Properties of Semantic Coherence Measures - Case of Topic Models

Author: Pirkko Pietiläinen
University of Oulu
pirkkoptlnn@gmail.com

UNIVERSITY OF OULU

A short resume: Some of the citations to my former work and a selection

- Feedback in information retrieval.
  Annual Review of Information Science and Technology
  by A Spink, RM Losee

- Machine learning: applications in expert systems and information retrieval
  dl.acm.org
  by R Forsyth, R Rada

- Abstracts of Articles in the Information Retrieval Area Selected by **Gerard Salton**
  ACM SIGIR Forum https://doi.org/10.1145/24634.1096830
  by Gerard Salton

- QUESTQUORUM A New Online Assistance tool from ESA-IRS (European Space Agency - Information Retrieval Services)
  by Sergio D'Elia, Pier Giorgio Marchetti

Coherence Measures – Topic Models

Applications of Topic Models in NLP:

IR, Classification, Content Analysis,
Data Mining, Sentiment Analysis,
Social Media Analysis, Word Sense Induction

Topics of size 10 and k=6

son	country	father	ancient	god	king	great	family	name	daught-er
album	band	song	music	series	film	releas-ed	video	featur-ed	movie
educa-tion	univer-sity	law	natio-nal	state	public	govern ment	elected	council	college
system	type	engine	sys-tems	can	using	use	used	stan-dard	struct-ure
war	army	forces	force	navy	naval	british	troops	military	fleet
park	located	road	area	league	country	south	city	railway	club

Measures used in prominent investigations include:

- NPMI, PMI
- cosine, Jaccard, Dice
- UCI
- UMass
- WordNet-based

Mixed results +

New Measures in 2015: Palmetto

- C_v best against human ratings

16 measures: 10 WN+ 6 Palmetto

- Topic Model: LDA + GloVe
- Corpus: 12 random samples from Wikipedia, 3 sizes
- Results are given as means of the 12 samples

