第十六届全国机器翻译大会(CCMT2020)

同源语料增强的印尼语汉语神经机器翻译

Cognate-Corpus-Enhanced Indonesian-Chinese Neural Machine Translation

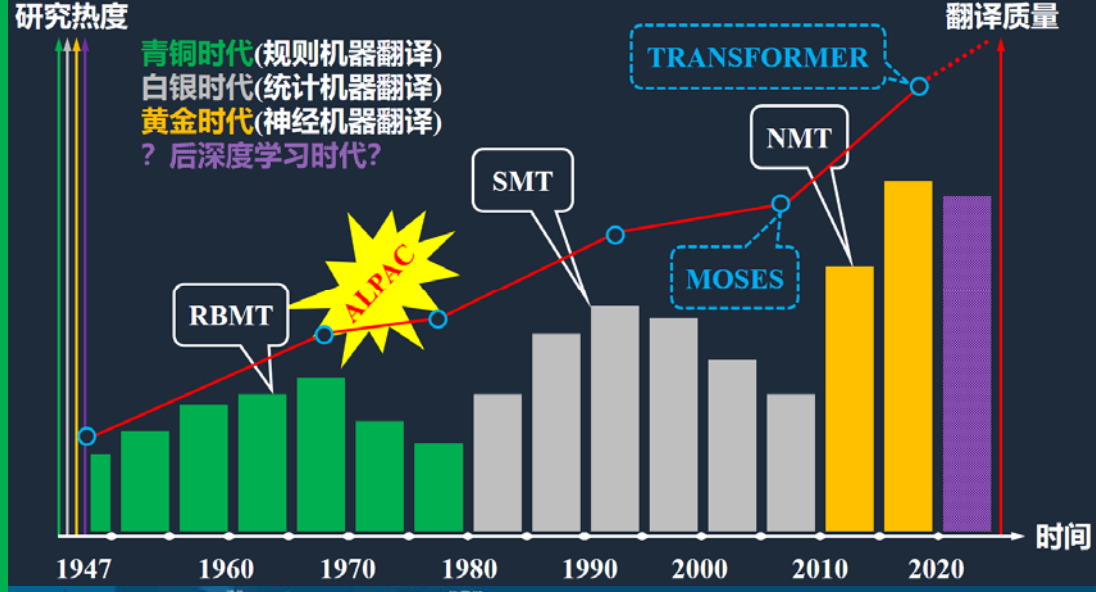
王　琳(lwang@xdsisu.edu.cn)

刘伍颖(wyliu@gdufs.edu.cn)

研究动机

- 机器翻译从青铜时代(规则机器翻译)、白银时代(统计机器翻译)、黄金时代(神经机器翻译)一路走来，诞生了系列优秀成果

- 神经机器翻译=超强的算力+超深的算法+超多的算料，当算力和算法趋同时，算料成为重要的能动研究
  - 对于动辄几千万甚至几亿句对算料的语言对，似乎已经解决机器翻译难题
  - 全世界尚存7,000多种语言，大部分非通用语言语料稀缺，低资源机器翻译依然是问题
  - 语言在进化过程中，形成远近亲疏的谱系树，同源语言之间在语音、形态、语义方面存在相似性，是否能用同源语料缓解低资源机器翻译问题



研究对象

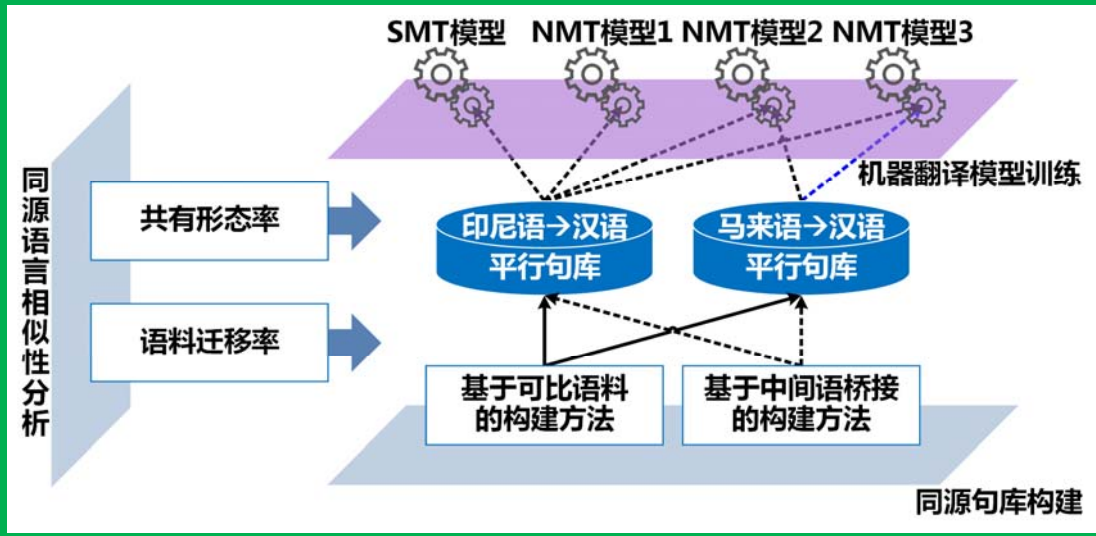
- 把印尼语和马来语作为研究对象语言
  - 印尼语和马来语同属南岛语系马来波利尼西亚语族
  - 广义而言印尼语也是一种马来语，类似于大陆汉语与新加坡华语、英国英语与美国英语之间的关系
  - 印尼语和马来语具有极强的同源关系，使得在语音、形态、语义等方面十分相似，拉丁形态字母表完全相同，就是英语字母表

- 选取印尼语汉语作为机器翻译目标语言对
  - 尽管网络印尼语资源远多于马来语资源，但较之英汉句对数量，印尼语汉语句对仍然是低资源的

据统计，目前母语为印尼语的人数有4,500万，全球印尼语使用者有1.6亿；而母语为马来语的人数超过8,000万，主要分布在文莱、印度尼西亚、马来西亚、新加坡等地，且全球马来语使用者超过3亿

总体框架

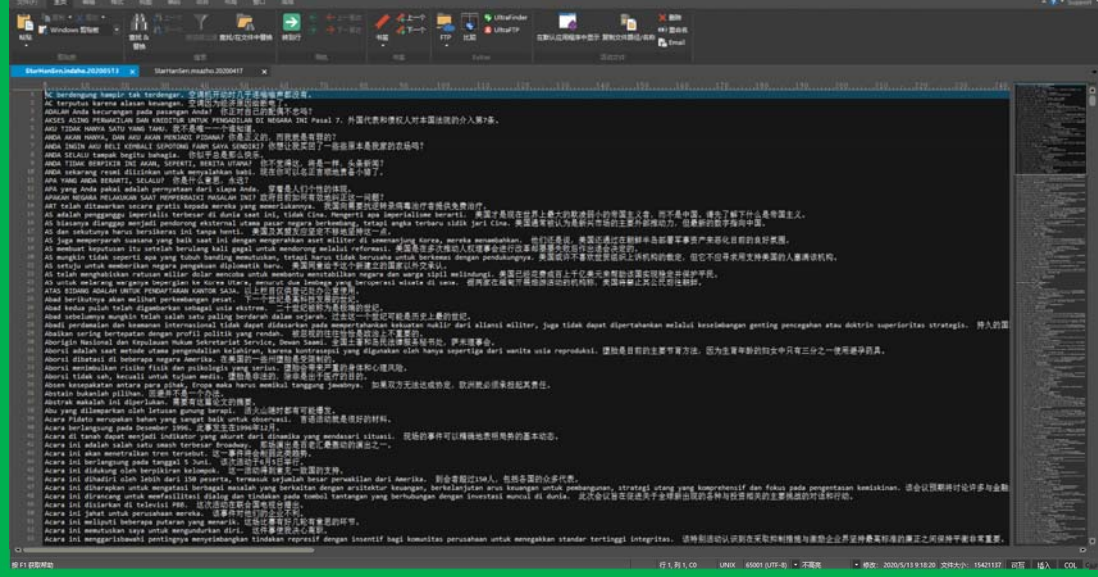
- 为了训练出高效的印尼语汉语机器翻译模型，我们设计了右图所示的语言资源扩建总体框架
  - 同源句库构建：我们采用基于可比语料的平行句库构建方法和基于中间语桥接的平行句库构建方法
  - 同源语言相似性分析：我们定量分析了马来语到印尼语的共有形态率和语料迁移率
  - 根据同源语言相似性分析结果，混合同源句库构建部分生成的印尼语汉语平行句库和马来语汉语平行句库，能够支持各种机器翻译模型的训练



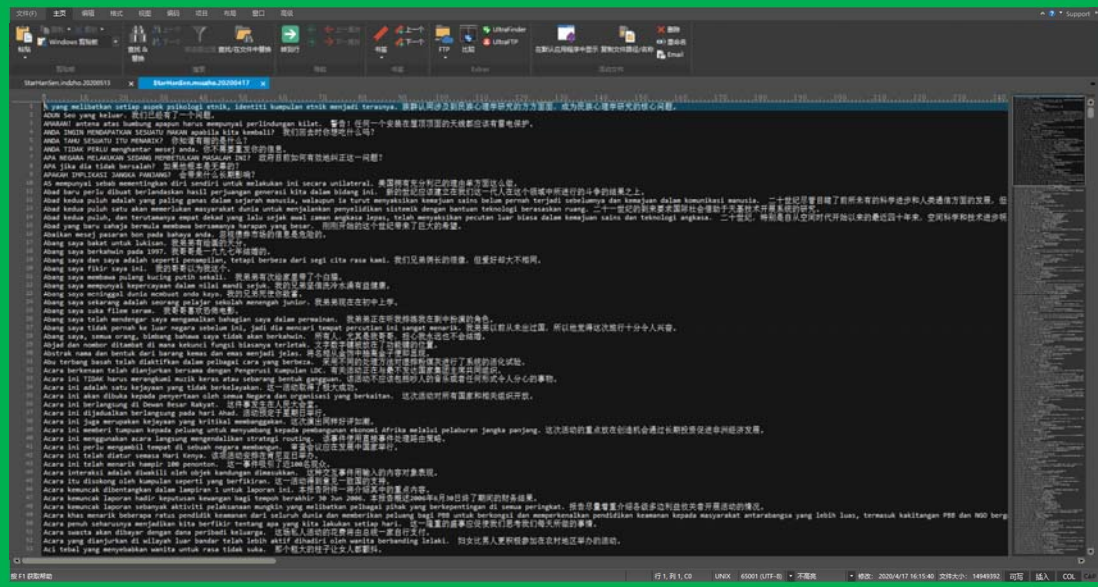
同源句库构建

- 基于可比语料的平行句库构建方法
  - 大脑的认知机能决定自然语言语义的等价性，当存在印尼语(马来语)到汉语的可比语料，可借助双语词典并根据标点符号、阿拉伯数字等语义不动点抽取语义等价的双语句子

- 基于中间语桥接的平行句库构建方法
  - 自然语言不仅存在语义等价性，且这种等价关系在不同形态表示的自然语言之间还具备传递性，当存在印尼语(马来语)到中间语以及中间语到汉语的大规模平行句库，利用形态传递性，计算中间语句子的相似度桥接构建印尼语(马来语)到汉语的平行句库



