

The Web Library of Babel: evaluating genre collections

Serge Sharoff,[†] Zhili Wu,[†] Katja Markert[‡]

Centre for Translation Studies,[†] School of Computing[‡]

{S.Sharoff, Z.Wu, K.Markert}@leeds.ac.uk

1. Our tasks against existing research

1. the accuracy of AGI on various collections using various feature sets;
2. the distance between similar categories in different collections;
3. the accuracy across collections after mapping to a shared set of genres;
4. the agreement of human judgement on individual collections.

Source	# texts	# genres	Format
HGC (Stubbe and Ringstetter, 2007)	1412	34	HTML only
I-EN-Sample (Sharoff, 2009)	250	7	TXT from HTML
KI-04 (Meyer zu Eissen and Stein, 2004)	1205	8	HTML only
KRYS I (Berninger et al., 2008)	6200	70	PDF
MGC (Vidulin et al., 2007)	1536	20	HTML + images
SANTINIS (Santini, 2009)	1400	7	HTML only
Combined (Santini and Sharoff, 2009)	9849	8	TXT from HTML
Brown Corpus (Kučera and Francis, 1967)	500	10	TXT
BNC (Lee, 2001)	4053	70	TXT

Examples of genres No compatibility between collections

MGC (20 genres)	KI-04 (8 genres)	Santinis (8 genres)	I-EN-S (8 genres)
adult	informative	article	blog
blog	journalistic	discussion	eShop
childrens	official	download	faq
commercial	personal	help	frontpage
community	poetry	linklists	hotlist
content delivery	prose fiction	portrait-non priv	PHP
entertainment	scientific	portrait-priv	sitemap
error message	shopping	shop	SPage
FAQ	user input		notext
gateway			
index			