Results: C_v is different

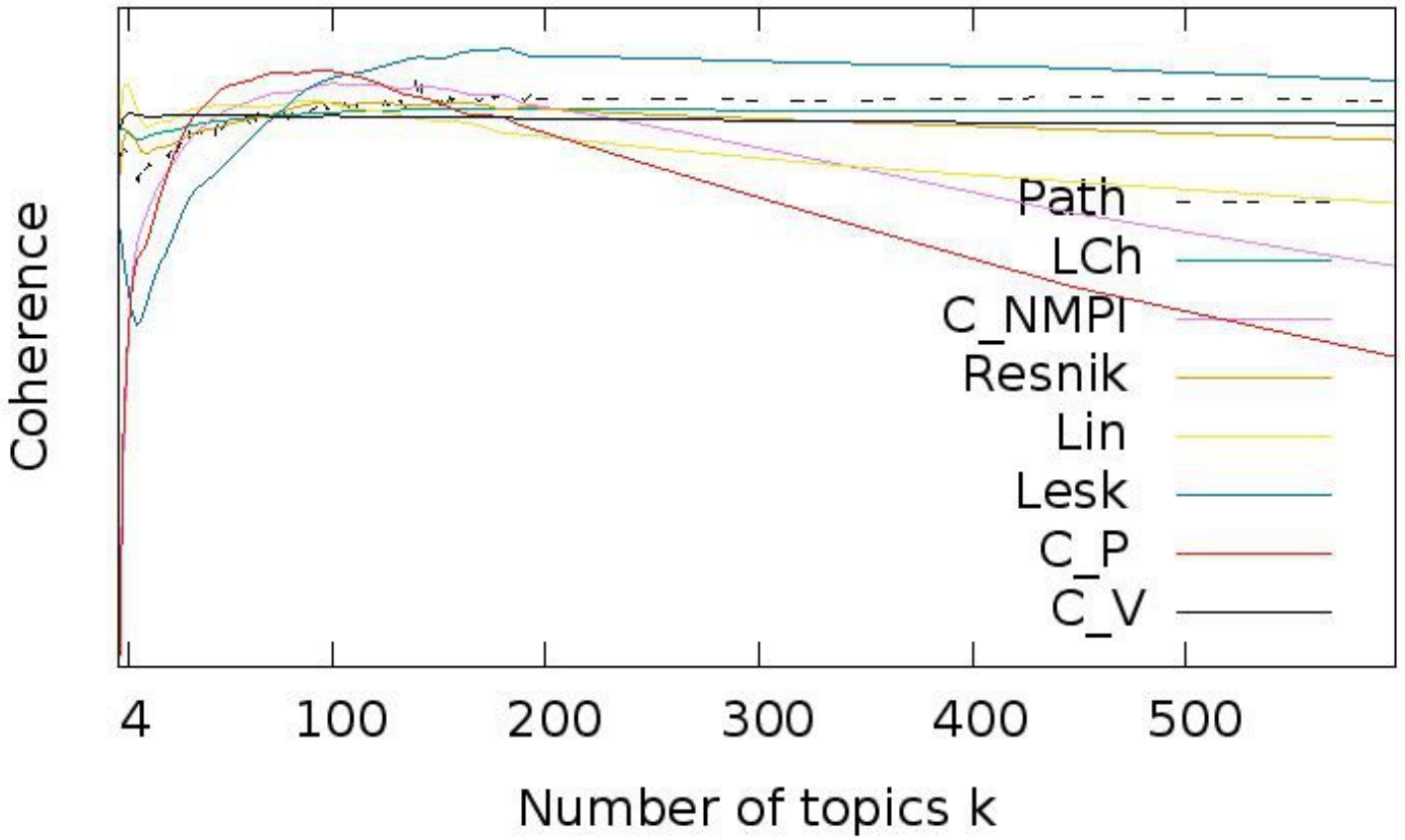
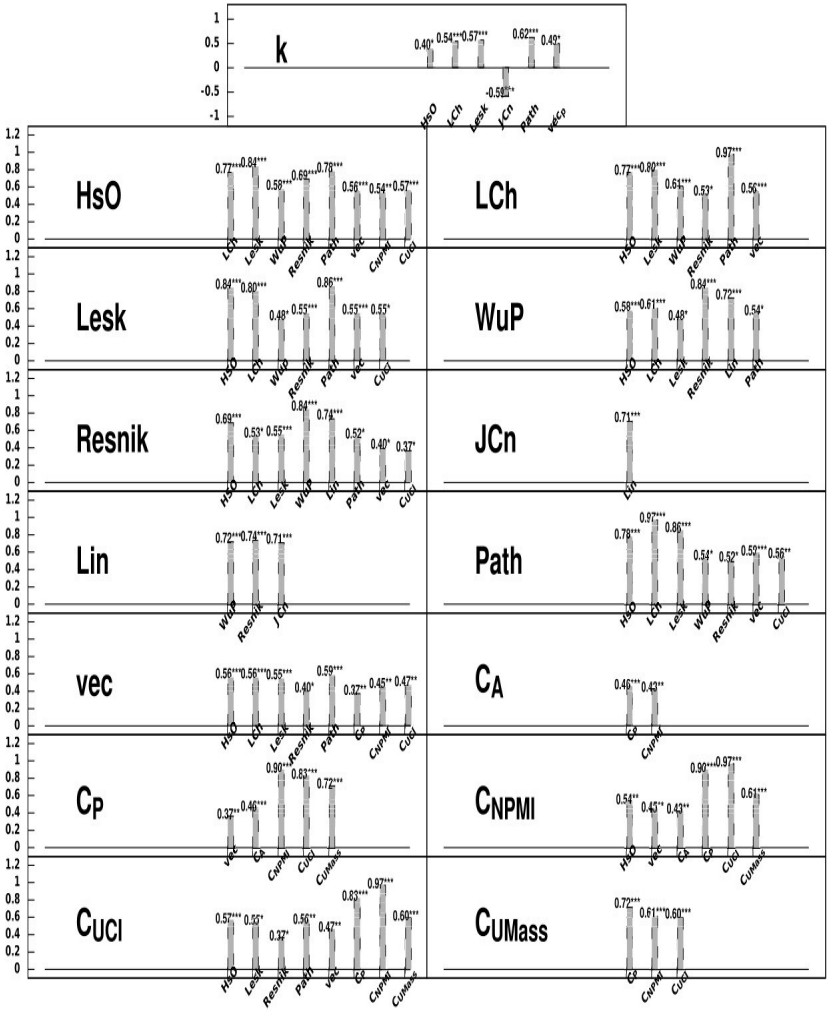


TABLE I. k-VALUES OF THREE HIGHEST COHERENCE VALUES FOR 12 CORPORA (A - D10) GIVEN BY 16 COHERENCE MEASURES ($HsO - C_{UMass}$).

