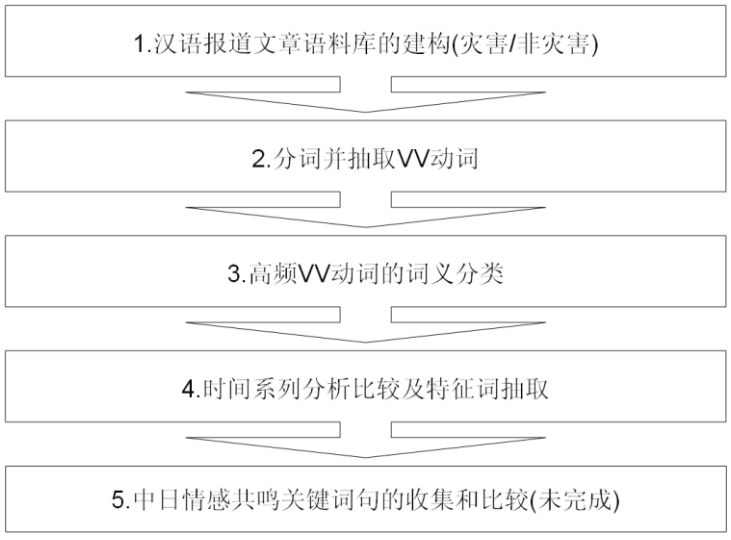
1. 收集一定量的有关自然灾害的文本数据，并构建成文本语料库。
2. 保证收集的语料信息的时间跨度足够大(实际跨度达到 3 年)，且保持各时间段的数据量平衡。
3. 遵循传统的研究流程，即：数据采集，特征词抽取，时间系列分析。
4. 以动词（含形容词）为线索进行特征词抽取并进行词义分类。因为动词（含形容词）的数量和用法比较稳定，不会受到时事动态变化的影响，不像名词那样不稳定。

具体的细节将会在之后的章节中加以详细介绍。

3 方法

笔者先收集了一系列汉语灾害报道文章来构建“自然灾害报道语料库”，并为便于特征词抽取而附加了各种信息。然后从该语料库中抽取出动词（以下简称 VV[1]），并对词频及词义概念进行了分析。为了判断该手法的合理性，笔者还另外收集了一些普通的汉语报道文章并同样把 VV 抽取出来进行了定量对比。在参照第 2 章中总结的相关研究的基础上，笔者认为作为实词的 VV，能够较稳定地反映出人类的各种活动，因此选择其作为分析对象。

【图 1】 整体研究工序



参照【图 1】中所显示的工序，对本研究的手法进行顺序说明。本章的 1 到 4 节分别对应【图 1】中的 1 到 4 的工序,【图 1】中的工序 5:中日情感共鸣关键词句的收集和比较(未完成)将在第 4 章中进行详细介绍。

[1] VV 是除去 VA 谓词性形容词，VC copula 动词“是”，VE 作为动词“有”的复合动词的总称。请参照中文依存句法分析器 CNP 的词性标注说明和 Chinese Tree Bank-CTB 的规格。

3.1 汉语报道文章语料库的建构(灾害/非灾害)

笔者首先收集了两种汉语的报道文章。一种是四川大地震的相关灾害报道文章（以下简称“四川”或“S”)，方法： ①传达受灾群众的心声及行动的文章，②从对政府的官方报道持保留意见的报章杂志中提取，③以保持地震发生后每个时间段的字数的平衡为前提，主观地选取出来。文章的出处请参照砂冈和子(2012)。另一种是在与“四川”语料库保持同时期及同篇幅为前提。最终笔者成功收集了四川地震发生后不久，半年后，1(0.5)年后，2 年后，3 年后，总计约 8 万汉字的汉语报道文章并构建成库（表 1)。

【表 1】 两种中文报道文章的详细信息

四川地震报道文章			人民日报报道文章		
发表时间	篇数	字数	发表时间	篇数	字数
地震发生后不久	5	18200	2008 年上半年	20	17807
发生后 6 个月	5	16000	2008 年下半年	16	15774
发生后 1 年	5	15400	2009 年	9	15550
发生后 2 年	4	15000	2010 年	13	14774
发生后 3 年	5	16400	2011 年	18	16488
总计	24	81000	**总计**	76	80393
1 篇 1000－8000 字			1 篇 147—3768 字		

3.2 分词并抽取 VV 动词

接着笔者利用中文依存句法分析器(a CNinese dependency Parser，简称 CNP)[1]对上述的两种汉语报道文章进行分词并抽取出 VV 类词语。该分析器是由高度语言信息融合论坛 ALAGIN[2]开发，并免费为科研提供服务。该分析器使用了由宾夕法尼亚大学 Linguistic Data Consortium(LDC)[3]所提供的中文树库 Chinese Tree Bank4.0，5,0 和 6.0[4]，使用的学习数据也是由该机构提供的 Chinese Gigaword[5]语料库。根据 Chen(2009)的报告，该分析器的分词精度达到 97.79%(F 值)，词性标注精度达到 93.60%(F 值)，依存句法解析精度达到 89.80%(F 值)。该分析器的性能在现有的中文句法依存分析器里名列前茅。【表 2】中是部分分词和词性标注的结果，使用的是 Chinese Tree Bank-CTB6.0 中所采用的 CoNLL 2007 格式[6]。在对乱码或专有名词的错误分析进行了手工删除及校对[7]后，最终笔者分别从“四川”语料库和“人民”语料库中成功抽取出 8461 和 8668 个 VV 类词语（两者数量基本持平)。汉语中常用动词多由添加补语及复合等形式所派生，而在 CNP 的词库中诸如“上”和“上去”之类的 VV 大都以独立形式存在。因此笔者还对同义词进行了整理和整合使得 VV 词语的数量

[1] 中文依存句法分析器 CNP: http://alaginrc.nict.go.jp/cnp/index.html
[2] 日本高度语言信息融合论坛: ALAGIN:http://alaginrc.nict.go.jp/index.php
[3] Linguistic Data Consortium-LDC: http://www.ldc.upenn.edu/
[4] Chinese Tree Bank: http://www.cis.upenn.edu/~chinese/
[5] LDC-Chinese Tree Bank-CTB: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2003T09
[6] CoNLL 2007 format: http://www.cnts.ua.ac.be/conll2007/
[7] 手工删除的部分只占整体的 3%左右。

减少了一半左右。

【表 2】 部分 CNP 解析结果

ID	FORM	LEMMA	GPOS	PPOS	PPOSS	HEAD	DEPREL	PRED	ARG	TIME
1	有	有	VE	VE	-	34	VMOD	-	-	0
2	人	人	NN	NN	-	6	SUB	-	-	0
3	在	在	P	P	-	6	VMOD	-	-	0
4	乐观	乐观	VA	VA	-	5	DEP	-	-	0
5	地	地	DEV	DEV	-	6	VMOD	-	-	0
6	估计	估计	VV	VV	-	1	VMOD	-	-	0
7	，	，	PU	PU	-	34	P	-	-	0
8	“	“	PU	PU	-	34	P	-	-	0
9	说	说	VV	VV	-	11	VMOD	-	-	0
10	不	不	AD	AD	-	11	VMOD	-	-	0
11	定	定	VV	VV	-	34	VMOD	-	-	0
12	再	再	AD	AD	-	13	VMOD	-	-	0
13	走	走	VV	VV	-	23	SBAR	-	-	0
14	一	一	CD	CD	-	15	AMOD	-	-	0
15	段	段	M	M	-	22	NMOD	-	-	0
16	就	就	AD	AD	-	18	VMOD	-	-	0
17	碰到	碰到	VV	VV	-	18	VC	-	-	0
18	出来	出来	VV	VV	-	19	SBAR	-	-	0
19	的	的	DEC	DEC	-	22	NMOD	-	-	0
20	家	家	NN	NN	-	21	DEP	-	-	0
21	里	里	LC	LC	-	22	NMOD	-	-	0
22	人	人	NN	NN	-	13	OBJ	-	-	0
23	了	了	SP	SP	-	11	VMOD	-	-	0
24	！	！	PU	PU	-	34	P	-	-	0
25	”	”	PU	PU	-	34	P	-	-	0

