汉字语用信息的复杂性与文本分级的可行性对策

山东大学文学院、澳门科技大学（兼职教授）
盛玉麒
香港城市大学
2017-10-23

一、问题的提出

- 1.1汉语文本分级的对象和目标
- 1.1.1对象：汉字文本
- 文体：多样化：
- 寓言、故事、神话、小说、科幻、艺术、悬疑、推理
- 1.1.2目标：知识+能力
- 知识：多学科：自然、地理、动物、社会、事理、观念、情感、态度
- 能力：识文字—知词语—懂语感—通篇章—达文意

1.1.3、“文本+知识+能力”相关性

- 1.1.3.1 知识相关性信息
- 感觉、知觉、直接经验、客观性；
- 情景语境依存线性推理、
- 因果知识
- 逻辑事理
- 作用与关系
- 1.1.3.2 能力相关性信息
- 分析认知、跨域重组；主观化；
- 情景语境“多维度”假设推理；虚拟建构、
- 联想类推
- 隐喻认知
- 具体表现在：原型与语境域的“跨域认知”，
- 能力级差表现在知识域概念语义空间距离的认知上。

1.2 汉语文本分级的挑战

- 关键：透过汉字的假象；
- 表现：汉字静态属性与动态功能的不一致
- 1.2.1一字多音
- 字数与音节数相比2-20倍；
- 若考虑变读因素多音字更多：
- 例如：干什么都差不了
- 文白异读：
- 差chā（文）不～什么、偏～、～别、视～误～、～池、～错、～等、～额、～异
- 差2 chà（语）～不多 ～不离 ～点儿
- 差3 cī 参～

1.2.2字义与词义的不一致性

- 语素化：字义转化
- 上：桌子上、火车上、飞机上、问题上、方法上、态度上、表现上、会上、心上、认识上……
- 奇葩—奇葩说、控制—管控—操控—**控
- 粉丝—钢丝—玉米—姜豆……
- “**粉”、“**饭”
- 符号化：字义消失
- 连绵词、音译词、
- “变形词”：酱紫、酿紫、童鞋

1.2.3 字级与文本分级不一致

- 1.2.3.1蒙学读物的繁难字
- 人之初，性本善。性相近，习相远。苟不教，性乃迁。教之道，贵以专。（三字经）
- 赵钱孙李 周吴郑王 冯陈诸卫 蒋沈韩杨 朱秦尤许 何吕施张（百家姓）
- 天地玄黄　宇宙洪荒、日月盈昃　辰宿列张（千字文）
- 弟子规　圣人训　首孝悌　次谨信（弟子规）
- 天子重英豪，文章教尔曹；万般皆下品，唯有读书高。（神童诗-劝学）
- 春眠不觉晓 处处闻啼鸟 夜来风雨声 花落知多少（千家诗：春眠－孟浩然）
- 混沌初开，乾坤始奠。（《幼学琼林》-天文）
- 天对地，雨对风。大陆对长空。山花对海树，赤日对苍穹。（《笠翁对韵》）

1.2.3.2童话故事的冷僻字

- 龟兔赛跑、女娲补天、夸父追日、羿射九日
- 龜兔賽跑、女媧補天、夸父追日、羿射九日
- 嫦娥奔月、掩耳盗铃、揠苗助长、为虎作伥
- 嫦娥奔月、掩耳盜鈴、揠苗助長、為虎作倀
- 精卫填海、愚公移山、凿壁偷光、黔驴技穷
- 精衛填海、愚公移山、鑿壁偷光、黔驢技窮
- 杯水车薪、铁杵磨针、郑人买屦、南辕北辙
- 杯水車薪、鐵杵磨針、鄭人買屨、南轅北轍

1.2.3.3成语典故的易误字

- “舍不得孩子套不住狼”——“鞋子”
- “狗屁不通”——“狗皮不通”
- “王八蛋”——“忘八端”。
- “跳进黄河——洗不清”——歇后语
- 无毒不丈夫——无度不丈夫
- 民可使由之，不可使知之——断句
- 不刊之论、明日黄花、醍醐灌顶、细大不捐
- 三人成虎、不易之论、罪不容诛、屡试不爽