	HsO	LCh	Lesk	WuP	Resnik	JCn	Lin	Path	vec_p	vec	C_A	C_P	C_V	C_{NPMI}	C_{UCI}	C_{UMass}
A	95	179	112	23	23	7	6	179	116	116	14	93	7	172	172	6
	112	450	115	7	6	23	7	450	178	102	9	138	9	95	181	25
	102	139	174	6	7	6	23	139	144	108	12	95	21	181	162	23
A20	164	146	143	7	7	6	7	143	124	36	12	99	4	95	95	77
	19	143	164	8	6	59	6	146	144	87	10	64	16	99	77	64
	90	132	173	10	12	7	93	144	146	60	7	151	24	59	183	62
A10	175	187	93	37	37	8	37	187	129	4	20	51	6	51	120	32
	93	145	164	93	93	93	8	145	163	6	51	102	76	88	51	52
	37	175	137	175	112	4	93	175	4	129	24	114	80	120	88	61
B	150	198	198	108	62	58	90	198	250	85	69	69	5	69	69	11
	196	147	164	89	90	54	62	147	92	89	7	11	6	81	158	10
	117	161	196	109	89	52	89	161	280	135	6	46	22	158	11	26
B20	170	143	129	33	33	33	33	149	5	67	10	101	5	66	101	10
	171	149	149	8	109	25	109	143	153	60	4	66	12	101	102	9
	190	187	171	5	63	48	5	170	181	142	14	10	6	95	95	25
B10	73	127	127	73	73	116	31	127	4	4	73	11	14	80	80	10
	146	146	181	64	105	31	73	147	5	135	23	80	7	88	68	11
	175	175	164	105	175	21	105	146	135	6	48	10	20	68	88	12
C	140	7	140	7	140	32	70	7	144	133	9	68	5	107	107	26
	155	8	193	70	113	104	98	5	143	106	17	107	11	140	140	41
	113	5	192	140	70	80	140	8	160	132	12	40	28	126	126	35
C20	50	153	188	11	11	11	11	133	132	111	8	140	5	157	157	33
	157	133	180	50	48	50	50	166	86	67	13	67	9	140	96	8
	144	166	144	48	50	10	48	153	133	152	14	81	7	96	140	14
C10	66	164	66	6	12	121	6	157	64	37	21	42	4	21	21	29
	69	189	103	17	140	4	12	189	117	48	9	9	8	16	22	19
	90	145	185	152	16	28	89	164	25	99	8	112	5	19	69	74
D	100	166	188	6	6	6	6	166	135	83	113	103	12	92	92	17
	149	7	191	7	10	183	7	143	185	146	9	113	113	113	189	19
	188	6	100	9	7	54	8	188	198	167	8	32	103	162	196	24
D20	97	116	107	48	48	4	48	144	7	72	12	78	41	109	109	14
	118	144	144	73	73	48	73	116	29	106	4	73	39	73	73	7
	107	169	184	20	91	91	46	169	197	91	16	55	40	100	100	16
D10	90	69	77	22	77	32	36	69	26	15	12	69	7	57	58	5
	79	141	79	36	90	36	12	280	39	45	6	57	8	58	57	64
	93	67	69	12	67	29	22	141	41	57	18	58	47	69	20	67


k
HsO
LCh
Lesk
WuP
Resnik
JCn
Lin
Path
vec
CA
CP
CNPMI
CUCI
CUMass

Human ratings

TABLE V. PEARSON AND SPEARMAN CORRELATIONS BETWEEN FOUR HUMAN RATINGS (MC - SIMLEX NOUNS) AND 16 COHERENCE MEASURES ($HsO - C_{\text{UMASS}}$). NOTE: HERE VALUES **without any** ASTERISKS ARE STATISTICALLY HIGHLY SIGNIFICANT WITH P<0.001. AND ** : P<0.01, AND * : P<0.05 , – : P>0.05 AND N.D. MEANS NO DATA.

	HsO	LCh	Lesk	WuP	Resnik	JCn	Lin	Path	vec_p	vec	C_A	C_P	C_V	C_{NPMI}	C_{UCI}	C_{UMass}
MC(P)	–	0.57*	–	0.55*	0.59	–	0.53*	–	0.60	**0.88**	–	0.79	–	0.77	0.67	–
MC(S)	–	0.58*	0.60	0.55*	0.68	–	0.56*	0.56*	0.70	**0.90**	–	0.81	0.65	0.82	–	–
RG(P)	0.54	0.60	0.44	0.53	0.61	–	0.54	0.54	n.d.	n.d.	–	0.75	–	**0.77**	0.71	–
RG(S)	0.49	0.56	0.55	0.51	0.55	–	0.46	0.54	n.d.	n.d.	–	**0.85**	0.50	0.84	0.83	0.45
Lau(P)	0.19	–	0.15	0.18	0.25	0.33	0.29	–	n.d.	n.d	0.38	**0.61**	0.31	0.55	0.51	0.28
Lau(S)	0.25	–	0.19	0.20	0.31	0.39	0.37	–	n.d.	n.d.	0.39	**0.52**	0.33	0.49	0.46	0.26
Simlex n.(P)	0.35	**0.52**	0.25	0.45	0.41	0.35	0.51	0.51	0.28	0.35	–	0.24	0.13	0.17	0.18	–
Simlex n.(S)	0.36	0.49	0.31	0.47	0.41	**0.51**	**0.51**	0.48	0.22	0.33	–	0.22	0.21	0.16	0.18	–

Conclusions

- The method used here is based on large data, is consistent and statistically tested
- WordNet-based and Palmetto-measures differ
- Large samples, different sizes + statistical testing → sample size to produce statistically significant results : 8000 documents / 2 million words
- Optimal number of topics k>100, except C_v

Further work

- Have a closer look at the human ratings studies and investigate why different data sets differ so much in respect of these 16 measures studied here
- Anonymous reviewer's suggestion:Try to find out what could explain the differences and similarities of the measures
- Data and R-code used in this study are available here.

Properties of Semantic Coherence Measures
-
Case of Topic Models

Author: Pirkko Pietiläinen
University of Oulu
pirkkoptlnn@gmail.com

Hello everybody!

This presentation is about measures, which are used in topic modeling.

Why topic modeling?
Because it has many applications in connection with the natural language processing: like information retrieval, classification, content analysis, data mining, sentiment analysis, social media analysis, word sense induction. There are more and more large volumes of data/text, and one of the methods to analyze them is topic modeling.

A short resume: Some of the citations to my former work and a selection

- Feedback in information retrieval.
  Annual Review of Information Science and Technology
  by A Spink, RM Losee
- Machine learning: applications in expert systems and information retrieval
  dl.acm.org
  by R Forsyth, R Rada
- Abstracts of Articles in the Information Retrieval Area Selected by **Gerard Salton**
  ACM SIGIR Forum https://doi.org/10.1145/24634.1096830
  by Gerard Salton
- QUESTQUORUM A New Online Assistance tool from ESA-IRS (European Space Agency - Information Retrieval Services)
  by Sergio D'Elia, Pier Giorgio Marchetti

Coherence Measures – Topic Models

Applications of Topic Models in NLP:

IR, Classification, Content Analysis,
Data Mining, Sentiment Analysis,
Social Media Analysis, Word Sense Induction

This presentation is about measures, which are used in topic modeling.

Why topic modeling?
Because it has many applications in connection with natural language processing: like information retrieval, classification, content analysis, data mining, sentiment analysis, social media analysis, word sense induction.
There are more and more large volume of data/text, and one of the methods to analyze them is topic modeling.

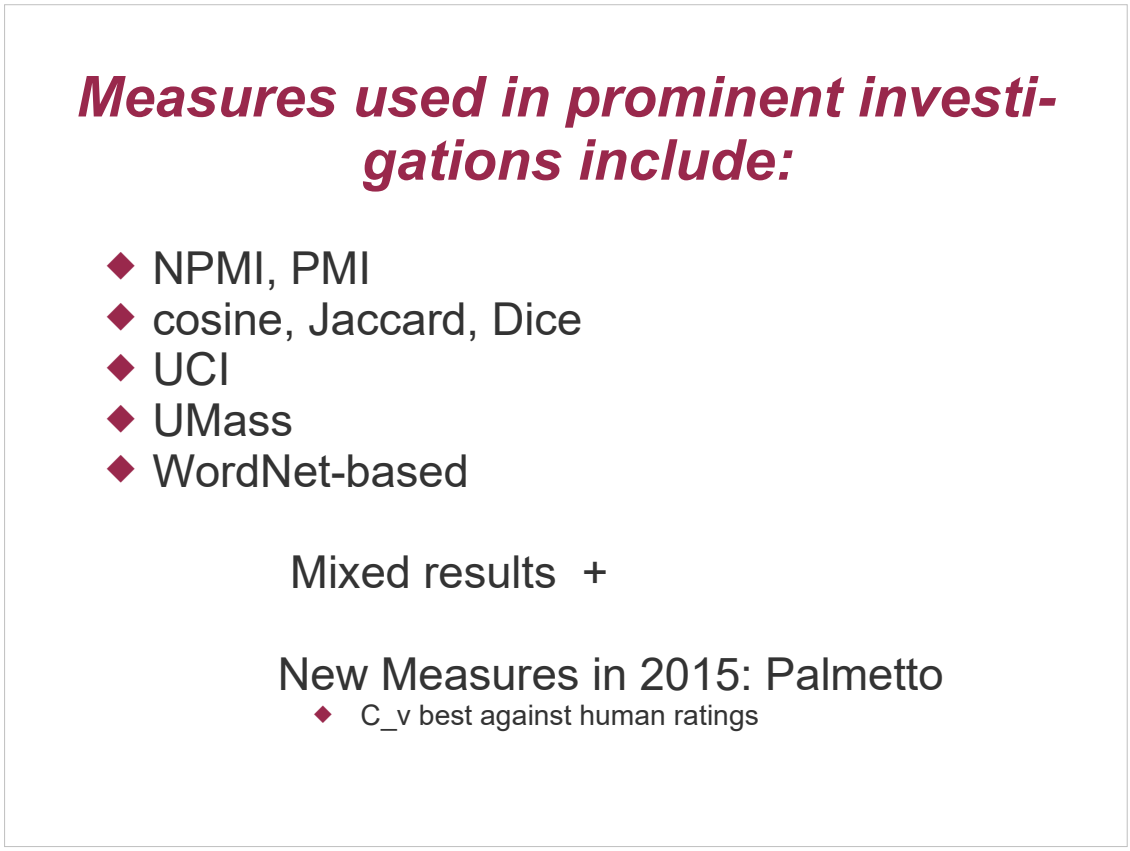
Topics of size 10 and k=6

son	country	father	ancient	god	king	great	family	name	daughter
album	band	song	music	series	film	released	video	featured	movie
education	university	law	national	state	public	government	elected	council	college
system	type	engine	systems	can	using	use	used	standard	structure
war	army	forces	force	navy	naval	british	troops	military	fleet
park	located	road	area	league	country	south	city	railway	club

An output from a Topic model is typically sets of 5 or 10 or 15 words. We use 10 words.

An important parameter is the number of topics, k. This example has 6 topics.

And to measure how well a model works we use measures of the coherence of these topics.
The more similar or related the topic words are, the more coherent a topic is.



There are many measures, which can be seen in the literature of the field.

One reason to do this investigation was the mixed results obtained in different studies.

Also there was recently introduced a new set of measures, which was especially designed to measure the topic coherence.

16 semantic coherence measures were selected to this study.
10 WordNet-based measures and 6 Palmetto measures.

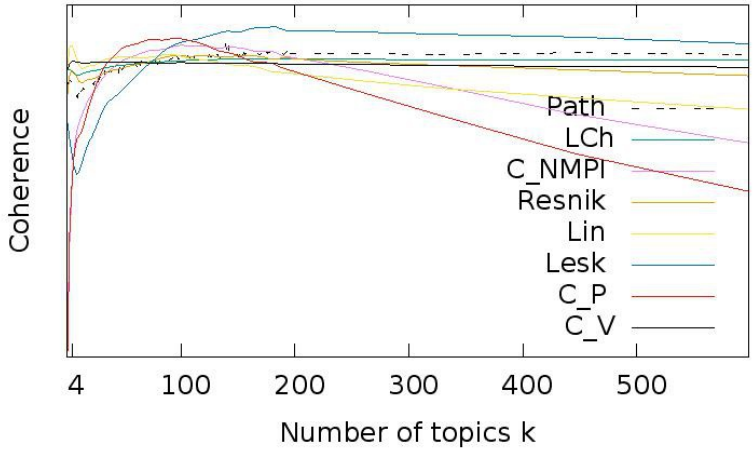
One detail we follow throughout this presentation is the Palmetto measure C_v, which has gained the highest human ratings of coherence in the evaluations done by the designers of the Palmetto measures.

16 measures: 10 WN+ 6 Palmetto

- Topic Model: LDA + GloVe
- Corpus: 12 random samples from Wikipedia, 3 sizes
- Results are given as means of the 12 samples

Our experimental set up was:
Latent Dirichlet Allocation (LDA) improved by Global Vectors.
Corpus was Wikipedia, from where 12 random samples was extracted, and samples had 3 different sizes
Results are given as averages of the 12 sample.

Results: C_v is different



Many of the measures reach their maximum around k =100 and even at higher k.
A notable exception is C_v, which has maximum close to k=40.

We note that measures Path and LCh behave very similarly, as is theoretically expected.

TABLE I. k-VALUES OF THREE HIGHEST COHERENCE VALUES FOR 12 CORPORA (A - D10) GIVEN BY 16 COHERENCE MEASURES ($HsO - C_{UMass}$).

	HsO	LCh	Lesk	WuP	Resnik	JCn	Lin	Path	vec_p	vec	C_A	C_P	C_V	C_{NPMI}	C_{UCI}	C_{UMass}
A	95	179	112	23	23	7	6	179	116	116	14	93	7	172	172	6
	112	450	115	7	6	23	7	450	178	102	9	138	9	95	181	25
	102	139	174	6	7	6	23	139	144	108	12	95	21	181	162	23
A20	164	146	143	7	7	6	7	143	124	36	12	99	4	95	95	77
	19	143	164	8	6	59	6	146	144	87	10	64	16	99	77	64
	90	132	173	10	12	7	93	144	146	60	7	151	24	59	183	62
A10	175	187	93	37	37	8	37	187	129	4	20	51	6	51	120	32
	93	145	164	93	93	93	8	145	163	6	51	102	76	88	51	52
	37	175	137	175	112	4	93	175	4	129	24	114	80	120	88	61
B	150	198	198	108	62	58	90	198	250	85	69	69	5	69	69	11
	196	147	164	89	90	54	62	147	92	89	7	11	6	81	158	10
	117	161	196	109	89	52	89	161	280	135	6	46	22	158	11	26
B20	170	143	129	33	33	33	33	149	5	67	10	101	5	66	101	10
	171	149	149	8	109	25	109	143	153	60	4	66	12	101	102	9
	190	187	171	5	63	48	5	170	181	142	14	10	6	95	95	25
B10	73	127	127	73	73	116	31	127	4	4	73	11	14	80	80	10
	146	146	181	64	105	31	73	147	5	135	23	80	7	88	68	11
	175	175	164	105	175	21	105	146	135	6	48	10	20	68	88	12
C	140	7	140	7	140	32	70	7	144	133	9	68	5	107	107	26
	155	8	193	70	113	104	98	5	143	106	17	107	11	140	140	41
	113	5	192	140	70	80	140	8	160	132	12	40	28	126	126	35
C20	50	153	188	11	11	11	11	133	132	111	8	140	5	157	157	33
	157	133	180	50	48	50	50	166	86	67	13	67	9	140	96	8
	144	166	144	48	50	10	48	153	133	152	14	81	7	96	140	14
C10	66	164	66	6	12	121	6	157	64	37	21	42	4	21	21	29
	69	189	103	17	140	4	12	189	117	48	9	9	8	16	22	19
	90	145	185	152	16	28	89	164	25	99	8	112	5	19	69	74
D	100	166	188	6	6	6	6	166	135	83	113	103	12	92	92	17
	149	7	191	7	10	183	7	143	185	146	9	113	113	113	189	19
	188	6	100	9	7	54	8	188	198	167	8	32	103	162	196	24
D20	97	116	107	48	48	4	48	144	7	72	12	78	41	109	109	14
	118	144	144	73	73	48	73	116	29	106	4	73	39	73	73	7
	107	169	184	20	91	91	46	169	197	91	16	55	40	100	100	16
D10	90	69	77	22	77	32	36	69	26	15	12	69	7	57	58	5
	79	141	79	36	90	36	12	280	39	45	6	57	8	58	57	64
	93	67	69	12	67	29	22	141	41	57	18	58	47	69	20	67

