Exploring the Space of Topic Coherence Measures

Michael Röder
Leipzig University
R&D, Unister GmbH
Leipzig, Germany
roeder@informatik.uni-leipzig.de

Andreas Both
R&D, Unister GmbH
Leipzig, Germany
andreas.both@unister.de

Alexander Hinneburg
Martin-Luther-University
Halle-Wittenberg, Germany
hinneburg@informatik.uni-halle.de

ABSTRACT

Quantifying the coherence of a set of statements is a long standing problem with many potential applications that has attracted researchers from different sciences. The special case of measuring coherence of topics has been recently studied to remedy the problem that topic models give no guaranty on the interpretablity of their output. Several benchmark datasets were produced that record human judgements of the interpretability of topics. We are the first to propose a framework that allows to construct existing word based coherence measures as well as new ones by combining elementary components. We conduct a systematic search of the space of coherence measures using all publicly available topic relevance data for the evaluation. Our results show that new combinations of components outperform existing measures with respect to correlation to human ratings. Finally, we outline how our results can be transferred to further applications in the context of text mining, information retrieval and the world wide web.

Categories and Subject Descriptors

[**Document representation**]: Document topic models

General Terms

Measurement

Keywords

topic evaluation; topic coherence; topic model

1. INTRODUCTION

A set of statements or facts is said to be coherent, if they support each other. Thus, a coherent fact set can be interpreted in a context that covers all or most of the facts. An example of a coherent fact set is "the game is a team sport", "the game is played with a ball", "the game demands great physical efforts". A long standing open question is how to quantify the coherence of a fact set [5]. Approaches proposed in scientific philosophy have formalized such measures as functions of joint and marginal probabilities associated with the facts. Bovens and Hartmann [5] discuss many examples that lead to a demanding set of complex properties such a measure needs to fulfill. An example is the non-monotonic behavior of coherence in case of growing fact sets. The coherence of the two fact sets "the animal is a bird" and "the animal cannot fly" can be increased by adding the fact "the animal is a penguin". The non-monotonicity becomes apparent when the coherence is lowered again by adding non-related facts [5]. The discussion of coherence measures in that community deals mainly with schemes that estimate the hanging and fitting together of the individual facts of a larger set. Examples of such schemes are (i) to compare each fact against the rest of all other fact, (ii) compare all pairs against each other, and (iii) compare disjoints subsets of facts against each other. Such theoretical work on coherence from scientific philosophy—see [7] for an overview— has potential to be adapted in computer science, e.g., coherence of word sets.

Interest into coherence measures has arisen in text mining, as unsupervised learning methods like topic models give no guarantees on the interpreteability of their output. Topic models learn topics—typically represented as sets of important words—automatically from unlabeled documents in an unsupervised way. This is an attractive method to bring structure to otherwise unstructured text data. The seminal work of [13] proposed automatic coherence measures that rate topics regarding to their understandability. The proposed measures treat words as facts. This important restriction will apply to all analyses presented in this paper. An example of such set is {*game, sport, ball, team*}, which we will use throughout the paper to illustrate the methods. Furthermore, [13] restricts coherence to be always based on comparing word pairs. Our analyses will go beyond this point.

Evaluations in [13] based on human generated topic rankings showed that measures based on word co-occurrence statistics estimated on Wikipedia outperform measures based on WordNet and similar semantic resources. Subsequent empirical works on topic coherence [12, 18, 10] proposed a number of measures based on word statistics that differ in several details: definition, normalization and aggregation of word statistics and reference corpus. In addition, a new method based on word context vectors has been proposed recently [1].

WSDM'15, February 2–6, 2015, Shanghai, China.

ACM 978-1-4503-3317-7/15/02 ...$15.00.
http://dx.doi.org/10.1145/2684822.2685324.

Looking at the two lines of research on coherence, scientific philosophy and topic modelling, we note that the contributions are mainly complementary. While the former proposed a good number of schemes for comparing facts or words, the latter proposed useful methods for estimating word probabilities and normalizing numeric comparisons. However, a systematic, empirical evaluation of the methods of both worlds and their yet unexplored combinations is still missing.

Human topic rankings serve as the gold standard for coherence evaluation. However, they are expensive to produce. There are three publicly available sources of such rankings: first, Chang et al. [6] that have been prepared by Lau et al. [10] for topic coherence evaluation, second, Aletras and Stevenson [1] and third, Rosner et al. [16]. A systematic, empirical evaluation should take all these sources into account. For this reason, we choose the concept of a framework providing an objective platform for comparing the different approaches. Following our research agenda, this will lead to completely new insights of the behavior of different algorithms with regard to the available benchmarks. Hence, it will be possible to finally evaluate the reasons for specific behavior of topic coherences on a comparable basis.

Our contributions are these: First, we propose a unifying framework that spans a configuration space of coherence definitions. Second, we exhaustively search this space for the coherence definition with the best overall correlation with respect to all available human topic ranking data. This search empirically evaluates published coherence measures as well as unpublished ones based on combinations of known approaches. Our results reveal a coherence measure based on a new combination of known approaches that approximates human ratings better than the state of the art.[1] Finally, we note that coherence measures are useful beyond topic modelling. We discuss applications to search, advertising and automatic translation.

The rest of the paper is structured as follows: In section 2, we briefly review related work. Our unifying framework is introduced in section 3. Section 4 describes data preparation and evaluation results and section 5 discusses our findings. In section 6 the runtimes of the measures ar analyzed. Section 7 explains how coherence measures can be applied beyond topic modelling. Our conclusions are stated in section 8.

2. RELATED WORK

The evaluation of topic models needs next to holdout perplexity an additional measure that can rate topics with respect to understandability and interpretability by humans [6]. Measures based on evaluation of topic model distributions can produce useful results [2, 11]. However, they are difficult to link with human generated gold standards.