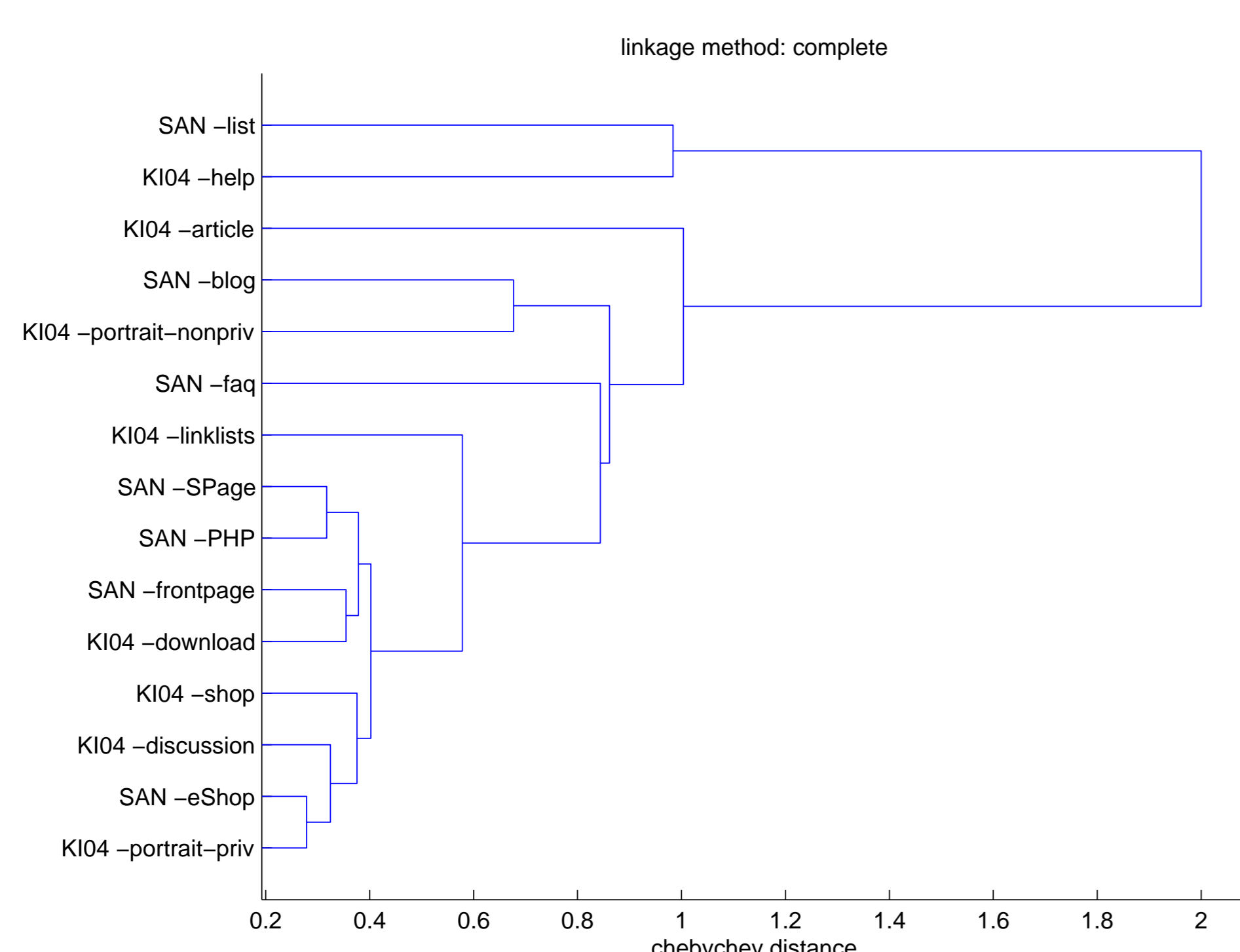
2. Accuracy for features

Features	HGC	I-EN-S	KI-04	KRYS-I	MGC	SAN	Comb	BNC	Brown
POS1	32.79	49.20	52.03	18.84	26.89	70.29	35.52	51.27	57.20
POS2	47.66	49.20	59.25	33.82	36.78	82.71	55.28	64.89	57.80
POS3	49.50	50.00	63.40	34.58	41.60	85.79	55.82	65.66	59.40
POS4	47.10	44.00	63.32	31.11	39.84	85.07	55.71	63.39	54.20
Char2	54.39	49.60	76.10	44.63	42.77	90.93	59.76	69.95	56.60
Char2b	53.33	41.20	73.03	38.58	43.29	93.07	49.80	55.81	54.00
Char3	59.99	54.80	80.00	51.35	49.54	93.93	63.58	72.49	65.40
Char3b	63.31	54.40	81.91	57.77	53.26	96.21	62.36	71.80	62.60
Char4	59.91	52.40	79.25	50.90	50.91	94.43	65.22	73.62	64.80
Char4b	65.51	55.20	85.81	61.87	55.14	97.14	66.89	74.54	65.80
Char5	57.65	52.00	78.42	49.40	49.87	94.21	66.54	72.59	64.20
Char5b	65.72	56.80	85.48	61.85	56.45	97.14	68.90	75.33	65.40
W1	61.69	54.80	81.83	54.02	51.63	94.79	67.41	71.50	61.60
W1b	59.06	60.40	84.15	59.05	51.63	95.86	63.31	75.03	64.00
W2	54.11	44.00	79.17	49.97	48.89	91.29	64.18	67.33	50.00
W2b	57.15	49.60	79.17	53.55	47.59	92.86	65.52	73.60	55.40

How do you know the genres? Features with significant weights:

MGC, shop	zon. trax garm rex. armi .?10 e et etr wayp waa waas peic alk. pei y aa bc16 lkak eice psu. bci
SAN, eshop	offe ord ifer rder tome orde ent. news only the .int pric mer. ine bask te m poun ment &pou und ;
KI-04, shop	lsen usab book osau oddl ord todd pric aur. fist droi rice nosa. you dino .boo hop. shop d bo s pf
HGC Help	acit epai fus q: w cito ac. faq volt adap otok okan kata dojo. doj ob r redm shod dapt kara a: t
I-EN-S, instr	., y .yo or d rem g it liv ? w ? wh ., e er w anot ly a et t ften eing houil ll t .oft bein iden
KI-04, FAQs	n't . which on't stio ? th doe of does do. tc. houil que frit tc a tc w llad pall .tr 'tru ete-
MGC, FAQs	12; i usi ivex vuln lner .tcp .sue -wri .tc. tc w tc a e tc t tc .tr 'tru f tc frit ritz e . " e "
SAN, FAQ	ces: : pu ing. do i opic lica . ho orm . tro cyc urri must rric pica clon trop lone cycl tax . yclo

3. Comparing labels



4. Cross-testing the collections

test on →	HGC	I-EN-S	KI04	KRYS-I	MGC	SAN	BNC	BROWN
# mapped	1329	250	1205	4360	1305	1400	755	436
HGC	63.31	34.00	33.86	38.10	41.99	40.71	39.34	38.30
I-EN-S	35.59	54.40	25.31	25.64	27.82	28.07	47.15	29.13
KI04	35.14	33.60	81.91	33.07	32.26	56.36	34.83	24.08
KRYS-I	38.68	29.60	27.80	57.77	32.03	21.79	47.81	55.50
MGC	46.95	37.60	34.52	36.33	53.26	38.00	55.23	44.04
SAN	31.98	22.80	40.41	13.03	23.60	96.21	20.53	4.13
BNC	37.77	42.00	23.82	29.13	34.02	19.36	71.80	58.95
BROWN	28.07	21.60	20.83	34.01	25.44	11.29	45.17	62.60

Examples of mapping MGC{adult, official, user input} → ∅;
MGC{childrens, poetry, prose} → recreation;

5. Human agreement

alpha chance-corrected agreement measure that allows annotation with multiple labels for a single item

Source	annotated	PA	alpha	# reliable cats?
KI-04	single	?	?	?
SANTINIS	single	?	?	?
KRYS I	double	0.5-0.6	?	?
HGC	double for 70 texts	0.76	?	?
MGC all	double	0.59	0.71	16 of 20
MGC Targeted	double	0.67	0.81	16 of 20
MGC Zeitgeist	double	0.47	0.56	4 of 20
MGC Random	double	0.50	0.55	5 of 20
I-EN-Sample	double	0.60	0.55	1 of 8

- Often lack of double annotation or no chance-corrected agreement measures.
- Reliability low for portions of the web that are randomly extracted.

This entails that current schemes are not representative for the whole Web!

6. Conclusions

The collections are not comparable Even when categories in two collections are described in a very similar way (e.g., FAQ, Help, Instruction), their actual content is considerably different. The accuracy of cross-classification is quite low.

The best set of features useful for AGI Character n-grams can capture many relevant generalisations not possible for other feature types, but their efficiency is often related to the ability to identify *topics* exemplifying particular genres in available collections.

Options for further AGI research We need a large reference corpus:

- collected from a diverse range of sources, and
- accompanied with a set of genre labels allowing consistently reliable annotation.
- We also need more research into detecting features which perform across a large number of texts.
- We need a range of corpora for several languages.
- We need methods for domain adaptation (Government of Canada vs. University of Leeds vs. Amazon).

7. Acknowledgements

We would like to thank the authors of each collection, who invested a lot of effort into producing them. We are also grateful to Google Inc for supporting this research via their Google Research Awards programme.

Resources: <http://corpus.leeds.ac.uk/serge/webgenres/>