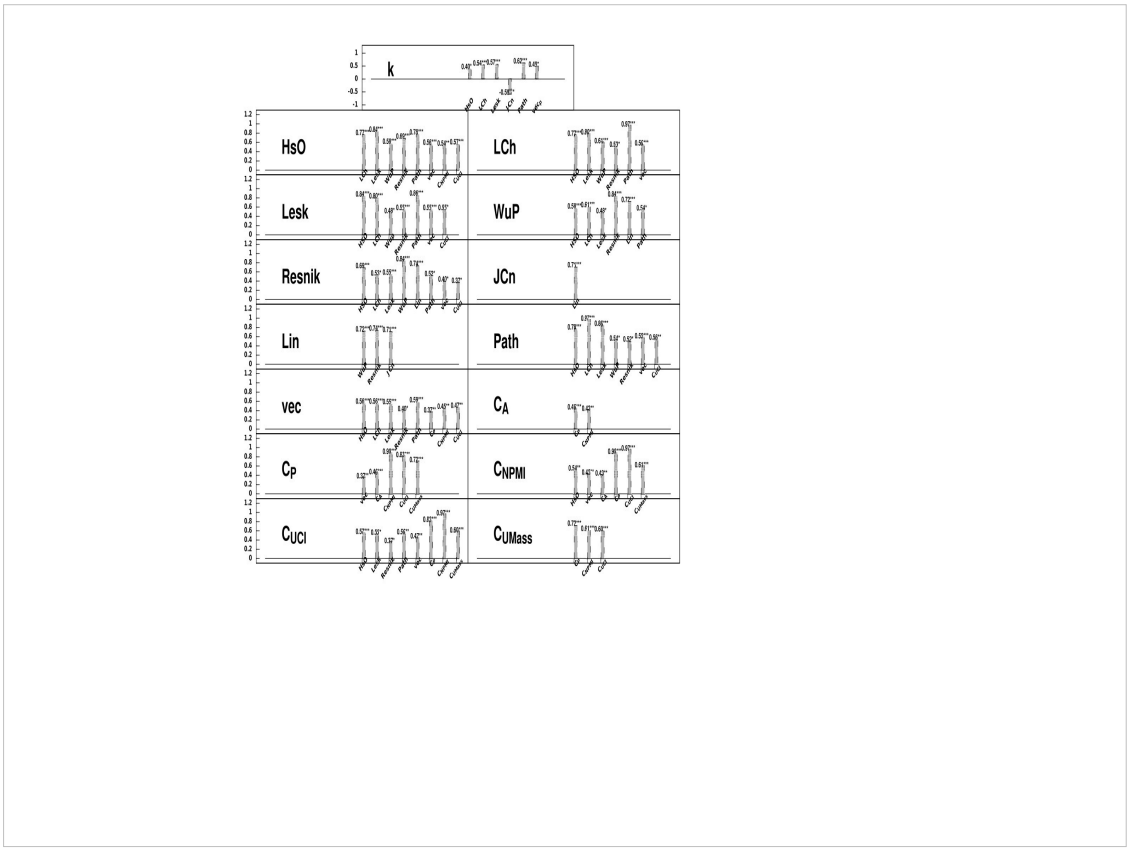
Because maximum values of a measure is often very close to the highest and the second highest local maxima, we list here the k-values of the 3 highest values of each measure in 12 corpora.

Colored cases indicate that there are no coincidences of k with any other measure (yellow for WordNet-based measures and green for Palmetto measures), and k-values are underlined when coincidences between the measure types occur.

We note again the similarity of LCh and Path telling about consistency of our method.

C_v and vec_p have less coincidences with any other measures.

Data and R-code to produce the table: https://www.pp.ouluresearch.fi



Data and R-code to produce the correlations and the tests of statistical significance: https://www.pp.ouluresearch.fi

There are only few statistically significant correlation between WordNet- and Palmetto-measures. We note also that C_v does not correlate with any other measure.

Relatively high correlation of number of topics k and some of the measures is an unexpected result.

Human ratings

TABLE V. PEARSON AND SPEARMAN CORRELATIONS BETWEEN FOUR HUMAN RATINGS (MC - SIMLEX NOUNS) AND 16 COHERENCE MEASURES ($HsO - C_{UMASS}$). NOTE: HERE VALUES **without any** ASTERISKS ARE STATISTICALLY HIGHLY SIGNIFICANT WITH P<0.001. AND ** : P<0.01, AND * : P<0.05 , – : P>0.05 AND N.D. MEANS NO DATA.

	HsO	LCh	Lesk	WuP	Resnik	JCn	Lin	Path	vec_p	vec	C_A	C_P	C_V	C_{NPMI}	C_{UCI}	C_{UMass}
MC(P)	–	0.57*	–	0.55*	0.59	–	0.53*	–	0.60	**0.88**	–	0.79	–	0.77	0.67	–
MC(S)	–	0.58*	0.60	0.55*	0.68	–	0.56*	0.56*	0.70	**0.90**	–	0.81	0.65	0.82	–	–
RG(P)	0.54	0.60	0.44	0.53	0.61	–	0.54	0.54	n.d.	n.d.	–	0.75	–	**0.77**	0.71	–
RG(S)	0.49	0.56	0.55	0.51	0.55	–	0.46	0.54	n.d.	n.d.	–	**0.85**	0.50	0.84	0.83	0.45
Lau(P)	0.19	–	0.15	0.18	0.25	0.33	0.29	–	n.d.	n.d	0.38	**0.61**	0.31	0.55	0.51	0.28
Lau(S)	0.25	–	0.19	0.20	0.31	0.39	0.37	–	n.d.	n.d.	0.39	**0.52**	0.33	0.49	0.46	0.26
Simlex n.(P)	0.35	**0.52**	0.25	0.45	0.41	0.35	0.51	0.51	0.28	0.35	–	0.24	0.13	0.17	0.18	–
Simlex n.(S)	0.36	0.49	0.31	0.47	0.41	**0.51**	**0.51**	0.48	0.22	0.33	–	0.22	0.21	0.16	0.18	–

C_v does not show very high correlations with any of the human ratings data sets studied here.

Again LCh and Path behave vary similarly.

Further studies why different data sets give so different results needs to be studied further.

Conclusions

- The method used here is based on large data, is consistent and statistically tested
- WordNet-based and Palmetto-measures differ
- Large samples, different sizes + statistical testing → sample size to produce statistically significant results : 8000 documents / 2 million words
- Optimal number of topics k>100, except C_v

Further work

- Have a closer look at the human ratings studies and investigate why different data sets differ so much in respect of these 16 measures studied here
- Anonymous reviewer's suggestion:Try to find out what could explain the differences and similarities of the measures
- Data and R-code used in this study are available here.