3.3 高频 VV 动词的词义分类

由于利用机器自动处理 VV 动词的词义分析及分类比较困难，笔者采用了启发式的方法进行。

为了在大体上把握特征词的倾向，以“四川”和“人民”各个时间段的前 100 位高频 VV 为对象，就其词频及排位的变化进行了观察。结果发现，对于两个语料库的前 100 位高频 VV 中的非共通 VV，两个语料库有着显著的差别。接着笔者优先对“四川”语料库的非共通 VV 进行了词义分类。

在进行词义分类时，笔者采用了日语的分类词汇表(“国立国语研究所”，2004)的分类体系。其中和汉语 VV 吻合的部分的日语分类号可以通用。【表 3】中是部分高频 VV 的词义分类结果。通过对从“四川”和“人民”中抽取出的 VV 进行高精度的分类，从而使得大体上的概念归属关系得到了明确。

【表 3】 高频 VV 动词的词义分类

sichuan				
VV	分類			freq(t=0.
拍摄	用,活動,事業,練り・塗り・撃ち・録	2.3851	M	5
复课	用,活動,待遇,教育・養成	2.3640	K	4
可信	用,活動,心,信仰・宗教	2.3047	F	3
叫	用,活動,心,声	2.3031	F	11
喜欢	用,活動,心,好悪・愛憎	2.3020	F	4
背	用,活動,心,学習・習慣・記憶	2.3050	F	5
怕	用,活動,心,恐れ・怒り・悔しさ	2.3012	F	3
负责	用,活動,心,自信・誇り・恥・反省	2.3041	F	7
哭	用,活動,心,表情・態度	2.3030	F	10
发现	用,活動,心,見る	2.3091	F	6
见	用,活動,心,見る	2.3091	F	7
睡	用,活動,心,飢渇・酔い・疲労・睡眠	2.3003	F	6
遇难	用,活動,生活,人生・禍福	2.3310	H	6
住	用,活動,生活,住生活	2.3333	H	12
打工	用,活動,生活,労働・作業・休暇	2.3320	H	4

3.4 时间系列分析比较及特征词抽取

经过 3.3 节中的词义分类之后，笔者接着通过手工方法对非共通 VV 的前后同现词进行了词频统计。在对前 100 位以内的非共通动词进行了词义分类后，发现了以下的特征。

（1）“四川”语料库中有关“行动”和“心理活动”的动词词频要比“人民”语料库中的高很多（【图 2】中 D,F）。特别是在地震发生后不久至地震发生后两年内十分显著。譬如（/后的数字表示地震发生后的经过年数），“得知/0”，“找到/0”，“发现/0.5”，“哭/0.5”，“叫/0.5”，“考虑/1”，“回忆/2”，“找/3”。不难发现，有关“行动”和“心理活动”的动词数量的增加是灾害报道的特征之一。此外，虽然排位靠后，但诸如“衣食住”(【图 3】中 H),“生死”（【图 2】中 N）分类中的“死，死亡，亡，去世/0-3”，“怀孕，生育，生/1-3”，“时间”（【图 2】中 O）分类中的动词都以较高频率出现。在“人民”语料库的前 100 位 VV 里并没有出现的这类动词，表明了对于受灾群众而言，情感上的接受及应对和灾后的重建及复兴是当务之急。

（2）“人民”语料库中主要以“活动，行为”，“交际”，“经济活动”，“事业”（分别为【图 3】中 I,J,L,M）等描述日常的社会生活的动词居多（【图 3】）。由于在“人民”语料库中经常出现有关刑事案件及维持治安的新闻报道，所以与这类题材相关的分类“教育救援”（【图 3】中 K）的动词（譬如“释放”，“抓获”，“安置”，“维护”，“安排”之类）反比“四川”语