1.2.3.4人教版七年级语文成语例样

- 不毛之地、参差不齐、叱咤风云、当之无愧、风餐露宿、
- 锋芒毕露、妇孺皆知、孤立无援、浑身解数、家喻户晓、
- 坚持不懈、尽态极妍、进退维谷、精打细算、精疲力竭、
- 迥乎不同、鞠躬尽瘁，死而后已、慷慨淋漓、可歌可泣、
- 来势汹汹、聊以自慰、略胜一筹、马革裹尸、毛骨悚然、
- 念念有词、疲惫不堪、气冲斗牛、锲而不舍、人迹罕至、
- 人声鼎沸、姗姗来迟、石破天惊、叹为观止、忘乎所以、
- 畏缩不前、闻所未闻、鲜为人知、相得益彰、寻欢作乐、
- 眼花缭乱、洋洋得意、养精蓄锐、怏怏不乐、耀武扬威、
- 一如既往、义愤填膺、勇往直前、忧心忡忡、语无伦次、

1.2.3.5少儿读物的符号字

- 宙斯、欧罗巴、普罗米修士、斯巴达克
- 舒克贝塔、皮皮鲁、
- 《小猪唏哩呼噜》、
- 《宝葫芦的秘密》
- 《查理和巧克力工厂》（英）、
- 《男生贾里》、
- 《随风而来的玛丽阿姨》（英）、
- 《绿山墙的安妮》

1.2.3.6生活词汇的罕见字

- 亲属称谓：叔伯姑舅、爷奶姨姥
- 身体部位：膝踝胯髋、肩肘胳膊
- 衣食住行：咳嗽喷嚏、搔痒肿痛、
- 果蔬食品：葡萄菠萝、蛇果榴莲
- 跌摔磕碰、迷眼魇梦、呛嗳呕吐

二、字级参数的不完备性

- 字频统计结果显示，降频表前2500字在现代汉语通用文本中累计覆盖率高达99.9%；
- HSK等级大纲学界公认具有科学性、实用性；
- 被多种能力水平测试直接使用或主要参照；
- 新HSK用汉字2624个，覆盖《现汉》（6版）31000“熟字生词”；
- 理论上认识2624个纲内字，就能“认识”31000条生词；
- 但实际上，即使是全甲级字的成语、熟语，多数人会认却未必能理解其语用含义。

表 5：新 HSK 词根语素“纲外”双音“熟字生词”量表

首字 / 次字		首语素						
		1 级	2 级	3 级	4 级	5 级	6 级	合计
次语素	1 级	884	622	815	909	928	608	4766
	2 级	742	554	687	782	818	549	4132
	3 级	905	631	1110	1267	1189	889	5991
	4 级	964	689	1184	1434	1272	973	6516
	5 级	882	679	1053	1347	1394	945	6300
	6 级	608	401	657	790	878	772	4106
	合计	4985	3576	5506	6529	6479	4736	31811

2.1全甲级字成语理解难度

493	非亲非故	非	1	亲	1	非	1	故	1
494	非同小可	非	1	同	1	小	1	可	1
495	非意相干	非	1	意	1	相	1	干	1
496	飞短流长	飞	1	短	1	流	1	长	1
497	飞流短长	飞	1	流	1	短	1	长	1
498	分别门户	分	1	别	1	门	1	户	1
499	分而治之	分	1	而	1	治	1	之	1
500	分工合作	分	1	工	1	合	1	作	1
501	久必合，合久必分	分	1	久	1	必	1	合	1
502	分门别户	分	1	门	1	别	1	户	1
503	分情破爱	分	1	情	1	破	1	爱	1
504	分三别两	分	1	三	1	别	1	两	1
505	分文不取	分	1	文	1	不	1	取	1
506	分文不直	分	1	文	1	不	1	直	1

1000	论长说短	论	1	长	1	说	1	短	1
1001	论短道长	论	1	短	1	道	1	长	1
1002	论黄数白	论	1	黄	1	数	1	白	1
1003	论黄数黑	论	1	黄	1	数	1	黑	1
1004	论千论万	论	1	千	1	论	1	万	1
1005	论世知人	论	1	世	1	知	1	人	1
1006	马放南山	马	1	放	1	南	1	山	1
1007	马面牛头	马	1	面	1	牛	1	头	1
1008	马牛其风	马	1	牛	1	其	1	风	1
1009	马如流水	马	1	如	1	流	1	水	1
1010	马如游鱼	马	1	如	1	游	1	鱼	1
1011	马上得天下	马	1	上	1	得	1	天	1
1012	马上房子	马	1	上	1	房	1	子	1
1013	马上看花	马	1	上	1	看	1	花	1
1014	马上墙头	马	1	上	1	墙	1	头	1
1015	马中关五	马	1	中	1	关	1	五	1
1016	买东买西	买	1	东	1	买	1	西	1

全甲级俗语例样

1	有把门的可没有把嘴的	有	1	把	1	门	1	的	1	可	1	没	1	有	1	把	1	嘴	1	的	1										
2	每家都有一本难念的经	每	1	家	1	都	1	有	1	一	1	本	1	难	1	念	1	的	1	经	1										
3	墙头一棵草，风吹两边倒	墙	1	头	1	一	1	棵	1	草	1	，	1	风	1	吹	1	两	1	边	1	倒	1								
4	说着红脸的，便来了关公	说	1	着	1	红	1	脸	1	的	1	，	1	便	1	来	1	了	1	关	1	公	1								
5	手中没把米，叫鸡鸡不来	手	1	中	1	没	1	把	1	米	1	，	1	叫	1	鸡	1	鸡	1	不	1	来	1								
6	一个唱红脸，一个唱白脸	一	1	个	1	唱	1	红	1	脸	1	，	1	一	1	个	1	唱	1	白	1	脸	1								
7	一个是半斤，一个是八两	一	1	个	1	是	1	半	1	斤	1	，	1	一	1	个	1	是	1	八	1	两	1								
8	河边没青草，不要多嘴驴	河	1	边	1	没	1	青	1	草	1	，	1	不	1	要	1	多	1	嘴	1	驴	3								
9	没吃过猪肉，也见过猪跑	没	1	吃	1	过	1	猪	1	肉	1	，	1	也	1	见	1	过	1	猪	1	跑	1								
10	出门三步远，又是一层天	出	1	门	1	三	1	步	1	远	1	，	1	又	1	是	1	一	1	层	1	天	1								
11	海里的水，到哪儿哪儿咸	海	1	里	1	的	1	水	1	，	1	到	1	哪	1	儿	1	哪	1	儿	1	咸	3								
12	三九天也别想跟他要点雪	三	1	九	1	天	1	也	1	别	1	想	1	跟	1	他	1	要	1	点	1	雪	1								
13	三打不回头，四打连身转	三	1	打	1	不	1	回	1	头	1	，	1	四	1	打	1	连	1	身	1	转	2								
14	笑到最后，谁才笑得最	谁	1	笑	1	到	1	最	1	后	1	，	1	谁	1	才	1	笑	1	得	1	最	1	好	1						
15	要马儿好，又要马儿不吃	又	1	要	1	马	1	儿	1	好	1	，	1	又	1	要	1	马	1	儿	1	不	1	吃	1	草	1				
16	你的阳关道，我走我的独	你	1	走	1	你	1	的	1	阳	1	关	1	道	1	，	1	我	1	走	1	我	1	的	1	独	2	木			
17	白门走四方，黑夜点灯补	白	1	天	1	游	1	门	1	走	1	四	1	方	1	，	1	黑	1	夜	1	点	1	灯	1	补	2	衲			
18	坐立得正，哪怕和尚尼姑	坐	1	得	1	正	1	来	1	立	1	得	1	正	1	，	1	哪	1	怕	1	和	1	尚	3	尼	4	姑			
19	子吃了十八里地，还吃不	一	1	个	1	包	1	子	1	吃	1	了	1	十	1	八	1	里	1	地	1	，	1	还	1	吃	1	不			
20	胖子不是一口吃起来的	胖	2	子	1	不	1	是	1	一	1	口	1	吃	1	起	1	来	1	的	1										
21	把吃奶的劲都使出来了	把	1	吃	1	奶	1	的	1	劲	2	都	1	使	1	出	1	来	1	了	1										
22	朝山的不是全为了敬神	朝	1	山	1	的	1	不	1	是	1	全	1	为	1	了	1	敬	2	神	1										
23	一个鸡蛋算出四两骨头	一	1	个	1	鸡	1	蛋	1	算	1	出	1	四	1	两	1	骨	2	头	1										
24	刚学会理发碰上大胡子	刚	1	学	1	会	1	理	1	发	1	碰	1	上	1	大	1	胡	2	子	1										
25	此处不留人，自有留人处	此	2	处	1	不	1	留	1	人	1	，	1	自	1	有	1	留	1	人	1	处	1								

