Gromoteur 软件使用手册

Copyright@ Professor Jiang's Linguistic Group, Xi'an Jiaotong University Gromoteur 软件是由巴黎第三大学 Kim Gerdes 教授开发的一款集数据 统计、词性标注、语义划分于一身的语料处理软件,为方便大家使用 此软件,灿灿和九菊特做此使用手册,供大家参阅。

1. 语料格式

该软件支持格式为 UTF-8, 具体转换方式为:

1) 首先,将 Word 格式另存为 TXT 格式,在转换前注意将 word 中 所有文字合并为一段(此方法适用于做单个语料研究)。如果想对比 两个或以上语料,则只需要将每段开始的空格去掉,将需要对比的语 料单独成段即可,见下图。

单个语料研究文字格式

举行。来自四十多个国家的千余名断裂领域的专家、学者参加会议。西安交大孙 军教授应邀在大会上作了50分钟大会特邀报告,西安交大校友、美国工程院院士 ;锁志刚教授、高华健教授出席会议。国际断裂大会每四年举行一次,这是这一具 有重要影响的国际盛会首次在中国召开。其内容涉及断裂学科的方方面面,如断 裂力学、断裂与疲劳机制、断裂实验技术、工程应用等,是材料断裂学科领域最 高级别的国际学术会议。此次大会由中国力学学会、香港断裂组、中国机械工程 学会、中国材料研究学会、中国腐蚀与防护学会、中国航空学会、中国金属学会、 中国结构完整性联盟联合承办。此次大会云集了相关领域内国际上最具影响力的 一批学术大师,例如J积分的提出者J.R.Rice教授、Paris公式的提出者P.C. Paris 教授、HRR场的提出者之一 John W. Hutchinson 教授,美国国家自然科学 基金委员会前任主席 Subra Suresh 教授、西安交大校友、美国工程院院士锁志 刚教授、澳大利亚两院院士 Y.W. Mai 教授等。我国国家自然科学基金委主任杨卫 院士、香港科技大学孙庆平教授、西安交大孙军教授等与会,并与西安交大校友、 美国工程院院士高华健教授等10位国外著名学者分别作50分钟大会特邀报告。大 会共设54个主题论坛和15个专题研讨会。|

两个语料的对比研究

第13届国际断裂大会(ICF-13)于6月16日至21日在北京国家会议中心(CNCC)隆重举行。来自四十多个国家的千余名断裂领域的专家、学者参加会议。西安交大孙军教授应邀在大会上作了50分钟大会特邀报告,西安交大校友、美国工程院院士锁志刚教授、高华健教授出席会议。国际断裂大会每四年举行一次,这是这一具有重要影响的国际盛会首次在中国召开。其内容涉及断裂学科的方方面面,如断裂力学、断裂与疲劳机制、断裂实验技术、工程应用等,是材料断裂学科领域最高级别的国际学术会议。此次大会由中国力学学会、香港断裂组、中国机械工程学会、中国材料研究学会、中国腐蚀与防护学会、中国航空学会、中国金属学会、中国结构完整性联盟联合承办。 此次大会云集了相关领域内国际上最具影响力的一批学术大师,例如J积分的提出者J. R. Rice教授、Paris公式的提出者 P. C. Paris教授、HRR场的提出者之一John W. Hutchinson教授,美国国家自然科学基金委员会前任主席 Subra Suresh教授、西安交大校友、美国工程院院士锁志刚教授、澳大利亚两院院士Y. W. Mai 教授等。我国国家自然科学基金委主任杨卫院士、香港科技大学孙庆平

教授、西安交大孙军教授等与会,并与西安交大校友、美国工程院院士高华健教 授等10位国外著名学者分别作50分钟大会特邀报告。大会共设54个主题论坛和15 个专题研讨会。