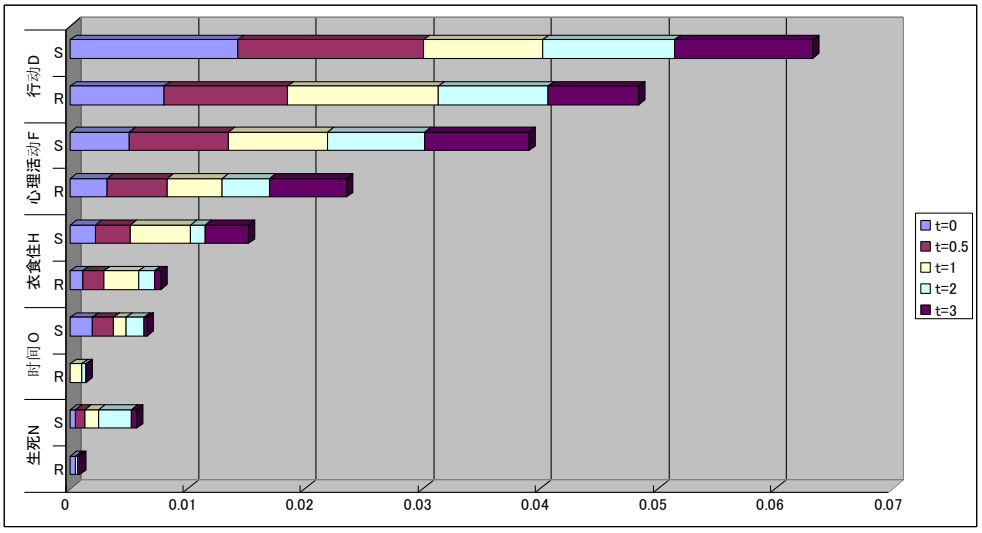
料库中出现的多。但是，“四川”语料库中的“教育救援”类动词中有和教育复兴密切相关的“复课”“复学”，因此，笔者认为在抽取特征词的同时如“教育救援”这样排名靠后的分类也是需要予以考虑的。

3.5 创伤后压力恢复模型

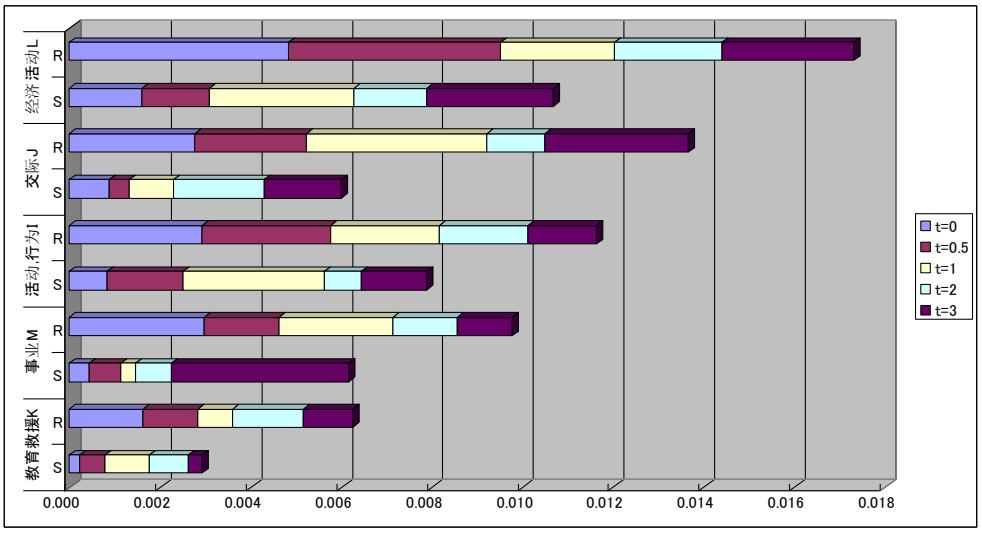
通过比对普通报道文章，我们总结出了灾害报道文章中几个和有关词义的特点，但是，究竟“四川”语料库中动词的整体特征又是怎么样的呢？保坂隆(2011)认为：一般而言，人类在遭遇危机时，会经过“受打击/冲击”，“战略/防御性避退”，“接受现实” 及“适应/

习惯”这 4 个阶段来克服危机。同样，受灾群众内心的平复也与其相类似，会经过“***茫然失去自我期***”，“***蜜月期***”，“***幻灭期***”及“***复兴期***”这四个阶段。把“四川”语料库中各个时间段的高频动词罗列出来后发现，中国受灾群众身心所受到的创伤后压力及其恢复过程基本和保坂模型的四个阶段相吻合（【表 4】）。

在“四川”语料库中，从第二，第三阶段中诸如“埋”，“死”，“遇难”等描述受灾之后“受打击/冲击”时期的动词显示出的高频率中我们可以知道，受灾群众们内心的平复需要一段时间。林春男(2001)指出，灾后的复兴支援可细分为“人命救助活动（72 小时内）”，“保证社会流通的安定活动（100 天）”，“社会储备资源的再建设（10 年）”，“信息资源管理（物资流动）（0~10 年）”，并需要得到同时开展。通过接收在报道文章中传达出的来自受灾群众内心压力的信号，我们能够更有效地进行对其早期恢复的支援活动。



【图 2】（ 四川/S >人民/R）词义分类后的词频比较



【图 3】(人民/R >四川/S）词义分类后的词频比较

【表 4】危机四阶段模型和“四川”语料库中的高频 VV 动词[1]

遭遇危机的经过	创伤后压力恢复经过	救助活动的重心	
第一阶段	茫然失去自我期（受打击/冲击）	生命救助	信息/资源管理
第二阶段	蜜月期（战略/防御性避退）	社会流通的安定	
第三阶段	幻灭期（接受现实）	社会储备资源的再建设	
第四阶段	复兴期（适应/习惯）		
“四川”语料库的时间分期	各时期高频动词	共通高频动词	
0-0.5年	“活动”“等候”“打饭”“挂职”	“找”“逃”“跑”“送”	
1年	“等待”	“读”“哭”“怕”	
2年	“怀孕”“结婚”“搬”“抑郁”“跳楼”	“希望”“失去”“打工”	
3年	“死”“失踪”“遇难”	“笑”“喜欢”	

【表 5】课堂实践中收集到的信息样本

	汉语课	日语课
报道文章	<孩子你慢慢来—汶川地震灾后学校重建纪实> 2010年5月,南方周末	「分享复兴支援的苦痛」 2011年7月9日,朝日电子新闻
选取理由	本来以为中国有着跟日本不同的国民性， 但是看了这篇报道，觉得和日本的对待态度也十分相似。	对从地震发生后到渐渐被人遗忘的过程中， 由普通的大学生发起的强烈控诉的言辞感到印象深刻。
关键词句	我们一定要战胜困难	当事人意识
	看到孩子们跳舞，我们心里好受了很多。	复兴遥遥无期
	多难兴邦	请不要觉得和你们无关

4 语料库应用及课堂实践

利用“四川”语料库，笔者在教学课上进行了课堂实践。【图 1】中的工序 5 正是利用在课堂实践中收集到的信息进行汉语和日语中能够激发情感共鸣的关键词的比较。

在汉语课(学生多为日语母语者)中，笔者要求学生们从“四川”语料库中任意选取一篇汉语报道文章。在日语课(学生多为汉语母语者)中，笔者要求学生们任意推选一篇有关 2011 年日本东北大地震的日语报道文章。同时还要求学生们总结出选取的理由，报道文章的内容概要并选出 3 个学生们认为能够激发情感共鸣的关键词或句。所有的内容都在 BBS 上进行提交，再由笔者进行翻译和整理。【表 5】中是部分样本。

通过比对中日关键词句，我们能够发现跨语言跨文化间激发情感共鸣的共同特征，并把其导入实际的汉语教学中，从而更有效地促进不同文化背景的学习者对中国社会及文化的理解。由于情感分析的过程比较复杂，该工序仍在进行中。