印尼语汉语平行句库包含2,489,442句对



马来语汉语平行句库包含761,373句对

同源语言相似性分析

- 以Wikipedia为统计分析源，20180501版包含印尼语文档439.3MB、马来语文档173.8MB

- 去重后词汇级*N*-gram串数如右表所示
  - *1*-gram数据表明印尼语文档是由657,409个不同的印尼语词汇组成
  - 马来语文档只由395,365个不同的马来语词汇组成
  - 两种语言的共有词汇数量多达211,453

表2. 词汇级*N*-gram串数

Table 2. Number of Word-Level *N*-gram Token

N-gram	Indonesian (MOR)	Malay (MOR)	Overlap
1-gram	657,409 (32.16%)	395,365 (53.48%)	211,453
2-gram	8,035,771 (15.36%)	3,994,868 (30.89%)	1,233,931
3-gram	19,614,857 (6.51%)	8,756,914 (14.58%)	1,276,456
4-gram	23,038,089 (2.84%)	9,708,373 (6.73%)	653,356
5-gram	20,797,228 (1.66%)	8,658,728 (4.00%)	345,955

共有形态率

- 两种语言之间共有形态串数对每种语言的效力是不一样的
- 把具有方向性的A语言到B语言的N-gram共有形态率(*Morphological Overlap Ratio*，简记为$MOR_{N-gram}^{A\to B}$)定义为A语言和B语言共有*N*-gram串数除以B语言的*N*-gram串总数得到的百分比
  - $MOR_{N-gram}^{A\to B}$ 数值反映了A语言语料的*N*-gram形态学习结果用于B语言的有效程度
  - 计算可知马来语(M)到印尼语(I)的*1*-gram共有形态率$MOR_{1-gram}^{M\to I}$=32.16%

表2. 词汇级*N*-gram串数

Table 2. Number of Word-Level *N*-gram Token

N-gram	Indonesian (MOR)	Malay (MOR)	Overlap
1-gram	657,409 (32.16%)	395,365 (53.48%)	211,453
2-gram	8,035,771 (15.36%)	3,994,868 (30.89%)	1,233,931
3-gram	19,614,857 (6.51%)	8,756,914 (14.58%)	1,276,456
4-gram	23,038,089 (2.84%)	9,708,373 (6.73%)	653,356
5-gram	20,797,228 (1.66%)	8,658,728 (4.00%)	345,955

语料迁移率

- 我们把A语言到B语言的语料迁移率(*Corpus Transfer Ratio*，简记为$CTR_N^{A\to B}$)定义为$\sum_{i=1}^{N}\alpha_i MOR_{i-gram}^{A\to B}$
  - 系数α_i表示A语言语料集中未去重*i*-gram串数除以全部未去重*N*-gram串数总和的百分比
- 由于语言的长程相关性相对较弱，通常只需要简化计算*N*=5时的$CTR_N^{A\to B}$用于表示A语言语料到B语言语料的增强程度

表2. 词汇级*N*-gram串数

Table 2. Number of Word-Level *N*-gram Token

N-gram	Indonesian (MOR)	Malay (MOR)	Overlap
1-gram	657,409 (32.16%)	395,365 (53.48%)	211,453
2-gram	8,035,771 (15.36%)	3,994,868 (30.89%)	1,233,931
3-gram	19,614,857 (6.51%)	8,756,914 (14.58%)	1,276,456
4-gram	23,038,089 (2.84%)	9,708,373 (6.73%)	653,356
5-gram	20,797,228 (1.66%)	8,658,728 (4.00%)	345,955

实验

- 实验语料分成训练、验证、测试三个集合，其中验证集和测试集各包含20,000印尼语汉语句对，是从2,489,442句对印尼语汉语平行句库中采用不放回采样技术随机抽取而成
  - 把2,489,442句对的印尼语汉语平行句库简记为IdChSens
  - 把除去验证集和测试集后剩下的2,449,442句对命名为印尼语汉语部分句库IdChSens-P
  - 把761,373句对的马来语汉语平行句库简记为MsChSens
- 统计MsChSens和IdChSens中马来语和印尼语的1-gram共有形态集OverlapWordSet
- 检测MsChSens中每个马来语句子，若该句中的每个词都属于OverlapWordSet，则将包含该句子的马来语汉语句对加入马来语汉语部分句库MsChSens-P
  - 可见MsChSens-P是MsChSens的一个子集，实际检测结果MsChSens-P包含696,840句对