2) 将处理好的 word 文档转换为 TXT 文档后,将 TXT 文档另存为

UTF-8格式(具体方法如下)。

🗈 材料会议 - 记事本	_ D 🗙
文件(2) 編集(2) 格式(0) 查看(2) 帮助(3)	
第13届国际断裂大会(10F-13)于6月46百至21日在北京国家会议中心(IONC)隆重举行。来自四十多个国家的千余名断裂领域的专家、学者参加会议。西安交大孙 大会特邀报告,西安交大校友、美国工程院院士彻志制教授、高华健教授出席会议。国际断裂大会每四年举行一次,这是这一具有重要影响的国际盛会首次在F 的方方面面,如断裂力学、断裂与疲劳机制、断裂天垫技术、工程反用等,是材料断裂学科领域最高级别的国际学术会议。此次大会由中国力学学会、春巷都有 材料研究学会、中国旗性与防护学会、中国航空学会、中国全国学会、中国结构完整性难塑散夺承办。 坦次大会会集了相关领域内国际上最具影响力的一批学术大师、例如叫外的提出者3.8.8.8ice教授、Paris公式的提出者P.C.Paris教授、HRR场的提出者之一 国国家自然科学基金委员会前任主席Subra Surest教授、西安交大校友、美国工程院院士彻志则教授、澳大利亚两院院士V.M.Hai教授等,我国国家自然科学基 学孙庆平教授、西安交大外军教授等与会,并与西安交大校友、美国工程院院士家任命和规模学18位国外来至全要公司供知道你的生产专用的资格。	電對使应邀在大会上作了50分钟 中国召开。其內容特及断裂爭判 覺里、中国制成准律学会、中国 -John W. Hutchinson對授、美 全委主任杨卫院士、香港科技大 遞论坛和15个专题研讨会。
保存在 印: 🞯 貞面 🔍 😗 😰 🖽 -	
武力の文档 武力の文 武力の 武力の 武力の 武力の 武力の 武力の 武力 武力の 武力	
文件名 @): 材料会议 🔍 保存 ©)	
保存类型(1): 文本文档(4. txt) ▼ 取消	
#第号(注): ANST (* MSCT (* Ubiteda Ubiteda Utreda big entian 切子の	
	~
La 8, Cal 1	
🛂 开始 🚯 🔮 👔 🛇 " 🔟 gronoteur使用手 🗢 MAGABET (I:) 🖤 lizi (兼容模式) 🚺 材料合议 - 记事本	🖮 🛙 👔 🛛 🕵 🔊 😕 🐧 😨 🔁 16:21

2. 软件使用

2.1 软件基本操作

该软件只能在电脑的 C 盘中运行,所以切记将软件考入 C 盘,并 且不能重命名文件夹或者再添加新的文件夹

1) 双击g图标,打开软件。

点击工具栏第一个按钮,将会弹出一个对话框,根据自己需要, 进行重命名,选择 OK,此时该软件将为用户自动建立起一个语料库, 接下来用户所有产出的数据和图表将存在该库中(可重复建多个不同 的语料库),具体见下图。

Database Edit Help	
🤪 🥮 jiuju jjj	Database content
zn	database information current filter selections pages 0 sent. 0 vords 0.0 Mb 1970-1-1 00:00:00 consents:

2) 点击工具栏中第四个按钮,导入处理好的语料。

Database Edit Help	
😝 🔐 📾 📾 S 🐰 🎬 🐄 🚿 😹 📾 📾 😐	
Database content stadard stadard	lext
Choose the tab-separated file to import	
査技花園(1): 🕝 桌面 🔹 🖛 💼 🔶 🛅 -	
現地的文档 0.47 現地的文档 DI LTUTTI Lock Stock and Two Sweking Barrds 1998 Bladay 720p x264 DTC-+ ロ LTUTTI Lock Stock and Two Sweking Barrds 1998 Bladay 720p x264 DTC-+ (1) 0.00000000000000000000000000000000000	

3) 软件将显示以下状态

Growoteur	
Database Edit Halp	
e 🖓 e 🔚 e 🦸 🗱 🛤 🦻 🥷 😑 📾 e	
Database content	
diministrator corpus (0) 3 06 24 10 44 55	0
rovid vul time linkOrigin_rovid title_nbCharacters title_nbSentences title text_nbCharacters text_nbSentences text_n	es ser-
👔 database information 🖳 cell content 🌱 current filter 🔷 selections	
column = condition remove	
$1 rowid \qquad \Rightarrow = \Rightarrow \qquad O$	
19 开始 5 @ * O * M gronotew #用手 > gronotew III Chronotew III Gronotew III Gronotew III	(3) € 16:51