Newman et al. [13] represented topics by sets of top words and asked humans to rate these sets as good, neutral or bad. Several automatic topic ranking methods that measure topic coherence are evaluated by comparison to these human ratings. The evaluated topic coherence measures take the set of N top words of a topic and sum a confirmation measure over all word pairs. A confirmation measure depends on a single pair of top words. Several confirmation measures were evaluated. The coherence based on *pointwise mutual information* (PMI) gave largest correlations with human ratings. *UCI coherence* is calculated by[2]:

$$C_{\mathrm{UCI}} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{PMI}(w_i, w_j) \quad (1)$$

$$\mathrm{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \quad (2)$$

Probabilities are estimated based on word co-occurrence counts. Those counts are derived from documents that are constructed by a sliding window that moves over the Wikipedia, which is used as external reference corpus. Each window position defines such a document. For our example topic from section 1 we would calculate:

$$\begin{aligned} C_{\mathrm{UCI}} = \frac{1}{6} \cdot \big(&\mathrm{PMI}(\mathit{game}, \mathit{sport}) + \mathrm{PMI}(\mathit{game}, \mathit{ball}) \\ &+ \mathrm{PMI}(\mathit{game}, \mathit{team}) + \mathrm{PMI}(\mathit{sport}, \mathit{ball}) \\ &+ \mathrm{PMI}(\mathit{sport}, \mathit{team}) + \mathrm{PMI}(\mathit{ball}, \mathit{team})\big) \end{aligned} \quad (3)$$

Mimno et al. [12] proposed to use an asymmetrical confirmation measure between top word pairs (smoothed conditional probability). The summation of *UMass coherence* accounts for the ordering among the top words of a topic.[2]

$$C_{\mathrm{UMass}} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \quad (4)$$

Word probabilities are estimated based on document frequencies of the original documents used for learning the topics. The calculation for our example would be:

$$\begin{aligned} C_{\mathrm{UMass}} = \frac{1}{6} \Big(&\log(P(\mathit{sport}|\mathit{game})) + \log(P(\mathit{ball}|\mathit{game})) \\ &+ \log(P(\mathit{ball}|\mathit{sport})) + \log(P(\mathit{team}|\mathit{game})) \\ &+ \log(P(\mathit{team}|\mathit{sport})) + \log(P(\mathit{team}|\mathit{ball}))\Big) \end{aligned} \quad (5)$$

Stevens et al. [18] found that both—UCI and UMass coherence—perform better if parameter ϵ is chosen to be rather small instead of $\epsilon = 1$ as in respective original publications.

Aletras and Stevenson [1] introduced topic coherence based on context vectors for every topic top word. A context vector of a word w is created using word co-occurrence counts determined using context windows that contain all words located ± 5 tokens around the ocurrences of the word w. Largest correlation to human topic coherence ratings were found when defining the elements of these vectors as *normalized PMI* (NPMI) [4]. Additionally, they showed that restricting the word co-occurrences to those words that are part of the same topic performs best (*top word space*). Thus, the j-th element of the context vector $\vec{v_i}$ of word w_i has NPMI:

$$v_{ij} = \mathrm{NPMI}(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} \right)^{\gamma} \quad (6)$$

[1] Data and tools for replicating our coherence calculations are available at `https://github.com/AKSW/Palmetto` .

[2] ϵ is added to avoid logarithm of zero.
C_{UCI} as well as C_{UMass} can be used with the arithmetic mean or only with the sum of the single elements, because the mean calculation doesn't influence the order of evaluated topics that have the same number of top words.

For our example topic, the vector of its top word *game* would be calculated as:

$$\vec{v}_{game} = \big\{\text{NPMI}(game, game)^{\gamma}, \text{NPMI}(game, sport)^{\gamma}, \text{NPMI}(game, ball)^{\gamma}, \text{NPMI}(game, team)^{\gamma}\big\} \quad (7)$$

An increase of γ gives higher NPMI values more weight. Confirmation measures between pairs of context vectors are vector similarities like cosine, Dice or Jaccard that are averaged over all pairs of a topics top words like in [13].

$$C_{cos} = \frac{1}{6} \cdot \big(cos(\vec{v}_{game}, \vec{v}_{sport}) + cos(\vec{v}_{game}, \vec{v}_{ball}) + cos(\vec{v}_{game}, \vec{v}_{team}) + cos(\vec{v}_{sport}, \vec{v}_{ball}) + cos(\vec{v}_{sport}, \vec{v}_{team}) + cos(\vec{v}_{ball}, \vec{v}_{team})\big) \quad (8)$$

Alternatively, topic coherence is computed as average similarity between top word context vectors and their centroid $\vec{v}_c$.

$$\vec{v}_c = \vec{v}_{game} + \vec{v}_{sport} + \vec{v}_{ball} + \vec{v}_{team} \quad (9)$$

$$C_{cen} = \frac{1}{4} \cdot \big(cos(\vec{v}_{game}, \vec{v}_c) + cos(\vec{v}_{sport}, \vec{v}_c) + cos(\vec{v}_{ball}, \vec{v}_c) + cos(\vec{v}_{team}, \vec{v}_c)\big) \quad (10)$$

Additionally, [1] showed that the UCI coherence performs better if the PMI is replaced by the NPMI.

Lau et al. [10] structured the topic evaluation in two different tasks—word intrusion and observed coherence. In the first task, an intruder word has to be identified among the top words of a topic. For the second task, topics have to be rated regarding their coherence, while ratings are compared to human ratings. Both tasks can be done for single topics or the whole topic model. [10] confirmed that the UCI coherence performs better with the NPMI.