表3. 机器翻译实验结果

Table 3. Experimental Results of Machine Translation

Model	Train Set	Validation Set	Test Set	BLEU4
SMT	IdChSens-P 2,449,442	20,000	20,000	19.58
NMT1	IdChSens-P 2,449,442	20,000	20,000	17.13
NMT2	IdChSens-P+MsChSens 3,210,815	20,000	20,000	20.30
NMT3	IdChSens-P+MsChSens-P 3,146,282	20,000	20,000	20.30

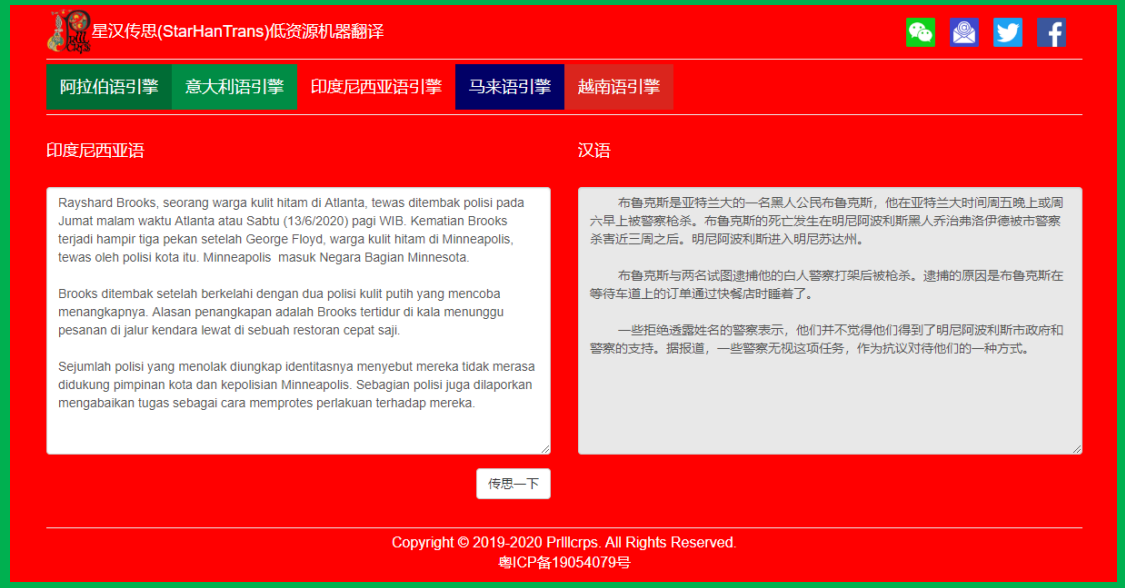
印尼语源文	汉语参考译文	45	我们的神经机器翻译译文	56	谷歌机器翻译译文	54	必应机器翻译译文	47
dengan kata lain, untuk efesiensi dan mengelola diri sendiri.	也就是自足和自理。	1	也就是自足和管理自我。	1	换句话说，就效率和自我管理而言。	3	换言之，为了有效地使用和管理自己。	4
dude, anda tampak konyol.	哥们，你穿得太荒唐了。	2	伙计，你看起来很傻。	4	老兄，你看起来很傻。	4	伙计你看起来很可笑	3
haruskah kita masuk?	那我进去了？	1	我们要进去吗？	4	我们应该进入吗？	2	我们应该进去吗？	3
ini adalah pertama kali saya mendengar kabar darinya dalam setahun.	这是一年来我第一次有他的消息。	4	这是我一年来第一次听到他的消息。	4	这是我一年中第一次听到他的消息。	3	这是我一年来第一次收到他的来信。	2
istri polisi mungkin akan berguna.	一个警察的妻子也许有点用。	2	警察老婆可能有用。	3	警察的妻子可能会有用。	3	警察的妻子也许有用。	4
itu seperti rakun mabuk mendapat seluruh berlian hope.	就像一只醉酒的浣熊，手上拿着蓝宝石。	2	这就像一只醉醺醺的浣熊，得到了希望的钻石。	3	就像醉酒的浣熊得到了整个钻石的希望。	1	就像喝醉了的浣熊到处都是希望钻石	1
jangan ungkit lagi.	不要去想它。	3	别再提了。	4	不要备份。	2	不要再 ungkit 了。	2
kau punya kebenarannya? berkendara dari peru ke polinesia dengan rakit dari pohon balsa!	搭轻木木筏从秘鲁漂到玻里尼西亚！	3	从秘鲁开车到玻里尼西亚，从轻木木岛上开木筏！	2	你明白了吗？ 从一个巴尔萨树驱车从秘鲁到波利尼西亚！	2	你知道真相了吗？从秘鲁到波利尼西亚，用一筏轻的木筏开车！	2
lakukan apa yang anda anggap terbaik, sayang.	做你认为最好的事，亲爱的。	3	尽你所能，亲爱的。	1	做你认为最好的，亲爱的。	4	做你认为最好的，爱。	2
pintu belakang ke neraka.	进入地狱的后门。	4	后门进入地狱。	1	后门到地狱。	1	回到地狱的大门。	3
setelah bertengkar dengan dua wanita itu, aku menelepon penyelamatku.	但是那两妞都跑走了，弄的我浑身骚动。	1	和两个女人吵架之后，我就呼唤了我的救世主。	3	在和这两个女人打架之后，我打电话给我救援。	3	和两个女人吵架后，我给 penyelamatku 打了电话。	2
stasiun listrik siap, pak!	通电站准备好了！	2	电力站好了，长官！	3	电站准备好了，先生！	4	电车站准备好了，先生！	1
sudah lama sejak yang terakhir menungganginya selama 8 detik.	很少有人可以骑超过八秒。	4	已经很久没有人骑过它 8 秒了。	4	自从最后一次骑了 8 秒钟以来已经很久了。	3	从上一次坐了 8 秒后已经很久了。	3
tidak, tidak, sayang, tidak ada jalan.	不，不，不，没门！	1	不，亲爱的，没有路。	1	不，不，亲爱的，没办法。	4	不，不，亲爱的，没有办法。	4
sekarang, giliranmu membereskan semuanya.	现在，轮你来处理后事了。	1	现在轮到你收拾残局了。	4	现在，轮到你照顾好一切。	1	现在轮到你了，保重一切。	1
katakan, kapan?	说吧，什么时间？	3	告诉我，什么时候？	4	说什么时候？	1	说什么时候？	1
berpotensi, punya potensi untuk mengubah dunia.	能力，有能力就能改变天下。	2	潜在的潜力，能改变世界。	3	潜在的，有可能改变世界。	3	有可能改变世界。	3
apakah aku pernah membuat kita tersesat?	我有没有害我们迷路过？	4	我有没有让我们迷路？	3	我曾让我们误入歧途吗？	4	如果我让我们迷路了呢？	1
pak, kita sudah tiba dikoordinat yang telah kita hitung. tidak ada apa2 di sini, apa perintahmu?	长官，我们已经到达了你预定的坐标，这里什么都没有？	1	长官，我们已经到达预定座标了，这里什么都没有？	2	先生，我们已经到达了我们计算的坐标。 什么都没有，你的命令是什么？	3	我们已经计算过 dikoordinat 了这里没有 apa2，是什么吩咐你的？	1
aku akan menemukannya, dan kau akan membantuku menemukannya.	我要找到他，你要帮我找到他。	1	我会找到她，你会帮我找到她。	2	我会找到的，你会帮助我找到它。	3	我会找到的，你会帮我找到的。	4