因为我们之前导入的语料为分段语料,所以在上图右边将会出现 2个栏目。

4)点击工具栏最右边九宫格图表(该图标显示的是图表中包括的各项信息,包括编号、标题、字数、句子等信息,用户可根据自己的需要自行选择),选择 text 选项,具体操作为:点来九宫格图表 ⇒点击"text"选项 ⇒点击"invert selection",此时将出现下图:



这样,我们的前期工作就做好了,接下来,我们将进行语义划分 (segmentation)和词性标注(tag)的操作。

2.2 语义划分及标注

2.2.1 语义划分:

在九宫格中选定 Text,点击工具栏中第 8 个"be in"按钮,在弹出的对话框中选择"word segmenter"下的"urheen segmenter"选项,此选项将进行 segmentation,见下图(urheen segmenter 只进行 segmentation, urheen segpos 能同时 segmentation 和 tagging,大家可以根据自己需要自行选择)。

_ Gromoteur	- 7 🗙
Database Edit Help	
🗑 🥔 📾 💼 🕲 🦿 🧱 🧏 🖉 📾 📾 😑	
Database content	
administratorcorpus/2013 06.24,16.44,55 Istandard	:01
Cro Tools	
source column: title, text	
Lematizer Word Segmenter Replacer	
only for Chinese: • mase • wheen segmenter • wheen segment	
only for English: wheen tokenizer urheen tager	
added to column name: _seg	
O database inform	
standard is the input view	
\sim	
🛃 开始 🔰 🚳 🖞 🛇 " 💟 granoteur/拥并 🝃 granoteur/III 🔽 Chygranoteur/III Granoteur 🐂 Gro Tools 🛛 CH 📍 🔍 📢 🖓 🔧	3 🗇 🕕 17:14

点击 go 按钮,将会出现如下结果(注意要将中间的进度条拉至最 右,也可以通过右边的九宫格选择 text-seg 项)。



2.2.2 词性标注:

如果想进行 tag 过程,则点击工具栏中第 8 个"be in"按钮,在弹出的对话框中选择"word segmenter"下的"urheen segpos"选项,点击 go 就可以进行 tag,结果如下图:



该软件同样支持数据分析。接着上面的操作,在九宫格中选定 text-seg 项,点击上方第十个按钮³⁶, statistical analysis,进行数据分 析。结果如下图所示:



图表中, Total frequency 指的是该词在整个语料中所出现的次数, frequency 指的是该词在选定的文本中出现的次数。如图中,"到"在整个语料中出现 8 次,在第一部分(即第一个小方框)出现 7 次。最后一组数据 specificity 指的是前两项分布出现的概率,即 SP=1/10^x。 大多数情况下,这个数据为 0,即出现的概率为 100%,当出现 1 以上或是负数时,便是值得深究的点。

2.3 结果导出:

我们以词性标注后的结果导出为例子,给大家展示如何将标注好的结果进行导出。

首先,点击工具栏中第11个选项 "export current view to file"。便

会出现对话框,如下图:

Gromoteur		a 🛛 🔀
Database Edit Help		
	🛛 🛤 🧏 🖉 📾 📾 😐	
	ase content	
jiuju jjj vcc	Exporter	÷01
	• text export O source export	✔ Ē/CC 21日/NT 在/P 北京/NR 国家/NR 会
	export to multiple files export to combined file	
	Reporting current view of vcc / standard	
	file name export/results.html	
	Choose file	
	View file after export	
	✓ export text inside of an HTML file	
	surround text matches by markup	
	word 🗹 match whole words	
		会议/NN 中心/NN (/PU CNCC/NR)/PU 隆
	😵 expert	//W 罢议/MA 。/PU 西安/MK 交人/MK 孙 、/PU 美国/NR 工程院/MM 院士/NN 锁志
		,/PU 这/PN 是/WC 这/DT 一/CD 具有/WV EG 方方面面/NN ,/PU 如/AD 断裂/WV 力
学级	W Cancel	N 断裂/VV 学科/NN 领域/NN 最/AD 高/JJ /PH 中国/NR 机械/NN 工程/NN 学会/NN
		P国/NR 全属/NN 学会/NN、/PU 中国/NR
	%/ ЯЯ プC2至(土/ JJ 時(金/ ЯЯ 時(日/ AD / %//)/ ** 。 / FU	
🛃 开始 😽 🚳 😰 🛇 " 🕅 grom	steur使用手 🍃 gronoteurIII 🔃 C:\gronoteurIII\ 🥃 Gronoteur	CH 🗘 🗄 👯 🔊 🧶 🖗 🖓 🕂 19:26