Theoretical work on coherence of sets of statements in a broader sense are reviewed in [7]. We follow their notation but adapt presentation of measures to word coherence. Shogenji's [17] and Olsson's [14] coherences are defined as:

$$C_S = \frac{P(w_1, \ldots, w_N)}{\prod_{i=1}^{N} P(w_i)} \quad (11)$$

$$C_O = \frac{P(w_1, \ldots, w_N)}{P(w_1 \vee \ldots \vee w_N)} \quad (12)$$

The usage of these coherences for our example is straight forward:

$$C_S = \frac{P(game, sport, ball, team)}{P(game) \cdot P(sport) \cdot P(ball) \cdot P(team)} \quad (13)$$

$$C_O = \frac{P(game, sport, ball, team)}{P(game \vee sport \vee ball \vee team)} \quad (14)$$

Fitelson [8] evaluated a single word in the context of all subsets that can be constructed from the remaining words. The set of all subsets without word w_i is denoted by $S(i)$. Fitelson's coherence is defined by comparing the probability of the i-th word with every single set in $S(i)$:

$$C_F = \frac{\sum_{i=1}^{N} \sum_{j=1}^{2^{N-1}-1} m_f(w_i, S(i)_j)}{N \cdot (2^{N-1} - 1)} \quad (15)$$

$$m_f(w_i, S(i)_j) = \frac{P(W_i|S(i)_j) - P(W_i|\neg S(i)_j)}{P(W_i|S(i)_j) + P(W_i|\neg S(i)_j)} \quad (16)$$

Note that this approach takes relationships between word sets into account and goes beyond averaging confirmations between word pairs.[3]

Douven and Meijs [7] took this approach further by creating pairs of word subsets $S_i = (W', W^*)$. These pairs are tested whether the existence of the subset W^* supports the occurrence of the subset W'. This is done using several confirmation measures and has been adapted to the evaluation of topics by Rosner et al. [16]. The authors found that using larger subsets W' and W^* can lead to better performing coherence measures.

3. FRAMEWORK OF COHERENCE MEASURES

Our new unifying framework represents a coherence measure as composition of parts that can be freely combined. Hence, existing measures as well as yet unexplored measures can be constructed. The parts are grouped into dimensions that span the configuration space of coherence measures. Each dimension is characterized by a set of exchangeable components.

Coherence of a set of words measures the hanging and fitting together of single words or subsets of them. Thus, the first dimension is the kind of segmentation that is used to divide a word set into smaller pieces. These pieces are compared against each other, e.g., segmentation into word pairs. The set of different kinds of segmentation is $\mathcal{S}$. The second dimension is the confirmation measure that scores the agreement of a given pair, e.g., NPMI of two words. The set of confirmation measures is $\mathcal{M}$. Confirmation measures use word probabilities that can be computed in different ways, which forms the third dimension of the configuration space. The set of methods to estimate word probabilities is $\boldsymbol{\mathcal{P}}$. Last, the methods of how to aggregate scalar values computed by the confirmation measure forms the fourth dimension. The set of aggregation functions is Σ.

The workflow of our framework as shown in figure 1 comprises four steps. First, the word set t is segmentated into a set of pairs of word subsets S. Second, word probabilities P are computed based on a given reference corpus. Both, the set of word subsets S as well as the computed probabilities P are consumed by the confirmation measure to calculate the agreements φ of pairs of S. Last, those values are aggregated to a single coherence value c.

In summary, the framework defines a configurations space that is the cross product of the four sets $\mathcal{C} = \mathcal{S} \times \mathcal{M} \times \boldsymbol{\mathcal{P}} \times \Sigma$. In the following subsections, these four dimensions are explained in more detail.

3.1 Segmentation of word subsets

Following [7], coherence of a word set measures the degree that a subset is supported by another subset. The result of the segmentation of a given word set W is a set of pairs of subsets of W. The definition of a subset pair consists of two parts that are differently used by the following confirmation mearsures. The first part of a pair is the subset for which the support by the second part of the pair is determined. Most proposed coherence measures for topic evaluation compare pairs of single words, e.g., the UCI coherence. Every single

[3] In section 3.1, we will give an example for S^{one}_{any} which is the equivalent to the $(w_i, S(i)_j)$ pairs of Fitelson's coherence.

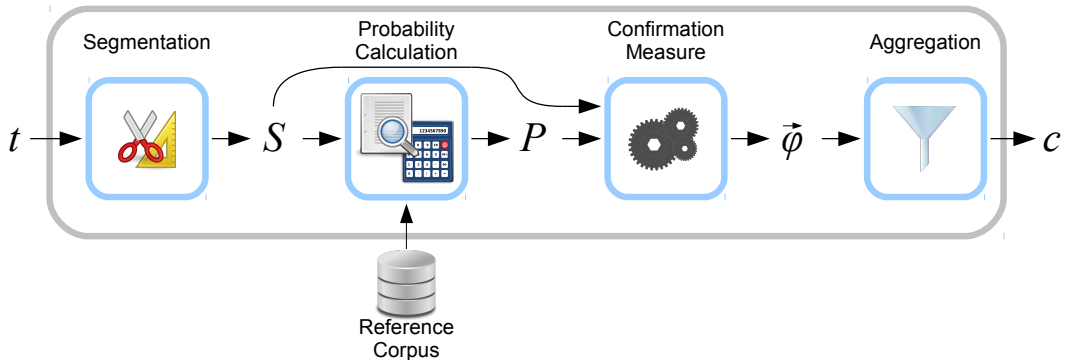

Figure 1: Overview over the unifying coherence framework—its four parts and their intermediate results.

word is paired with every other single word. Those segementations are called *one-one* and are defined as follows:

$$S^{one}_{one} = \Big\{(W', W^*)|W' = \{w_i\}; \\ W^* = \{w_j\}; w_i, w_j \in W; i \neq j\Big\} \quad (17)$$

$$S^{one}_{pre} = \Big\{(W', W^*)|W' = \{w_i\}; \\ W^* = \{w_j\}; w_i, w_j \in W; i > j\Big\} \quad (18)$$

$$S^{one}_{suc} = \Big\{(W', W^*)|W' = \{w_i\}; \\ W^* = \{w_j\}; w_i, w_j \in W; i < j\Big\} \quad (19)$$

The latter two are variations of the first one that require an ordered word set. They compare a word only to the preceding and succeeding words respectively, as done by the UMass coherence.

Douven and Meijs [7] proposed several other segmentations that have been adapted to topic evaluation by [16]. These definitions allow one or both subsets to contain more than one single word.

$$S^{one}_{all} = \Big\{(W', W^*)|W' = \{w_i\}; \\ w_i \in W; W^* = W \setminus \{w_i\}\Big\} \quad (20)$$

$$S^{one}_{any} = \Big\{(W', W^*)|W' = \{w_i\}; \\ w_i \in W; W^* \subseteq W \setminus \{w_i\}\Big\} \quad (21)$$

$$S^{any}_{any} = \Big\{(W', W^*)|W', W^* \subset W; W' \cap W^* = \emptyset\Big\} \quad (22)$$

S^{one}_{all} compares every single word to all other words of the word set. S^{one}_{any} extends S^{one}_{all} by using every subset as condition. S^{any}_{any} is another extension that compares every subset with every other disjoint subset. Figure 2 shows the different sets of subset pairs produced by applying the different segmentations to the running example.

The approach in [1] compares words to the total word set W using words context vectors. Therefore, we define another segmentation

$$S^{one}_{set} = \Big\{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W\Big\} \quad (23)$$

Note that this segmentation does not obey the requirement $W' \cap W^* = \emptyset$ stated in [7]. Therefore, it is only used together with coherence measures based on the ideas in [1].

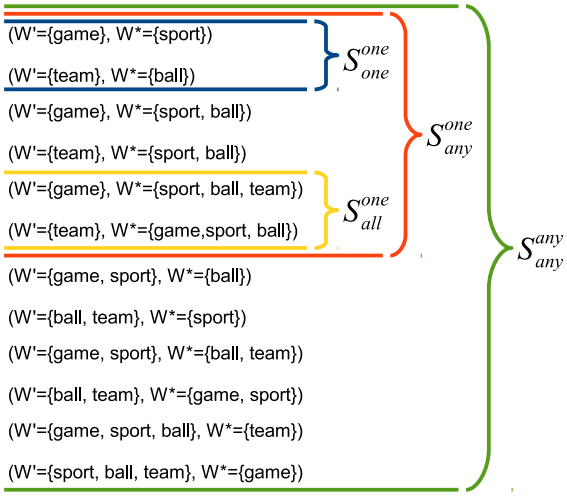

Figure 2: $S^{one}_{one}, S^{one}_{all}, S^{one}_{any}$ and S^{any}_{any} segmentations of the word set {game, ball, sport, team} and their hierarchy.

3.2 Probability Estimation

The method of probability estimation defines the way how the probabilities are derived from the underlying data source. *Boolean document* ($\mathcal{P}_{bd}$) estimates the probability of a single word as the number of documents in which the word occurs divided by the total number of documents. In the same way, the joint probability of two words is estimated by the number of documents containing both words divided by the total number of documents. This estimation method is called boolean as the number of occurrences of words in a single document as well as distances between the occurrences are not considered. UMass coherence is based on an equivalent kind of estimation [12]. Text documents with some formatting allow simple variations, namely the *boolean paragraph* ($\mathcal{P}_{bp}$) and *boolean sentence* ($\mathcal{P}_{bs}$). These estimation methods are similar to *boolean document* except instead of documents paragraphs or sentences are used respectively.

Boolean sliding window ($\mathcal{P}_{sw}$) determines word counts using a sliding window.[4] The window moves over the documents one word token per step. Each step defines a new virtual document by copying the window content. *Boolean document* is applied to these virtual documents to compute

[4]The window size is added to the name, e.g., $\mathcal{P}_{sw(10)}$ for a sliding window of size $s = 10$.

word probabilities. Note that *boolean sliding window* captures to some degree proximity between word tokens.

3.3 Confirmation Measure

A confirmation measure takes a single pair $S_i = (W', W^*)$ of words or word subsets as well as the corresponding probabilities to compute how strong the conditioning word set W^* supports W'. This could be done either directly as proposed in [7, 12, 13] or indirectly as done in [1].

3.3.1 Direct confirmation measures

Measures to directly compute the confirmation of a single pair S_i of words or word subsets are:

$$m_d(S_i) = P(W'|W^*) - P(W') \tag{24}$$

$$m_r(S_i) = \frac{P(W', W^*)}{P(W') * P(W^*)} \tag{25}$$

$$m_{lr}(S_i) = \log \frac{P(W', W^*) + \epsilon}{P(W') * P(W^*)} \tag{26}$$

$$m_{nlr}(S_i) = \frac{m_{lr}(S_i)}{-\log\left(P(W', W^*) + \epsilon\right)} \tag{27}$$

$$m_l(S_i) = \frac{P(W'|W^*)}{P(W'|\neg W^*) + \epsilon} \tag{28}$$

$$m_{ll}(S_i) = \log \frac{P(W'|W^*) + \epsilon}{P(W'|\neg W^*) + \epsilon} \tag{29}$$

$$m_c(S_i) = \frac{P(W', W^*)}{P(W^*)} \tag{30}$$

$$m_{lc}(S_i) = \log \frac{P(W', W^*) + \epsilon}{P(W^*)} \tag{31}$$

$$m_j(S_i) = \frac{P(W', W^*)}{P(W' \vee W^*)} \tag{32}$$

$$m_{lj}(S_i) = \log \frac{P(W', W^*) + \epsilon}{P(W' \vee W^*)} \tag{33}$$

In [7], the confirmation measures m_d, m_r and m_l are called difference-, ratio- and likelihood-measure. There, log-likelihood (m_{ll}) and log-ratio measure (m_{lr}) are also defined—the last is the PMI, the central element of the UCI coherence. Normalized log-ratio measure (m_{nlr}) is the NPMI. The log-conditional-probability measure (m_{lc}) is equivalent to the calculation used by UMass coherence [12]. The last two confirmation measures are the Jaccard and log-Jaccard measures.

A small constant ϵ is added to prevent logarithm of zero. Following [18], we set it to a small value ($\epsilon = 10^{-12}$).[5] Olsson's and Fitelson's coherences as well as a logarithmic variant of Shogenji's coherence (formulas 12, 16 and 11) are denoted by m_o, m_f and m_{ls}.

3.3.2 Indirect confirmation measures

Instead of directly computing the confirmation of $S_i = (W', W^*)$, indirect computation of confirmation assumes that given some word of W, direct confirmations of words in W' are close to direct confirmations of words in W^* with respect to this given word. Thus, indirect confirmation computes similarity of words in W' and W^* with respect to direct confirmations to all words.

[5]Additionally ϵ is used in m_l and m_{ll} for preventing division by 0.