2.2 原因：汉字语用信息复杂性

- 2.2.1具体情景话题语境下语用含义灵活多样；
- 2.2.2作家作品话语方式的个性化；
- 2.2.3 网络新媒体推波助澜，
- 误用、误读、误解与“新、奇、特”交互影响、难辨真伪，
- 2.2.4 阅读者认知方式的新常态

2.3 网红流行语的万象百态

- 厚德载雾、自强不吸、与食俱进、
- 厚颜五尺、逢堵必疏、钱程远大、
- 独具贿眼、勤捞致富、白收起家、
- 一身政气、一捞永逸、前腐后继；
- 成语新解
- 知书达礼、度日如年、杯水车薪
- 友谊的小船怎能说翻就翻？洪荒之力、
- 先定个小目标、厉害了我的哥。

三、文本分级对策

3.1基本判断

针对汉字语用功能的分化，挖掘标志性“语用”信息作为文本分级的可操作参数。

词字：

天、地、人、刀、口、手、马、牛、羊

词素字：

模样、气氛、活力、经济、法律

符号字：

阿弥陀佛、磺胺嘧啶、阿司匹林

3.2 分级原理基本假说

- 3.2.1 中文阅读文本是汉语动态语用系统的书面映射；
- 3.2.2 基于单音节词根孤立语的古代汉语而创制的汉字符号系统，在记录现代汉语时表现出语用信息复杂性的特点；
- 3.2.3 汉字语用信息复杂性源自“字-词-语-段”多维度、多层级相关作用的结果；
- 3.2.4 通过规则+统计的策略，挖掘和提取“特征标记”知识点作为文本分级参数体系具有可行性。

3.2.5“掘土机参数”的启发

- GDP是一个经济发展的衡量指标；
- 构成要素和经济行为具有复杂性系统的特点；
- 2014年以来有数据表明，“挖掘机”销售、工作量、分布、使用情况等与GDP曲线有明显的“正相关”关系；
- 于是有了“掘土机参数”的说法和共识。

3.3 字词语用信息知识点挖掘

3.3.1构词信息：词素化与词义泛化；

能否正确理解构词语素与所构词汇意义的关系，涉及到汉语词法知识和构词能力的评估参数；

分拆—拆分、快递—速递、速食面—即食面

难点：有效避免望文生义所致偏误

3.3.2 词汇习得应用的“家族相似性”认知模式，涉及阅读理解过程的认知语用能力程度。

宽带：宽衣解带（？）

带宽：衣带渐宽（？）

难点：认知习惯中知识与实践应用中的辨识能力关系，可以作为文本分级的参数。

3.3.3 同音近义语素辨析

词汇系统中存在许多同音近义词。细微的语用差异成为评价和判断文本难度和阅读理解能力的重要参照指标。例如

查看—察看、考查—考察、审查—审察

查———察

拼音	词条	词性	词长	词次	词频
Cha2kan4	查看	v	2	26	0.0013
	察看	v	2	20	0.001
Kao3cha2	考察	v	2	12	0.0006
	考察	vn	2	1	0
	考查	vn	2	2	0.0001
Shen3cha2	审查	v	2	8	0.0004
	审查	vn	2	7	0.0003
	审察	v	2	3	0.0001