然后,将 file name 格中的名字全部删除,自己命名,并将下方 "export text inside of an HTML file"前的小箭头去掉,然后点击 expert,下方会出现一系列选项卡,如下图:

Gromoteur		
Database Edit Help		
9999	📱 📲 🧏 🚳 🔚 🗐 🥹	
	base content	
jiuju jjj vec	Exporter	÷0
	text export ort	✔ Ē/CC 21日/NT 在/P 北京/NR 国家/NN 会
	export to multiple files export to combined file	
	R exporting current view of wcc / standard	
	file name FCC	
	Choose file	
	V 📵 view file after export	
	export text inside of an HTML file	
	surround text matches by markup	
	word	
	C concordancer	
重		会议/NN 中心/NN (/PU CNCC/NR)/PU 隆 D/VV 会议/NN 。/PU 西安/NR 交大/NR 孙
	🐼 expert	、/PU 美国/NR 工程院/NN 院士/NN 锁志 ,/PU 这/PN 是/VC 这/DT 一/CD 具有/VV
	Jon X Cancel	EG 方方面面/NN , /PU 如/AD 断裂/VV 力 N 断裂/VV 学科/NN 领城/NN 最/AD 高/JJ
\sim		/PU 中国/NR 机械/NN 工程/NN 学会/NN P国/NR 金属/NN 学会/NN 、/PU 中国/NR
结?	闷/MN 完整性/JJ 联盟/MN 联合/AD 承办/WV 。/PU	
🛃 开始 🔰 🚳 😰 🛇 🦉 🕬 🖉 🕬	oteur使用手 🍃 gromoteurIII 🔗 C: \gromoteurIII\ 🤤 Gromoteur	CH 🗘 📋 🎘 🐨 🧶 😵 🔁 19:29

Gronoteur	- - X
Detabase Edit Help	
jiuju jiji wee Exporter	× Ø 🛄
text export source export	✔ Ē/CC 21日/NT 在/P 北京/NR 国家/NN 会
export to multiple files export to combined file	1
Reporting current view of wcc / standard	
S expert	
✓ surround table items by column name	
✓ escape tags (< and >) in html output	
Guess language (Chinese, left to right) O Force left to right O Force right to left	
comcordancer length in characters: 100	
match surround: (/span>	
surround text pages by: (sage) (/page)	
surround html pages by: (div name="page") (/div)	
thai template file: sources/htmlTemplate.html 🔞 View template	HOLEN CONTRACTION (POCNCC/NR)/PO ME D/VV 会议/NN 。/PU 西安/NR 交大/NR 孙
	、/PU 英国/NK 工程院/NN 院志/NN 税志 ,/PU 这/PN 是/VC 这/DT 一/CD 具有/VV
Cancel	EG 万方面面/NN ,/PU 如/AD 断裂/VV)」 N 断裂/VV 学科/NN 领域/NN 最/AD 高/JJ
	PU 中国/NR 初機/NN 工程/NN 学会/NN
5474/NH 元型性/JJ 联盟/NH 联告/AD 東示/W 。/PU	
5	
	'

我们可以根据自己需要选择,一般我们会将 match surround, surround text pages by,和 surround HTML pages by 这三个选项后的 6 个空格清空,然后点击 ok,处理好的文件会自动弹出打开,结果如此下图:



处理好的文件就在 gromoteur 文件夹中,根据自己当时的命名 找到对应文件,或者直接将弹出的文件"另存为",方便寻找。 后续的画树需要将此时导出的文件上传到 Arborator(这是专门 用于画树的在线资源,其使用方法后续会给大家介绍)上进行画树。 如下图:

