Properties of Semantic Coherence Measures - Case of Topic Models

Pirkko Pietiläinen

University of Oulu
Oulu, Finland
Email: pirkkoptlnn@gmail.com

***Abstract*—Measures of semantic relatedness and coherence are used in several Artificial Intelligence (AI) applications. Topic models is one of the fields where these measures have a role. In evaluating topic models, it is important to know well the properties of the used measure or measures. In this paper, it is first shown how 16 proposed coherence measures behave in finding the highest coherence in Latent Dirichlet Allocation (LDA) processing. With the collected exceptionally large corpus data from Wikipedia, it was then determined the correlations of the measures and the number of topics in LDA. From the average behavior of the measures, it is possible to conclude the range where the maximum values of coherence probably occur. Approximation of the size of a corpus giving statistically significant results in these respects is possible. Comparisons to human ratings are also included. The data and the R-codes for the calculations are made public. This paper explains many of the features affecting the use of coherence measures, including the roles of corpus/sample size, number of topics and the existence of local maxima of the measures. Differences of the measures and their correlations are also described.**

Keywords–Measuring Topic Coherence; LDA; Wikipedia; WordNet; Palmetto.

I. INTRODUCTION

Topic models are used in a wide range of Natural Language Processing (NLP) applications. Examples of application fields where they have been found useful include information retrieval [1], classification [2], content analysis [3], data mining [4], sentiment analysis [1], social media analysis [5] and word sense induction [6].

The evaluation of the quality of topics and the quality of the whole model can be done using direct methods, e.g., coherence metrics, or indirect methods where the quality is observed after a task performed with produced topics, e.g., measuring classification accuracy variance when done with different topic models. Only the direct methods are examined here.

Coherence measures are based on the idea that the more relatedness there is in a topic, the more coherent the topic is. Relatedness can be, e.g., semantic or based on co-occurrences of the topic words in a reference corpus, or the measures can be combinations of different aspects of coherence quantification.

Aletras and Stevenson [7] investigate the correlation between several coherence measures and ratings given by human evaluators and find out that Normalized Point Mutual Information (NPMI) coherence measure gives the best correlation in a number of tasks. Lau et al. [8] conclude that especially cosine-measure as well as Jaccard and Dice-measures outperform the NPMI-measure, because they receive higher correlations with human ratings in several experiments. A coherence measure based on calculation of word statistics was proposed by Mimno et al. [9] and Wikipedia was used as the corpus in the studies by Newman et al. [10], where they found that measures using word co-occurrence statistics perform better than WordNet-based methods. Röder et al. [11] developed a set of coherence measures and tested them against human ratings.

Stevens et al. [12] studied topic coherence over many models and with large number of topics. They used coherence measures known as UCI and UMass measures to evaluate the models. Of the models studied, they concluded that each has its own strengths; LDA was one of the models studied.

Given these mixed results, the present study was designed to examine coherence measures more closely. The research question is: What can be learned from a large scale study of semantic topic coherence measures to guide their usage and explain the present mixed results? Recently developed new measures designed for exactly this purpose were included along with old ones, which has been widely used. So, altogether, 16 semantic coherence measures and their role in topic modeling were selected to be included to this study.

To produce the topics studied, a method among latest improved LDA model [13] [14] was selected. Because the number of topics is an important parameter in performance optimization of a topic model, the topics studied were produced with an exceptionally wide range of number of topics.The study consists of 16 coherence measures, most of which are widely used.

The main contribution of this paper is the description of the behavior of the selected coherence measures in an enhanced LDA topic learning. This is done using exceptionally large data from Wikipedia, where an approximation of the corpus size needed to perform statistically significant experiments can be given. In order to investigate the relation of the number of topics, k, to the coherence measurement results, our experiments cover a wide range of k-values. Average coherence curves of the measures are presented and the consequences discussed. In addition to the maximum coherence, the closest local maxima are examined as well. The same extensive data is also used to determine the correlations between all the measures and their correlation with the number of topics. Human ratings of the coherence measures are also presented. Finally, some recommendations to the users of topic models and coherence measures are made.

The structure of the paper is such that the topic model used is introduced in Section II, and then the ways to measure coherence are presented in Section III. Experiments are described in Section IV and after that the results of our experiments are listed in Section V. Correlations with human ratings are

 ISBN: 978-1-61208-813-6

reported in Section VI, and in the final Section VII conclusions are drawn and what remains to be studied is discussed.

II. Latent Dirichlet Allocation

Unsupervised learning methods can be used to find latent topics from text corpora. One of the most used is LDA [15] and its many variants. The latest developments in topic models include incorporating to the models word vector representations trained on very large corpora. Instead of using only the words in the documents, the semantically related words from the corresponding word vectors are imported to the LDA process.

A topic model called Latent Feature LDA (LF-LDA) developed by Nguyen et al. [14] shows significant improvements on topic coherences when external word vectors are incorporated. Improvements can also be seen in classification and clustering tasks. For these reasons, the LF-LDA model by Nguyen et al. [14] is used in the present study.

Our preliminary experiments showed that in terms of coherence it is more feasible to use similar corpora as both training and actual corpus. For example, using word vectors trained on Google News [16] with Wikipedia corpus produces topics having lower coherences than when Wikipedia has been used as the training corpus as well. So, throughout the experiments of this paper, GloVe (Global Vectors) [17] word representations, which are pre-trained on Wikipedia, were chosen to be used.

III. Measuring coherence

Topic coherence measures can be divided to two groups according to whether they are planned specially for measuring topic coherence, or adapting to this purpose measures developed for other purposes. The first type of measures are the set of coherence measures recently proposed by Röder et al. [11]. They are described in the Subsection B called Palmetto-measures.

The more a topic word set contains words, that are semantically related with each other, the more interpretable and coherent the topic is. With this in mind, measures of semantic relatedness available in WordNet [18], are used here as well and the expression (1) is applied.

Topic coherence is usually defined as the average similarity of each word pair in a set of top-n most probable words produced by the topic model in use. Coherence C is usually calculated with the expression, see, e.g., [7]

$$C = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f(w_i, w_j)}{\binom{n}{2}} \quad (1)$$

Here $\{w_1, w_2, ...w_n\}$ are the topic words, and f are measures of semantic relatedness, like measures in Section III-A.

In this paper, the value of the number of topics k, for which the average coherence between topic words is highest, is called the optimal or best number of topics.

A. WordNet-based relatedness measures

Topics are considered coherent when their most probable words are semantically related. For this reason, the measures of semantic relatedness have also been used as coherence measures. WordNet [18] is a central resource in lexical semantics. By using WordNet it is possible to measure semantic similarity and relatedness between two concepts [19]. Ten measures have been developed, which use WordNet as their central resource, and therefore they are called WordNet-based measures. Six of these measure similarity and four of them measure the more general relatedness.

Similarity measures in WordNet are based on the hierarchy of concepts, and half of them quantify the similarity of two concepts using the most specific common ancestor of the pair of concepts, namely Jiang and Conrath (JCn) [20], Lin [21], and Resnik [22]. The rest of the similarity measures are based on the lengths of the paths between two concepts. They have been developed by Wu and Palmer (WuP) [23], Rada et al. (Path) [24], and Leacock and Chodorow (LCh) [25].

The four measures of relatedness are: Hirst and StOnge (HsO) [26], Lesk [27], and two vector measures [19]. The first one, Hirst and StOnge, makes use of the path direction and length between the concepts. The vector measures and Lesk measure calculate the relatedness using the definition texts of the concepts.

All ten WordNet-measures were used in this study, namely measures of HsO, LCh, Lesk, WuP, Resnik, JCn, Lin, Path, vec_p and vec. The first eight were obtained using the WS4J-package (WordNet Similarity for Java) [28] version 1.0.1 . Measures vec and vec_p [29] were from WordNet::Similarity [19].

B. Palmetto-measures

Unlike the Wordnet-based measures, a set of new measures was developed especially for topic coherence purposes by Röder et al. [11]. They first studied all the ways how to quantify coherence. Out of these quantifications they made a large number of combinations, and then investigated which ones correlated best with human ratings.

The ways of quantifications were: a) how to evaluate permutations of the top probable topic words. b) ways of computing probabilities of single words as well as joint probabilities of word pairs in an external reference corpus, and the size of sliding window was one parameter here, c) probabilistic confirmation measurement applied to quantifications a) and b), and finally, d) a huge number of combinations resulting from the former phases are combined to one single coherence measure. These measures were tested against several human rating data sets, and the best measure is called C_{V} and the second best is called C_{P}.

Measure C_{V} combines the indirect cosine measure with NPMI and with a sliding window size of 110 words. The second best, C_{P} combines Fitelson's [30] confirmation measure with a sliding window of 70 words. Four other previously proposed coherence measures were described and tested against human ratings in the same framework as the new ones. They are in the order of human test results: C_{NPMI}, C_{A} both proposed by Aletras [8], C_{UCI} proposed by Newman [10] and C_{UMass} was proposed by Mimno [9]. C_{NPMI} uses a window size of 5 words and C_{UCI} has window size of 10 words. As an external reference corpus, the English Wikipedia is always used.

Röder et al. [11] have made available both Java software and web-service possibilities to calculate six Palmetto-measures and all of these six measures are used in this study.

 ISBN: 978-1-61208-813-6

IV. EXPERIMENTS

To discover latent topics from corpora, LF-LDA [14] latent model is used. It differs from ordinary latent models in that it is improved by incorporating word vector representations or embeddings [17] to the model. The LDA [15] model has two hyper parameters, which are kept constant in all of these experiments: $\alpha = 50/k$ where k is the number of topics, and $\beta = 0.01$ following, e.g., Fang [31]. The mixture weight λ was set to 1.0, because it is one of the values often used in this type of connections - $\lambda = 0.6$ is also frequently used. The number of the most probable topical words was always 10. The number of topics k varied from 4 to 200, and sparse points between 250 and 600. Note, that in all of these experiments, LF-LDA is the only part with randomness in addition to random selection of Wikipedia samples.

As pre-trained vector representations, the 50-dimensional `vectors.6B.50d.txt` from the GloVe-project trained on 6 billion token corpus containing a Wikipedia 2014 dump with 1.6 billion tokens and Gigaword5 repository [17] were used. More details, e.g., information on available versions, can be found on the GloVe-project's web site [17].

Four consecutive, equal-sized samples from a 2010 Wikipedia corpus [32] were extracted. This corpus contains the raw text of the articles in the English part of the Wikipedia, only shorter than 2000 character long documents, links and navigation texts and other irrelevant material removed. Note that the vocabulary of a 2010 Wikipedia is a subset of the vocabulary of a 2014 Wikipedia, not the other way around. Stop-words and the words that were not included in the used GloVe-vectors (on the average 6.5 % of the remaining words) were removed. No lemmatization was performed, so that the corpus remained closer to the natural language. The starting point of the first sample was randomly selected. Then, 20% and 10% samples from those four original samples were extracted in order to get information on the effect of sample size. The properties of the twelve samples are given in Table I. To

TABLE I. PROPERTIES OF THE TWELVE CORPORA EXTRACTED FROM WIKIPEDIA.

sample size	documents	words	vocabulary size	names of the corpora
	$4 * 10^4$	10^7	$1.7 * 10^5$	**A,B,C,D**
20%	$8 * 10^3$	$2 * 10^6$	$8 * 10^4$	**A20,B20,C20,D20**
10%	$4 * 10^3$	10^6	$6 * 10^4$	**A10,B10,C10,D10**

demonstrate that topics of neighboring k-values, here $k = 6$ and $k = 7$, can be very similar, they are presented in Figures 1 and 2. This feature has consequences on the results, as can be seen later.

Topic0: son century father ancient god king great family name daughter
Topic1: album band song music series film released video featured movie
Topic2: education university law national state public government elected council college
Topic3: system type engine systems can using use used standard structure
Topic4: war army forces force navy naval british troops military fleet
Topic5: park located road area league county south city railway club

Figure 1. An example set of six topics. The words of each topic are permuted pairwise and 16 measures are obtained for each pair.

The words of each topic are permuted pairwise and 16 measures are obtained for each pair. The topic group averages

Topic0: education university law research school social college students national based
Topic1: album band music song film released series songs featured video
Topic2: located area railway park river town county city north near
Topic3: war army force military forces british united states december union
Topic4: season league championship games team cup championships football champion game
Topic5: son father god king daughter her emperor mother his lord
Topic6: engine type system using can systems used use surface design

Figure 2. A set of seven topics, $k = 7$, corpus **A**, has also one of the highest coherences.

for these 16 measures are calculated. The example set of six topics of corpus **A**, $k = 6$, in Figure 1 is present also on the first, second and third row of Table IV, meaning that this set has the highest coherence when the measures Lin and C_{UMass} are applied, the second highest when measured by Resnik and the third highest value when measures Wup and JCn are applied.

Because the topics are almost the same in Figures 1 and 2, the example set of Figure 2 can also be found in Table IV. It has the highest coherence when the measures JCn and C_{V} are applied, the second highest with WuP and Lin, and the third highest with Resnik. White areas of Table IV indicate that k-values co-occur within a corpus. The underlined k-values occur in both groups of measures: WordNet-based (on the left) and Palmetto-measures (on the right), and colored areas have no co-occurrences.

V. RESULTS

At first, it is important to look at examples of the semantic coherence measures considered here. Normalization to one is used with all the measures so that it is possible to make comparisons between them, because this type of normalization preserves the proportional relationships of the data. It is done by dividing each data value by the sum of the data values of the same object. As an example, for k=[4,600] the values of the measure LCh ranges from 1.3345959 to 1.5078795 and those of Path from 0.109415 to 0.14231707. When they are normalized by dividing all LCh values by the sum of all values between k=[4,600], which is 296.99, and doing the same to the Path measure respectively with the sum of Path values 26.79, the ranges take the values from 0.0044936978 to 0.0050771585 (LCh) and from 0.0040834493 to 0.005311379 (Path).

The normalized values of measures LCh and Path, when k=[4,600], are presented in Figure 3. It can be seen that both measures have many local maxima close to the maximum coherence value. This means that there are several almost as optimal number of topics, whose coherence values differ only a little. This property is repeated in all of the studied semantic coherence measures, both WordNet- (Figure 3) and Palmetto-measures (Figure 4).

The example of LCh and Path in Figure 3 is important in another aspect, too. They can be seen to find their maxima at the same k-values. That happens because LCh and Path have very similar functional shape in the areas in question. So, the two measures, for which the theoretically predicted behavior is similar, really exhibit similar behavior in our experiments. That is an indication of reliability of the present approach,

meaning that the predicted behavior is not disturbed by any part of our data processing. Because of the existence of

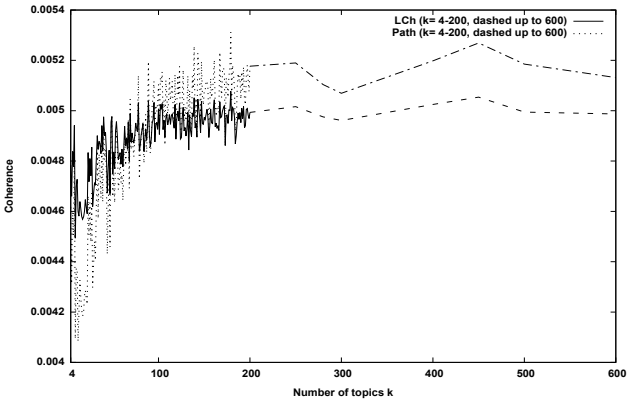

Figure 3. LCh and Path, both normalized to one, and as a function of Number of topics k in corpus **A.**

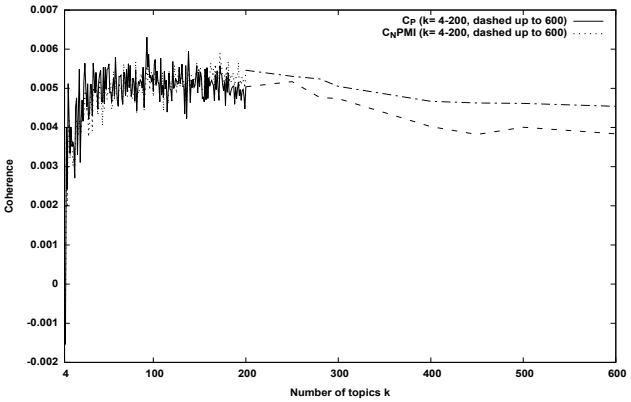

Figure 4. C_{P} and C_{NPMI}, both normalized to one, and as a function of Number of topics k in corpus **A.**

the many close local maxima in Table IV (see last page of this document) not only the maximum found by each measure but three highest coherence values of each measure are considered. Good examples supporting this decision are rows A10 columns C_{NPMI} and C_{UCI} in Table IV, where the three highest coherences are located at k = 51, 88, and 120 in this order for C_{NPMI}, but only the order differs from C_{UCI}. Six measures having most white areas in Table IV are listed in

TABLE II. SIX HIGHEST PERCENTAGES OF CO-OCCURRING NUMBER OF TOPICS k ON THREE TOPMOST COHERENCE VALUES IN TABLE IV.

C_{NPMI}	Lch	Path	Resnik	Lin	C_{UCI}
94%	92%	89%	89%	86%	81 %

Table II. These figures tell us that, e.g., 94% of top-3 k-values of measure C_{NPMI} occur also in some other measure's set of top-3 k-values.

A. Averages

First, the properties of the average behavior of the measures are presented.

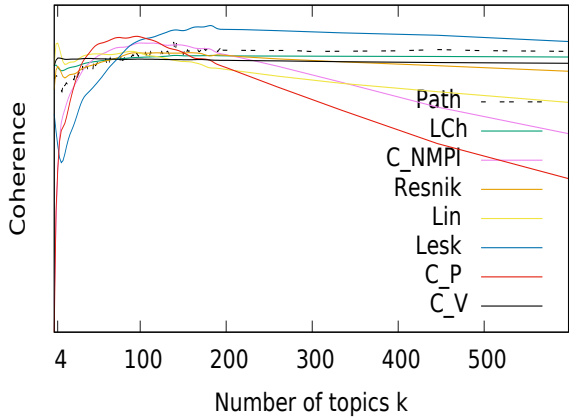

Figure 5. Averages over all twelve corpora of normalized C_{NPMI}, LCh, Path, Resnik, Lin, Lesk, C_{V} and C_{P} -measures.

Many typical properties of the semantic coherence measures are depicted in Figure 5. First, it can be noted that while there are similarities, there are big differences in the way they behave in the very low part in the k-axis. After about k=10, the curves show same type of behavior until after about k=100 their ways apart. Some seem to decrease, the others do not. The measures selected to Figure 5 are due to their appearance in Table II and Figure 6. In the light of this study, it is possible that the optimal number of topics can be found in the area of k = several hundreds.

B. Correlations

A good way of finding differences and similarities between the coherence measures is to examine their correlations with each other. A histogram of all statistically significant correlations between the semantic coherence measures is depicted in Figure 6. Note that the data is of general nature and includes an exceptionally large sample of word pairs. The data consists of N compared word pairs

$$N = \binom{10}{2} \sum_{k=4}^{K} k = 1\ 045\ 080, \qquad (2)$$

where $K = [200, 250, 280, 300, 400, 450, 500, 600]$. There are over one million measurements of similarity of word pairs for each corpus $\mathbf{A} - \mathbf{D20}$. To our knowledge, there is no other so large data collection used in finding the correlations of the semantic similarity measures.

The highest correlations, 0.97, occur between Path and LCh and between C_{NPMI} and C_{UCI}. These four measures are present also in Table II. The second best correlation 0.90 is found between C_{NPMI} and C_{P}. C_{P} does not appear in Table II, but is the third in Palmetto group with 61% of co-occurrences, or white areas in Table IV. Path is again participating with Lesk in the next highest correlation 0.86.

 ISBN: 978-1-61208-813-6

All correlation calculations with the tests of statistical significance are included in the additional material [33]. Both original data and the R-code for producing information in Figure 6 are included.

C. The effect of corpus size

The most surprising result was that there is so little statistically significant differences between variables in 100% corpus size and the smaller 20% or 10% sized, see Table I, corpora. The following properties against the corpus size were tested:

- **correlations between the sixteen measures**
  There is no significant difference of means of correlations between the groups of 100%, 20% and 10%.
- **average of three optimal number of topics in Table IV.**
  The biggest corpus has the highest average of the optimal number of topics 91.9 not differing from the next 20% sample significantly ($p = 0.07388$), where the average was 79.5. The smallest sample had the average 71.3, which does not differ significantly from the 20% but differs significantly ($p = 0.00263$) from the 100% group.
- **average co-occurrences of three optimal number of topics in Table IV.**
  The mean co-occurrence percentages were 60.9, 68.3 and 64.5, respectively, in the 100%, 20% and 10% groups, and there is no statistically significant difference between the groups.
- **WordNet- and Palmetto-measures as two separate groups.**
  The results are included in the next Section V-D.

D. Differences between WordNet- and Palmetto-measures

Co-occurrences in Table IV between WordNet- and Palmetto-measures do not have statistically significant differences, as WordNet-measures have on the average 67% co-occurences of the best three number of topics, and Palmetto-measures 60%, respectively. On the contrary, the means of three best number of topics of WordNet- and Palmetto- groups differ highly significantly ($p << 0.0001$).

There is only one highly significant correlation between any WordNet- and Palmetto-measures, namely between HsO and C_{UCI} with 0.57 correlation, as can be seen in Figure 6. That is also the highest correlation between WordNet- and Palmetto-measures. The correlations within each group are much higher. It is also noteworthy in Figure 6 that none of the Palmetto-measures have any correlation with the number of topics, whereas some of the WordNet-measures have relatively high correlations with the number of topics k. The effect of corpus size is also different in these groups. On the whole, when both types of measures are evaluated together, there is no difference of correlations between corpus sizes. The same is true with Palmetto-measures, but not with WordNet-measures, where correlations of 100% and 20% sized groups differ significantly ($p = 0.03$).

VI. Correlations with human ratings

Because coherence is measured using relatedness scores of word pairs, examples of data sets, which compare human judgements of the relations of two words are presented here.

Similarly as earlier in this study, the relatedness of word pairs of four well known human ratings data sets were measured. MC (Miller an Charles) [34] is the smallest one, consisting of only 28 word pairs, and there were 38 human annotators. RG (Rubenstein and Goodenough) [35] has 65 pairs and 51 annotators. Both data sets are available on the web [36]. Lau [37] collected coherence judgements for 600 topics using Amazon Mechanical Turk with a developed quality control of the annotations. Only the top-5 topic words data set was used here. Hill [38] collected human ratings of similarity of word pairs, and they had 500 annotators. These Simlex-datasets are also available on SemR-11 pages cited above. In our comparisons, Simlex subset of nouns, which consists of 666 noun pairs, was used.

There is the list of correlations between each coherence measure and human ratings in terms of MC, RG, Lau and Simlex in Table V. Pearson and Spearman correlations between human ratings RG, MC, Simlex nouns, and LAU data and ten WordNet-measures (HsO $-$ vec) and six Palmetto-measures ($C_{\mathrm{A}} - C_{\mathrm{UMass}}$) using the same measurement methods as earlier in this study. Statistical significance of the correlations are included in Table V.

Four examples indicate that the correlations tend to be lower with the bigger data sets, and the bigger the data set the more statistical significance is reached. Also there is no clear one measure with the highest human ratings. In addition it can be concluded that the behavior of WordNet- and Palmetto-measures differ with respect to human ratings data sets. For example Palmetto measures reach higher ratings with Lau data set, whereas WordNet-measures do the same with Simlex data set.

The average correlations of the data sets in Table V (at the end of the text of this document) were calculated in the same way as in Figure 6. Now, it is possible to compare the correlations in Figure 6, where the data consists of millions of word pairs, see Section V-B, to the statistically significant correlations of the measures in Table V, where the data is limited at most to 666 word pairs. Out of 16 measures five: Lesk, Lin, C_{P}, C_{NPMI} and C_{UCI}, have exact match with the results of Figure 6, when comparing the two highest correlation co-measure. For example, Lesk has the highest correlation with Path, and the second highest with HsO, just like in Figure 6, and the same is happening with the average correlations of the measures in Table V, and in the same order. As example of the consistency of the measures is WuP; it has the highest correlation with Resnik in both calculations, 0.84 in the Topic Model calculations, and 0.83 in the case of human ratings.

TABLE III. Average Pearson correlations of WuP in cases of Wikipedia data of Figure 6 and human ratings data sets of Table V.

WuP :	Resnik	Lin	LCh	HsO	Path
Figure 6	0.84	0.72	0.61	0.58	0.54
Table V	0.83	0.79	0.84	0.59	0.69

Seven of the measures have partial match, including dif-

ferent order of the highest and the second highest: HsO, LCh, WuP. Resnik, Path, C_{A} and C_{UMass}. Two of the rest of cases, vec_p and C_{A} did not reach statistical significance in results of Figure 6, and that's why they could not be compared. With JCn, only one co-measure reached statistical significance in Section V-B. The average correlation of measure vec is the highest with vec_p, and the second highest with Resnik. These results can be considered, for their part, to describe the consistency of the methods used in this study.

VII. Conclusion and Future Work

The paper analyzes the effect of different semantic coherence measures when determining a topic model. Of the other variables in topic modeling, this study addresses the variable corpus by calculating the results for twelve randomly chosen samples from Wikipedia, the variable of number of topics by using a wide range of number of topics (k= between 4 and 600). Word embeddings used, LDA parameters, average document length in corpora and other variables need to be taken into account in further studies.

The average coherence values of sets of ten topmost topic words show no clear maximum, as can be seen in Figure 3. Instead, there are many local maxima, which have very small differences in their coherence values. For these reasons, not only the number of topics corresponding to the maximum coherence, but also similarly k- values of the two second highest coherences, were listed in Table IV. It can be seen that many measures find the three highest coherence values, but not necessarily in the same order. This behavior supports methods that do not rely one the highest coherence value but use methods like coherence @n [31]. On average, these maxima appear mainly after $k \approx 100$, see Figure 5. So, conclusions made from studies using only smaller k-values might suffer from a lack of generality.

From our result in Section V-C, it can be concluded that increasing the sample size after a limit of 8000 documents with two million words, see Table I, does not have any effect on most of the results. So, an approximation for the minimum corpus size capable to produce general results in this respect can be given. For determining correlations and co-occurrences, this study shows that even a smaller corpus of 4000 documents is enough.

Although the used measure sets, Palmetto and WordNet, include similar elements, the results indicate, that there are differences as well, see Section V-D. The most notable difference is that correlations between these groups are substantially lower than correlations within each group. It is interesting that the measure reaching the highest correlation with human ratings in the study of Röder et al. [11], see Section III-B, does not correlate with any of the other 15 measures studied here, see Figure 6.

Users of the coherence measures studied here should also take into account the relatively high correlation between the number of topics k and some of the measures, as seen in Figure 6.

Different data sets of human ratings do not give similar results for the coherence measures studied here, see Table V. This leads us to think that further research with human ratings data sets is needed.

Appreciated is a comment pointing out that a more detailed discussion on why the measures studied show similarities and differences would be needed here. That is an excellent topic for a further investigation of the current topic.

 ISBN: 978-1-61208-813-6

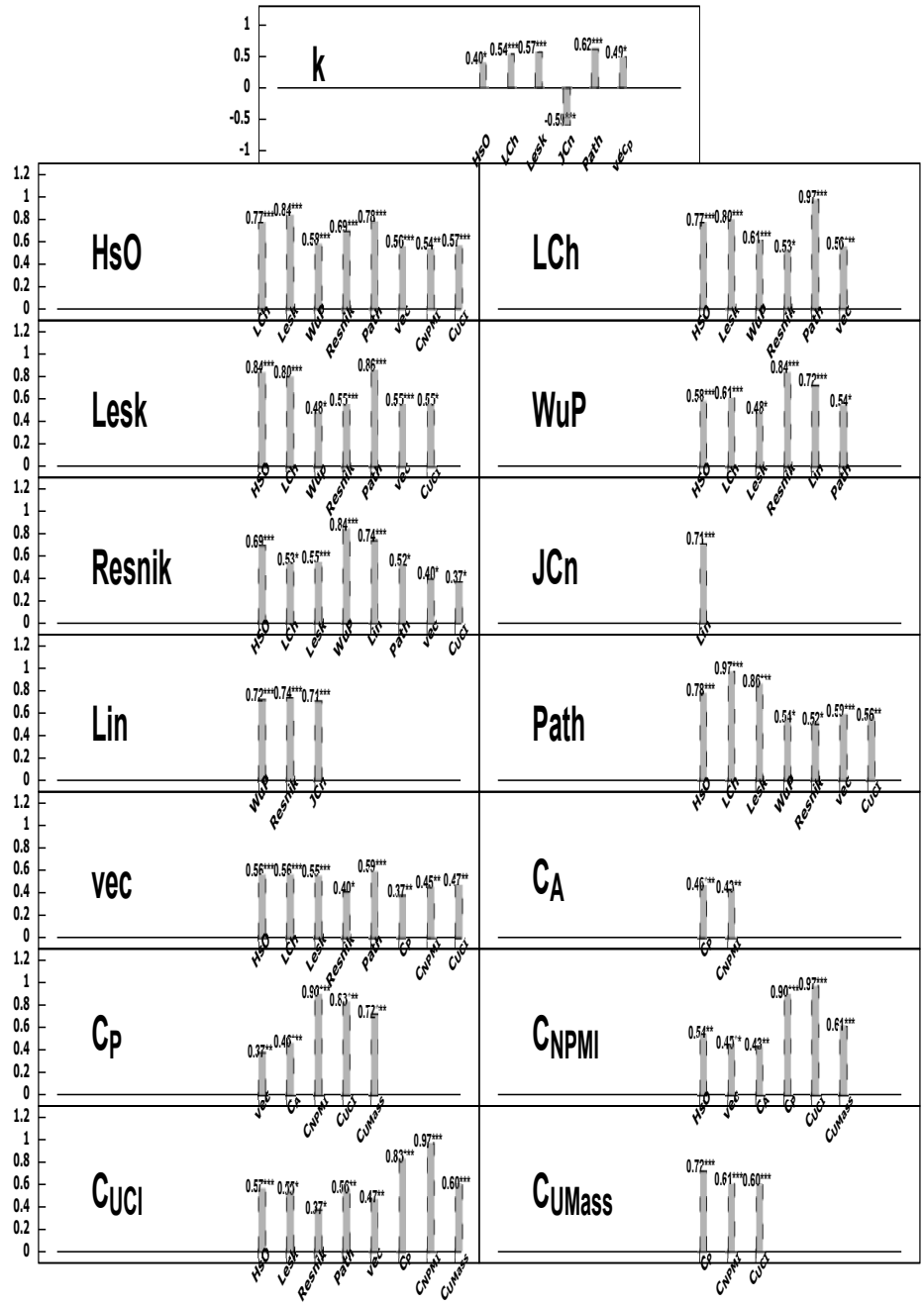

Figure 6. Average correlations of all 12 corpora between the coherence measures with each other and also with the number of topics k . *** means statistically highly significant with p<0.001, ** : p<0.01, and * : p<0.05.

TABLE IV. 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

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	–

 ISBN: 978-1-61208-813-6

REFERENCES

[1] J. L. Boyd-Graber, Y. Hu, and D. Mimno, Applications of topic models. now Publishers Incorporated, 2017, vol. 11.

[2] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," Machine learning, vol. 88, no. 1-2, 2012, pp. 157–208.