结论

- 马来语语料能够有效改进印尼语汉语神经机器翻译主要是源于同源语言之间的形态相似性和语义等价性
- 神经机器翻译能够从低质量语料中学出超过参考译文的翻译模型，这归功于深度学习细化了学习特征，并充分利用了巨量训练数据
- 未来平行句库构建中，希望在双语资源之外发挥单语资源的规模优势，通过高效机器翻译算法生成句对
- 此外针对神经机器翻译超参考质量的现象，需要探索平行句对的质量评价算法，优化训练语料质量和译文评价方法

星汉传思(StarHanTrans)低资源机器翻译

- http://lorelp.oicp.io/indzho



- http://lorelp.oicp.io/msazho

语言数据建设成果

- 同文术语(TWTerm)矩阵
  - 103种语言(104种文字)×100万术语对

	A	B	C	D	E	F	G	H	I	J
1	accumulated error	累计误差	累計誤差	تراكمت خطأ	accumulé erreur	kumulierte Fehler	накопленная ошибк	acumulada de error	累積誤差	오류를 축적
2	aircraft carrier	航空母舰	航空母艦	حاملة طائرات	porte-avions	Flugzeugträger	авианосец	portaaviones	空母	항공 모함
3	alternate load	交替载重	交替載重	البديل تحميل	Autres charges	Alternate Last	Знакопеременная н	Suplente de carga	交番荷重	대체로드
4	automatic control	自动控制	自動控制	التحكم الآلي	contrôle automatique	automatische Steuer	автоматический ко	control automático	自動制御	자동 제어
5	average dia	平均直径	平均直徑	متوسط دیا	dia moyenne	durchschnittlichen D	Средний диаметр	de diámetro promedi	平均直径	평균 디아
6	bikini	比基尼	比基尼	بيكيني	bikini	Bikini	бикини	bikini	ビキニ	비키니
7	borderline patient	临界病人	臨界病人	الشريط الحدودي المريض	borderline patient	Borderline-Patienten	пограничный паци	limite de la paciente	境界型患者	환자의 경계선
8	braking force	制动力	制動力	قوة الكبح	la force de freinage	Bremskraft	торможение	la fuerza de frenado	制動力	강제로 제동
9	break time	休息时间	休息時間	كسر الوقت	temps de pause	Pausenzeit	перерыв	tiempo de descanso	休憩時間	휴식 시간
10	Buckingham Palace	白金汉宫	白金漢宮	قصر باكنغهام	Buckingham Palace	Buckingham Palace	Букингемский двор	El Palacio de Buckir	バッキンガム宮殿	버킹엄 궁전
11	Buddha's relics	舍利子	舍利子	بوذا الآثار	Reliques de Bouddh	Buddhas Reliquien	Будды реликвии	reliquias de Buda	仏の遺物	부처의 유물
12	bulk dry milk	散装奶粉	散裝奶粉	معظم الحليب الجاف	lait en vrac sec	Bulk-Trockenmilch	часть сухого моло	leche en polvo a gra	一括粉乳	일괄 건조 우유
13	capital account	资本帐户	資本帳戶	حساب رأس المال	Compte de capital	Capital-Konto	Счет операций с ка	Cuenta de Capital	資本収支	자본 계정
14	capitalism	资本主义	資本主義	رأسمالية	Le capitalisme	Kapitalismus	Капитализм	El capitalismo	資本主義	자본주의
15	Chinese Medical As	中华医学会	中華醫學會	الجمعية الطبية الصينية	Association médicale	Chinese Medical As	Китайской медицин	Asociación Médica (	中国の医療協会	중국어 의사회
16	circuitry	电路	電路	الدارات	circuits	Schaltung	Схема	circuitos	回路	회로
17	civilized force	文明之师	文明之師	قوة حضارية	force civilisée	zivilisierte Kraft	цивилизованного си	fuerza civilizada	文明の力	문명 힘
18	conductive tissue	导电组织	導電組織	الأنسجة موصل	conductrice de tissus	Bindegewebe	проводящих тканей	tejido conductivo	導電性組織	전도성 조직
19	confidence region	置信区间	置信區間	منطقة الثقة	région de confiance	das Vertrauen der R	доверия область	región de confianza	信頼性領域	신뢰 영역
20	contradict	矛盾	矛盾	تناقض	contredisent	widersprechen	противоречить	contradecir	矛盾	모순
21	crown prince	王储	王儲	ولي العهد	Le prince héritier	Kronprinz	Наследный принц	Corona Principe	皇太子	황태자
22	cutting machine	切割机	切割機	آلة قطع	Machine de découpe	Schneidemaschine	станки	Máquina de corte	切断機	절단기
23	datum line	基准线	基準線	خط المسند	ligne de référence	Bezugslinie	базовая линия	linea de referencia	基準線	기준선

语言数据建设成果

- 星汉句库(StarHanSen)
  - 103种语言×(1200万句对~4亿句对)

南非语	afr	菲律宾语	fil	爪哇语	jav	毛利语	mri	索马里语	som	意第绪语	yid
阿姆哈拉语	amh	芬兰语	fin	日本语	jpn	马来语	msa	塞索托语	sot	约鲁巴语	yor
阿拉伯语	ara	法语	fra	卡纳达语	kan	缅甸语	mya	西班牙语	spa	祖鲁语	zul
阿塞拜疆语	aze	弗里西语	fry	格鲁吉亚语	kat	尼泊尔语	nep	阿尔巴尼亚语	sqi		
白俄罗斯语	bel	苏格兰盖尔语	gla	哈萨克语	kaz	荷兰语	nld	塞尔维亚语	srp		
孟加拉语	ben	爱尔兰语	gle	柬埔寨语	khm	挪威语	nor	巽他语	sun		
波斯尼亚语	bos	加利西亚语	glg	卢旺达语	kin	齐切瓦语	nya	斯瓦希里语	swa		
保加利亚语	bul	古吉拉特语	guj	吉尔吉斯语	kir	奥里亚语	ori	瑞典语	swe		
加泰罗尼亚语	cat	海地克里奥尔语	hat	韩国语	kor	旁遮普语	pan	泰米尔语	tam		
宿务语	ceb	豪萨语	hau	库尔德语	kur	波兰语	pol	鞑靼语	tat		
捷克语	ces	夏威夷语	haw	老挝语	lao	葡萄牙语	por	泰卢固语	tel		
科西嘉语	cos	希伯来语	heb	拉脱维亚语	lav	普什图语	pst	塔吉克语	tgk		
威尔士语	cym	印地语	hin	立陶宛语	lit	罗马尼亚语	ron	泰语	tha		
丹麦语	dan	克罗地亚语	hrv	卢森堡语	ltz	俄语	rus	土库曼语	tuk		
德语	deu	匈牙利语	hun	马拉雅拉姆语	mal	僧伽罗语	sin	土耳其语	tur		
希腊语	ell	亚美尼亚语	hye	马拉地语	mar	斯洛伐克语	slk	乌克兰语	ukr		
英语	eng	伊博语	ibo	马其顿语	mkd	斯洛文尼亚语	slv	乌尔都语	urd		
爱沙尼亚语	est	印度尼西亚语	ind	马尔加什语	mlg	萨摩亚语	smo	乌兹别克语	uzb		
巴斯克语	eus	冰岛语	isl	马耳他语	mlt	修纳语	sna	越南语	vie		
波斯语	fas	意大利语	ita	蒙古语	mon	信德语	snd	科萨语	xho		

Thank You!
Q&A

Prlllcrps 刘伍颖

广东 广州 18819407832

扫一扫上面的二维码图案，加我微信