5 总结和今后的课题

本文中介绍了笔者从约 8 万汉字的 2008 年中国四川大地震的相关报道文章中抽取出特征动词并对其进行了词频统计及词义分类，还通过比对同时期且同篇幅的普通《人民日报》的报道文章来抽取出灾害发生时动词的出现特征。分析结果表明，自然灾害报道文章中动词的词义分布与灾害发生时的“创伤后压力恢复模型”及受灾群众内心压力变化相类似。笔者认为，在汉语教学过程中积极地导入此类跨语言跨文化共识，有助于促进汉语学习者的情感共鸣，从而提高他们的汉语学习兴趣。

作为今后的课题，本文中介绍的研究手法还需要做一定的改进。

在精度方面，需要把现有的语料扩大至少 10 倍，即 80 万字，并利用 CNP 中句法依存关系的解析功能，以特征词的同现词为线索进行进一步的文本挖掘分析。另外，还需要对 CNP 的自动解析结果进行手工修正。在效率方面，为了能够更快更好地收集语料，需要实现自然灾害相关语料的定期自动攫取和归档。最后还需要实现对特征词词义分类的自动化。

[1]笔者参考保坂隆(2011),林春男(2001)的研究内容加工而成。

致谢

在本文的执笔过程中得到了中国北京语言大学的杨尔虹教授、刘鹏远博士、邹红建先生提供的《人民日报》的数据及大力协助。此外还由日本 Advanced LAnGuage INformation forum(ALAGIN)机构提供了中文依存句法解析器 CNP。在此表示衷心感谢。

本研究作为日本文部科学省的科研项目（基础 C 课题号：22520445,研究代表人：砂冈和子）及以罗凤珠代表的“历代语言知识库建置计划”九十八年度蒋经国国际学术交流基金会的科研项目的成果的一部分，得到了大力资助。

在本稿的修改过程中，得到了第八届中文电化教学国际研讨会与会的各位老师和专家的关注及提出的宝贵意见，在此表示感谢。

参考文献

[1] 保坂隆. 災害ストレス:直接被災と報道被害. 角川書店, 2001

[2] 砂岡和子, 沈睿. 災害報道文の特徴語抽出. 日本言語処理学会第 18 回年次大会（NLP2012）論文集, 2012: 579-582

[3] 宮部真衣, 荒牧英治, 三浦麻子. 東日本大震災における Twitterの利用傾向の分析. 情報処理学会研究報告, グループウェアとネットワークサービス研究会, 2011-GN-81, No.17: 1-7

[4] 宮部真衣, 梅島彩奈, 灘本明代, 荒牧英治. 流言情報クラウド:人間の発信した訂正情報の抽出による流言収集. 日本言語処理学会第 18 回年次大会（NLP2012）論文集, 2012

[5] Qu, Y., P.F. Wu, X. Wang. Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthquake. Proc. HICCS, 2009

[6] Qu, Y., Huang, C., Zhang, P., et al. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. In Proceedings of the ACM, 2011 conference on Computer supported cooperative work (CSCW ’11): 25-34

[7] 長尾光悦, 大内東. 観光地に対する風評被害の変遷と対応分析. 人工知能学会誌(特集:観光と知能情報), 26(3): 264-271, 2011

[8] 佐藤翔輔, 林春男, 牧紀男, 井ノ口宗成. TFIDF を用いた災害・危機に関する言語資料体からのキーワード自動抽出手法の外的妥当性の検証. 地域安全学会論文集, 2007 (9): 65-74

[9] Wenliang Chen, Jun’ichi Kazama, Kiyotaka Uchimoto, Kentaro Torisawa. Improving Dependency Parsing with Subtrees from auto-Parsed Data. EMNLP 2009

[10] “国立国語研究所”編. 分類語彙表（増補改訂版）. 大日本図書, 2004

[11] 林春男. 率先市民主義:防災ボランティア論講義ノート. 晃洋書房, 2001