[3] J. Chuang et al., "Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations," in Advances in Neural Information Processing Systems workshop on human-propelled machine learning, 2014, pp. 1–9.

[4] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," Int. J. Adv. Comput. Sci. Appl.(IJACSA), vol. 6, no. 1, 2015.

[5] J. W. Mohr and P. Bogdanov, "Introduction—topic models: What they are and why they matter," 2013.

[6] L. Li, B. Roth, and C. Sporleder, "Topic models for word sense disambiguation and token-based idiom detection," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 1138–1147.

[7] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality." in EACL, 2014, pp. 530–539.

[8] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers, 2013, pp. 13–22.

[9] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011, pp. 262–272.

[10] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010, pp. 100–108.

[11] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015, pp. 399–408, https://github.com/AKSW/Palmetto/wiki/How-Palmetto-can-be-used, accessed: 2020-09.

[12] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012, pp. 952–961.

[13] H. Jelodar et al., "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," Multimedia Tools and Applications, vol. 78, no. 11, 2019, pp. 15 169–15 211.

[14] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," Transactions of the Association for Computational Linguistics, vol. 3, 2015, pp. 299–313, https://github.com/datquocnguyen/LFTM, accessed: 2020-09.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," the Journal of machine Learning research, vol. 3, 2003, pp. 993–1022.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Advances in neural information processing systems, 2013, pp. 3111–3119, https://code.google.com/archive/p/word2vec/, accessed: 2020-09.

[17] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), vol. 12, 2014, pp. 1532–1543, http://nlp.stanford.edu/projects/glove/, accessed: 2020-09.

[18] Princeton University, ""About WordNet. WordNet Princeton University." 2010, http://wordnet.princeton.edu, accesed: 2020-09.

[19] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet:: Similarity: measuring the relatedness of concepts," in Demonstration Papers at HLT-NAACL 2004. Association for Computational Linguistics, 2004, pp. 38–41.

[20] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in Proc of 10th International Conference on Research in Computational Linguistics, ROCLING97, 1997, pp. 19–33.

[21] D. Lin, "An information-theoretic definition of similarity." in Proc. of the 15th International Conference on MachineLearning, vol. 98, 1998, pp. 296–304.

[22] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, pp. 448–453.

[23] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994, pp. 133–138.

[24] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," Systems, Man and Cybernetics, IEEE Transactions on, vol. 19, no. 1, 1989, pp. 17–30.

[25] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," WordNet: An electronic lexical database, vol. 49, no. 2, 1998, pp. 265–283.

[26] G. Hirst and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," WordNet: An electronic lexical database, vol. 305, 1998, pp. 305–332.

[27] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," in Proceedings of the 18th international joint conference on Artificial intelligence, vol. 3, 2003, pp. 805–810.

[28] H. Shima, "WS4J-package (WordNet Similarity for Java)," 2014, https://code.google.com/p/ws4j/, accessed: 2020-09.

[29] S. Patwardhan, "Incorporating dictionary and corpus information into a context vector measure of semantic relatedness," Ph.D. dissertation, University of Minnesota, Duluth, 2003.

[30] B. Fitelson, "A probabilistic theory of coherence," Analysis, vol. 63, no. 3, 2003, pp. 194–199.

[31] A. Fang, C. Macdonald, I. Ounis, and P. Habel, "Examining the coherence of the top ranked tweet topics," in Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016, pp. 825–828.

[32] C. Shaoul and C. Westbury, "The Westbury Lab Wikipedia Corpus," Edmonton, AB: University of Alberta, 2010, psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html, accessed: 2020-09.

[33] P. Pietilänen, "Data and r-code, additional material to this article," 2020, https://pp.ouluresearch.fi, accessed: 2020-10.

[34] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," Language and Cognitive Processes, vol. 6, no. 1, 1991, pp. 1–28, https://doi.org/10.1080/01690969108406936, accessed: 2020-09.

[35] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," Communications of the ACM, vol. 8, no. 10, 1965, pp. 627–633.

[36] S. Barzegar, B. Davis, M. Zarrouk, S. Handschuh, and A. Freitas, "Semr-11: A multi-lingual gold-standard for semantic similarity and relatedness for eleven languages," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018, https://tinyurl.com/yyz7jvem, accessed: 2020-09.

[37] J. H. Lau and T. Baldwin, "The sensitivity of topic coherence evaluation to topic cardinality," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 483–487, https://github.com/jhlau/topic-coherence-sensitivity, accessed: 2020-09.

[38] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," Computational Linguistics, vol. 41, no. 4, 2015, pp. 665–695.

 ISBN: 978-1-61208-813-6