Why is this an advantage? For example, assume word x semantically supports word z but they do not appear frequently together in the reference corpus and have therefore low joint probability. Thus, their direct confirmation would be low as well. However, the confirmations of x and z correlate with respect to many other words y in W. An example is that x and z are both competing brands of cars, which semantically support each other. However, both brands are seldom mentioned together in documents in the reference corpus. But their confirmations to other words like "road" or "speed" do strongly correlate. This would be reflected by an indirect confirmation measure. Thus, indirect confirmation measures may capture semantic support that direct measures would miss.

This idea can be formalized by representing the word sets W' and W^* as vectors with dimension of the total size of the word set W. Such vector can be computed with respect to any direct confirmation measure m. In case W' and W^* consist of single words, the vector elements are just the direct confirmations as suggested in [1]. In case W' and W^* are sets of more than one word, the vector elements are the sum of the direct confirmations of the single words. Following [1], the vector elements can be non-linearly distorted.

$$\vec{v}_{m,\gamma}(W') = \left\{ \sum_{w_i \in W'} m(w_i, w_j)^\gamma \right\}_{j=1,\ldots,|W|} \tag{34}$$

Given context vectors $\vec{u} = \vec{v}(W')$ and $\vec{w} = \vec{v}(W^*)$ for the word sets of a pair $S_i = (W', W^*)$, indirect confirmation is computed as vector similarity. Following [1], we equip our framework with the vector similarities cosine, dice and jaccard:

$$s_{cos}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \tag{35}$$

$$s_{dice}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} \min(u_i, w_i)}{\sum_{i=1}^{|W|} v_i + w_i} \tag{36}$$

$$s_{jac}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} \min(u_i, w_i)}{\sum_{i=1}^{|W|} \max(u_i, w_i)} \tag{37}$$

Thus, given a similarity measure sim, a direct confirmation measure m and a value for γ, an indirect confirmation measure $\tilde{m}$ is

$$\tilde{m}_{sim(m,\gamma)}(W', W^*) = s_{sim}(\vec{v}_{m,\gamma}(W'), \vec{v}_{m,\gamma}(W^*)) \tag{38}$$

3.4 Aggregation

Finally, all confirmations $\vec{\varphi} = \{\varphi_1, \ldots \varphi_{|S|}\}$ of all subset pairs S_i are aggregated to a single coherence score. Arithmetic mean (σ_a) and median (σ_m) have been used in the literature. Additionally, we evaluate geometric mean (σ_g), harmonic mean (σ_h), quadratic mean (σ_q), minimum (σ_n) and maximum (σ_x).

3.5 Representation of existing measures

Now we are ready to describe all coherence measures from section 2 as instances within the new framework. The coherences of [12, 13, 1] can be written as:

$$C_{\text{UMass}} = \left(\mathcal{P}_{bd}, S^{one}_{pre}, m_{lc}, \sigma_a\right) \tag{39}$$

$$C_{\text{UCI}} = \left(\mathcal{P}_{sw(10)}, S^{one}_{one}, m_{lr}, \sigma_a\right) \tag{40}$$

$$C_{\text{NPMI}} = \left(\mathcal{P}_{sw(10)}, S^{one}_{one}, m_{nlr}, \sigma_a\right) \tag{41}$$

Name	20NG	Genomics	NYT	RTL-NYT	RTL-Wiki	Movie
Topics	100	100	100	1095	1096	100
Top words	10	10	10	5	5	5
Documents	19 952	29 833	—	—	7 838	108 952
Paragraphs	155 429	2 678 088	—	—	319 859	2 136 811
Sentences	341 583	9 744 966	—	—	1 035 265	6 583 202
Tokens	2 785 319	114 065 923	—	—	13 679 052	86 256 415
Vocabulary	109 610	1 640 456	—	—	591 957	1 625 124

Table 1: Datasets used for evaluation.

The coherences defined in [7] and transformed in [16] can be written as:

$$C_{\text{one-all}} = (\mathcal{P}_{bd}, S^{one}_{all}, m_d, \sigma_a) \tag{42}$$

$$C_{\text{one-any}} = \left(\mathcal{P}_{bd}, S^{one}_{any}, m_d, \sigma_a\right) \tag{43}$$

$$C_{\text{any-any}} = \left(\mathcal{P}_{bd}, S^{any}_{any}, m_d, \sigma_a\right) \tag{44}$$

Shogenji's [17], Olsson's [14] and Fitelson's [8] coherences do not define how the probabilities are computed. Therefore, these measure definitions can be combined with every method of probability estimation[6]:

$$C_S = \left(\cdot, S^{all}_{one}, m_{ls}, \sigma_a\right) \tag{45}$$

$$C_O = \left(\cdot, S^{set}_{set}, m_o, \sigma_i\right) \tag{46}$$

$$C_F = \left(\cdot, S^{one}_{any}, m_f, \sigma_a\right) \tag{47}$$

Using the context-window-based probability estimation $\mathcal{P}_{cw}$ described in section 2, we are able to formulate the context-vector-based coherences defined in [1] within our framework[7]:

$$C_{\cos} = \left(\mathcal{P}_{cw(5)}, S^{one}_{one}, \tilde{m}_{cos(nlr,\gamma)}, \sigma_a\right) \tag{48}$$

$$C_{\text{dice}} = \left(\mathcal{P}_{cw(5)}, S^{one}_{one}, \tilde{m}_{dice(nlr,\gamma)}, \sigma_a\right) \tag{49}$$

$$C_{\text{jac}} = \left(\mathcal{P}_{cw(5)}, S^{one}_{one}, \tilde{m}_{jac(nlr,\gamma)}, \sigma_a\right) \tag{50}$$

$$C_{\text{cen}} = \left(\mathcal{P}_{cw(5)}, S^{one}_{set}, \tilde{m}_{cos(nlr,\gamma)}, \sigma_a\right) \tag{51}$$

This shows that the framework can cover all existing measures. However, it allows to construct new measures that combine the ideas of existing measures.

4. EVALUATION AND DATA SETS

The evaluation follows a common scheme that has already been used in [1, 10, 13, 16]. Coherence measures are computed for topics given as word sets that have been rated by humans with respect to understandability. Each measure produces a ranking of the topics that is compared to the ranking induced by human ratings. Following [10], both rankings are correlated using Pearson's r. Thus, good quality of a coherence measure is indicated by a large correlation to human ratings.