“查”构词动态分析

词条	词性	词长	词次	词频
查	v	1	98	0.0048
查看	v	2	26	0.0013
查明	v	2	11	0.0005
查询	v	2	8	0.0004
查账员	n	3	6	0.0003
查查	v	2	5	0.0002
查对	v	2	5	0.0002
查访	v	2	5	0.0002
查出	v	2	4	0.0002
查清	v	2	4	0.0002
查验	v	2	3	0.0001
查阅	v	2	3	0.0001
查	nr	1	2	0.0001
查禁	v	2	2	0.0001
查讫	v	2	2	0.0001
查找	v	2	2	0.0001
查查看	v	3	1	0
查处	vn	2	1	0
查考	v	2	1	0
查问	v	2	1	0
查询	vn	2	1	0
查证	v	2	1	0
合计			192	0.0089

词条	词性	词长	词次	词频
警察	n	2	85	0.0042
观察	v	2	84	0.0041
察觉	v	2	36	0.0018
觉察	v	2	30	0.0015
察看	v	2	20	0.001
察	vg	1	12	0.0005
考察	v	2	12	0.0006
观察	vn	2	9	0.0004
侦察	v	2	9	0.0004
警察局	n	3	7	0.0003
侦察	vn	2	7	0.0003
察言观色	i	4	6	0.0003
勘察	vn	2	6	0.0003
视察	v	2	6	0.0003
检察官	n	3	5	0.0002
洞察	v	2	4	0.0002
纠察队	n	3	4	0.0002
洞察力	n	3	3	0.0001
审察	v	2	3	0.0001
体察	v	2	3	0.0001
洞察	vn	2	2	0.0001
检察	b	2	2	0.0001
视察	vn	2	2	0.0001
察访	v	2	1	0
督察	v	2	1	0
监察	vn	2	1	0
检察长	n	3	1	0
勘察	v	2	1	0
考察	vn	2	1	0
侦察兵	n	3	1	0
侦察机	n	3	1	0

词条	词性	词长	词次	词频
察	vg	1	12	0.0005
察访	v	2	1	0
察觉	v	2	36	0.0018
察看	v	2	20	0.001
察言观色	i	4	6	0.0003
洞察	v	2	4	0.0002
洞察	vn	2	2	0.0001
洞察力	n	3	3	0.0001
督察	v	2	1	0
观察	v	2	84	0.0041
观察	vn	2	9	0.0004
合计			372	0.0174
监察	vn	2	1	0
警察	n	2	85	0.0042
警察局	n	3	7	0.0003
纠察队	n	3	4	0.0002
觉察	v	2	30	0.0015
勘察	vn	2	6	0.0003
勘察	v	2	1	0
考察	v	2	12	0.0006
考察	vn	2	1	0
明察秋毫	i	4	5	0.0002
审察	v	2	3	0.0001
视察	v	2	6	0.0003
视察	vn	2	2	0.0001
体察	v	2	3	0.0001
侦察	v	2	9	0.0004

3.4 目标阅读者等级常模量表

- 1 分年级教材生词用字量表
- 2 分年级常用词量表
- 3 分年级平均难字量表
- 4 分年级教材熟语量表
- 5 分年级固定短语量表
- 6 分年级外来词量表
- 7 分年级非汉字符号量表
- 8 分年级百科常用术语量表
- 9 分年级抽样测试常模量表

3.5 基于常模量表的难度系数

- 根据常模量表与实测指标的相关性曲线，计算平均值和平均差、标准差模型；
- 原型基本义—常用词汇义—义项引申义—修辞语用义—谐音双关义—认知隐喻义
- 近义词的语义距离：概念语义空间

讨论

- 1 阅读文本动态复杂性系统可控性策略；
- 2 基于语料库的“字-词-语-句”常模量表；
- 3 语用信息特征标记知识挖掘；
- 4 基于机器学习的文本分级专家系统可行性探索；
- 5 语用信息特征标记识别的规范化、标准化路线。

- 谢谢！

- 盛玉麒 鞠躬
- 2017-10-23