Word counts and probabilities are derived from Wikipedia. In case the corpus, which was used as training data for topic learning, was available, we computed coherence measures a second time using counts derived from that corpus.

During evaluation, we tested a wide range of different parameter settings. Window sizes varied between $[10, 300]$ for the sliding and $[5, 150]$ for the context window. The parameter γ varied in $\{1, 2, 3\}$. Overall, our evaluation comprises a total of 237 912 different coherences and parameterizations.

In literature, other evaluation methods has been used as well, e.g., humans were asked to classify word sets using different given error types [12]. However, since the data is not freely available, we can not use such methods for our evaluation.

A dataset comprises a corpus, topics and human ratings. Topics had been computed using the corpus and are given by word sets consisting of the topics top words. Human ratings for topics had been created by presenting these word sets to human raters. Topics are rated regarding interpretability and understandability using three categories—good, neutral or bad [13].

The generation of such a dataset is expensive due to the necessary manual work to create human topic ratings. Recently, several datasets have been published [1, 6, 10, 16]. Additionally, the creation of such a dataset is separated from the topic model used to compute the topics, since the humans rate just plain word sets without any information about the topic model [1, 6, 10, 15, 16]. This opens the possibility to reuse them for evaluation.

The *20NG* dataset contains the well known 20 Newsgroups corpus that consists of Usenet messages of 20 different groups. The *Genomics* corpus comprises scientific articles of 49 MEDLINE journals and is part of the TREC-Genomics Track[8]. Aletras and Stevenson [1] published 100 rated topics for both dataset, each represented by a set of 10 top words. Further, they published 100 rated topics that have been computed using 47 229 New York Times articles (*NYT*). Unfortunately, this corpus is not available to us.

Chang et al. [6] used two corpora, one comprising New York Times articles (*RTL-NYT*) and the other is a Wikipedia subset (*RTL-Wiki*). A number of 900 topics were created for each of these corpora. [10] published human ratings for these topics. Human raters evaluated word subsets of size five randomly selected from the top words of each topic. We aggregated the ratings for each word set. Word sets with

[6] S^{all}_{one} is used to formulate Shogenji's coherence and is defined like S^{one}_{all} but with W' and W^* swapped. $S^{set}_{set} = \{(W', W^*) | W', W^* = W\}$ is used for Olsson's coherence. Note that S^{set}_{set} contains only one single pair. Therefore, the aggregation function is the identity function σ_i. These special segmentation schemas are used only for these two coherences.

[7] We added window size to the model name, e.g., $\mathcal{P}_{cw(5)}$ for the window size $s = \pm 5$. The context window is only used for indirect coherence measures. For represent all measures mentioned in [1] instead of m_{nlr} the measures m_{lr} or $m_P = P(W', W^*)$ could be used respectively.

[8] `http://ir.ohsu.edu/genomics`

coherences	Name	C_V	C_P	C_{UMass}	$C_{\text{one-any}}$	C_{UCI}	C_{NPMI}	C_A
	$\mathcal{S}$	S^{one}_{set}	S^{one}_{pre}	S^{one}_{pre}	S^{one}_{any}	S^{one}_{one}	S^{one}_{one}	S^{one}_{one}
	$\mathcal{P}$	$\mathcal{P}_{sw(110)}$	$\mathcal{P}_{sw(70)}$	$\mathcal{P}_{bd}$	$\mathcal{P}_{bd}$	$\mathcal{P}_{sw(10)}$	$\mathcal{P}_{sw(10)}$	$\mathcal{P}_{cw(5)}$
	$\mathcal{M}$	$\tilde{m}_{cos(nlr,1)}$	m_f	m_{lc}	m_d	m_{lr}	m_{nlr}	$\tilde{m}_{cos(nlr,1)}$
	Σ	σ_a	σ_a	σ_a	σ_a	σ_a	σ_a	σ_a
using corpus	20NG	0.665	0.756	0.395	0.563	0.312	0.486	0.563
	Genomics	0.671	0.652	0.514	0.549	0.624	0.630	0.632
	RTL-Wiki	0.627	0.615	0.272	0.545	0.527	0.573	0.542
	Movie	0.548	0.549	0.093	0.453	0.473	0.438	0.431
	average	0.628	0.643	0.319	0.528	0.484	0.532	0.542
using the Wikipedia	$N=10$ 20NG	0.859	0.825	0.562	0.822	0.696	0.780	0.739
	$N=10$ Genomics	0.773	0.721	0.442	0.452	0.478	0.594	0.530
	$N=10$ NYT	0.803	0.757	0.543	0.612	0.783	0.806	0.747
	$N=10$ average	0.812	0.768	0.516	0.629	0.652	0.727	0.672
	$N=5$ RTL-NYT	0.728	0.720	0.106	0.438	0.631	0.678	0.687
	$N=5$ RTL-Wiki	0.679	0.645	0.350	0.499	0.558	0.606	0.602
	$N=5$ Movie	0.544	0.533	0.143	0.454	0.447	0.452	0.465
	$N=5$ average	0.650	0.633	0.200	0.464	0.545	0.579	0.585
	average	**0.731**	0.700	0.358	0.546	0.599	0.653	0.628

Table 2: Coherence measures with strongest correlations with human ratings.

less than three ratings or words with encoding errors are removed.[9]

The RTL-Wiki corpus is published in bag-of-words format that is unsuitable for paragraph-, sentence- or window-based probability estimations. Therefore, we have retrieved the articles in version of May 2009 from Wikipedia history records. Not all articles of the original corpus were available anymore. Therefore, the recreated corpus is slightly smaller than the original one.

Rosner et al. [16] published the *Movie* corpus—a Wikipedia subset—and 100 rated topics. Topics are given as sets of five top words. Like the RTL-Wiki corpus this corpus was recreated and is slightly smaller than the original one.

Table 1 shows an overview of sizes of the different datasets used for evaluation. All corpora as well as the complete Wikipedia used as reference corpus are preprocessed using lemmatization and stop word removal. Additionally, we removed portal and category articles, redirection and disambiguation pages as well as articles about single years.

5. RESULTS AND DISCUSSION

Table 2 shows the best performing coherence measures with respect to the different datasets.[10]

The largest correlations for all datasets (except for the Movie dataset) were reached, when the coherence measures relied on probabilities derived from the Wikipedia instead of the corpus used for topic learning. We focus our discussion on these calculations.

Most correlations computed for topics with 10 top words are higher than those of topics with 5 top words. Thus, evaluation of the topic quality is harder, if the number of top words N is small.

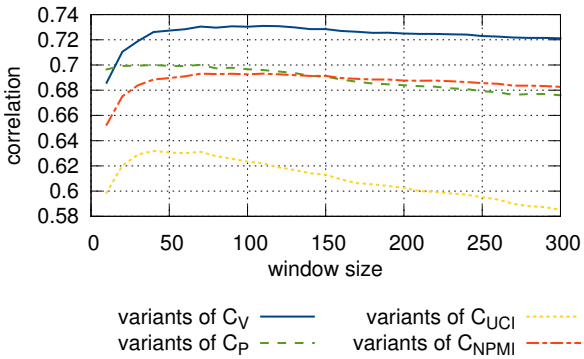

Figure 3: The influence of the sliding window's size on the correlation of variants of different coherences and human ratings.

Looking at already proposed coherence measures (five most right columns of table 2), our results confirm that on average UCI coherence performs better with NPMI. Among already proposed coherence measures, the NPMI showed best overall performance. Slightly lower correlations are obtained by the best performing vector-based coherence of those proposed in [1]. UMass coherence has lower correlations—especially in cases of small word sets. Shogenji's (0.034) and Olsson's (0.218) coherences (not shown in table 2) have low correlations, while Fitelson's coherence (0.541) is comparable to $C_{one-any}$ proposed in [16].

The best performing coherence measure (the most left column) is a new combination found by systematic study of the configuration space of coherence measures. This measure (C_V) combines the indirect cosine measure with the NPMI and the boolean sliding window. This combination has been overlooked so far in the literature. Also, the best direct coherence measure (C_P) found by our study is a new combination. It combines Fitelson's confirmation measure

[9]The RTL-Wiki dataset contains 23 word sets with 6 words and more than 3 ratings that were removed as well to ensure comparability of ratings

[10]Further results can be found in the supplementary material. Results differing to Aletras and Stevenson [1] might be caused by a different preprocessing and different versions of the Wikipedia.

Name	direct	indirect
$\mathcal{P}_{bd}$	0.648	0.664
$\mathcal{P}_{bp}$	0.679	0.698
$\mathcal{P}_{bs}$	0.664	0.691
$\mathcal{P}_{sw(70)}$	0.700	0.731
$\mathcal{P}_{sw(110)}$	0.668	0.731
$\mathcal{P}_{cw(50)}$	—	0.695
S^{any}_{any}	0.617	0.730
S^{one}_{all}	0.456	0.728
S^{one}_{any}	0.648	0.730
S^{one}_{one}	0.699	0.711
S^{one}_{pre}	0.700	0.711
S^{one}_{suc}	0.695	0.711
S^{one}_{set}	—	0.731
σ_a	0.700	0.731
σ_g	0.468	0.606
σ_h	0.457	0.590
σ_m	0.659	0.730
σ_n	0.482	0.573
σ_q	0.648	0.716
σ_x	0.513	0.670

Table 3: Best correlations for the probability estimations, segmentations and aggregations if they were combined with a direct or indirect confirmation measure.

with the boolean sliding window. This one is still better than published measures.

Among probability estimation methods, the boolean paragraph, boolean sentence and context window methods performed better than the boolean document (see table 3). The boolean sliding window performed best, but the window size should be larger than proposed in [13]. Figure 3 shows the correlation to human ratings achieved by variants of C_V, C_P, UCI and NPMI with different window sizes. It shows that only the C_P coherence has a very high correlation with a small window ($s = 10$). The correlation of C_V and NPMI remains on a high level, when the window size is larger than 50. The UCI coherence benefits from a larger window size, too, and reaches its best correlation at $s = 70$. An explanation for the good performance of the boolean sliding window is that it implicitly represents distances between word tokens within large documents. Further, large documents that are known to have good quality in Wikipedia, are implicitly up weighted because they contain more windows than smaller documents.

Among the segmentation methods, if a direct confirmation measure is used the single-word-based segmentation methods (S^{one}_{one}, S^{one}_{pre} and S^{one}_{suc}) have the highest and S^{one}_{all} has the worst correlations. This changes when an indirect confirmation measure is used. While all segmentation methods reach a very high correlation, the single-word-based segmentations have slightly lower correlations than those that take larger subsets into account (S^{any}_{any}, S^{one}_{any}, S^{one}_{all}, S^{one}_{set}).

The arithmetic mean is the aggregation with the highest correlations. Combined with indirect confirmation measures the median and the quadratic mean created high correlations, too.

Among the direct confirmation measures, only m_f, m_{nlr}, m_{lr} and m_d reach higher correlations (see table 4). The last

Name	direct	indirect		
		$\tilde{m}_{cos}$	$\tilde{m}_{dice}$	$\tilde{m}_{jac}$
m_c	0.581	0.619	0.628	0.626
m_d	0.623	0.675	0.676	0.664
m_f	0.700	0.685	0.453	0.539
m_j	0.478	0.582	0.592	0.576
m_l	0.381	0.388	0.380	0.380
m_{lc}	0.493	0.374	0.238	0.238
m_{lj}	0.574	0.249	0.210	0.205
m_{ll}	0.582	0.563	0.472	0.507
m_{lr}	0.632	0.714	0.672	0.670
m_{ls}	0.172	0.714	0.672	0.670
m_{nlr}	0.693	0.731	0.689	0.691
m_o	0.478	0.582	0.592	0.576
m_P	—	0.575	0.605	0.590
m_r	0.378	0.605	0.577	0.557

Table 4: Best correlations for the confirmation measures used directly or combined with an inderect confirmation measure.

corpus left out	coherence $\boldsymbol{\mathcal{P}}$	$\mathcal{S}$	$\mathcal{M}$	Σ	average correlation
20NG	$(\mathcal{P}_{sw110}, S^{any}_{any}, \tilde{m}_{cos(nlr,1)}, \sigma_m)$				0.708
Genomics	$(\mathcal{P}_{sw70}, S^{one}_{set}, \tilde{m}_{cos(nlr,1)}, \sigma_a)$				0.724
NYT	$(\mathcal{P}_{sw110}, S^{one}_{set}, \tilde{m}_{cos(nlr,1)}, \sigma_a)$				0.717
RTL-NYT	$(\mathcal{P}_{sw110}, S^{one}_{set}, \tilde{m}_{cos(nlr,1)}, \sigma_a)$				0.732
RTL-Wiki	$(\mathcal{P}_{sw110}, S^{one}_{set}, \tilde{m}_{cos(nlr,1)}, \sigma_a)$				0.741
Movie	$(\mathcal{P}_{sw70}, S^{one}_{set}, \tilde{m}_{cos(nlr,1)}, \sigma_a)$				0.769

Table 5: Coherence measures with the highest average correlations if one dataset has been left out.

three benefit from a combination with an indirect measure, while the correlation of m_f drops.

The small differences in correlation of coherences with a) different window sizes and b) segmenation methods that are very similar to each other, leads to a large number of coherences having a high average correlation that is only slightly lower than the best performing coherence C_V. Thus, there are many variants of C_V that are performing very good as long as they use

- a sliding window with a window size $s \geq 50$,
- a segmentation method that takes larger subsets into account and
- σ_a or σ_m as summarization method.

We confirm this by generating *leave one out averages* by calculating the average using only five of the six datasets. Table 5 shows that independently from the dataset left out, C_V or one of its variants fullfilling the points above achieves the highest average correlation. For every coherence in table 2, figure 4 shows the six averages. It can be seen that the set of averages of C_V only overlap with the averages of C_P and are clearly higher than those of the other coherences.

6. RUNTIMES

Next to the quality it is important to know the costs of a certain coherence measure. Therefore, we have analyzed the runtimes of all coherence measures. For this experiment, we used the 100 topics of the NYT dataset and the Wikipedia

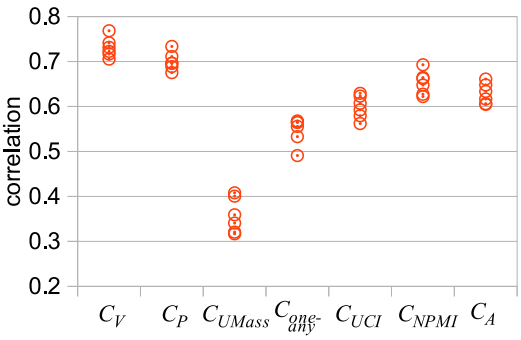

Figure 4: The leave one out averages of the coherence measures.

Name	runtime in s
C_V	315.2
C_P	317.7
C_{UMass}	13.7
$C_{\text{one-any}}$	13.7
C_{UCI}	356.1
C_{NPMI}	356.1
C_A	301.8

Table 6: Runtimes on the NYT dataset and the Wikipedia as reference corpus.

as reference corpus. Table 6 shows the overall runtimes of the coherence measures to compute the results presented in table 2.

For the runtime of a coherence measure, the most important component is the probability estimation. The fastest one is the boolean document that needed only 13.7s to retrieve all neccessary probabilities. The boolean paragraph and the boolean sentence based estimation methods needed 34.4s and 138.8s. Both suffer from the fact that there are much more paragraphs and sentences than single documents. But they are still faster than the window based approaches since the reference corpus can be divided into paragraphs or sentences while preprocessing the corpus. In contrast, both window based estimation methods had the highest runtimes. This is caused by the need of retrieving the single positions of the words inside the documents to check whether these positions are inside the same window. As parameters for the runtime of all probability calculation we can identify a) the number of topics, b) the number of top words per topic (N) and c) the size of the reference corpus. On the other hand, the size of the window does not have an influence on the runtimes.

Another important component is the segmentation. While the segmentation of a specific topic is very fast, it controls the number of confirmation values that have to be calculated. Thus, it has a high impact on the time needed by the confirmation measure and the aggregation component. Table 7 shows the number of subset pairs S_i that the different segmentations create and two examples of the influence of this number on the runtime of confirmation measures and aggregations. It can be seen that S^{one}_{any} and S^{any}_{any} have an exponential complexity.

7. APPLICATIONS

Coherence measures can be applied to automatically rate quality of topics computed by topic models. This can be

Name	$\lvert S\rvert$	runtimes in s	
		$m_{cos(m_{nlr},1)}$	σ_m
S^{one}_{all}	N	0.002	<0.001
S^{one}_{set}	N	0.002	—
S^{one}_{pre}	$\frac{N\cdot(N-1)}{2}$	0.002	<0.001
S^{one}_{suc}	$\frac{N\cdot(N-1)}{2}$	0.002	<0.001
S^{one}_{one}	$N\cdot(N-1)$	0.002	0.001
S^{one}_{any}	$N\cdot(2^{(N-1)}-1)$	0.140	0.023
S^{any}_{any}	$\sum_{i=1}^{N-1}\left(\binom{N}{i}\cdot(2^i-1)\right)$	1.436	0.322

Table 7: Different segmentation schemes, the number of subset pairs S_i they contain ($|S|$) and examples of their influence on runtimes of confirmation measures and aggregations.

used in data exploration systems like TopicExplorer[11] [9] that use word representations of topics to show the user an overview of large document collections. Topics that cannot be interpreted are filtered out using coherence measures. Thus, usability of the system is increased by hiding topics that would confuse users.

As example table 8 shows the five topics with highest and lowest coherence respectively of a total of 200 topics that are computed by Latent Dirichlet Allocation [3] using English Wikipedia. It clearly shows the difference between the quality of the topic representations as best topics allow to guess a general topic while the worst topics are not interpretable.

However, coherence measures have applications beyond topic models. First, precision of key word search in document collections can profit by filtering out documents that contain the query key words in some proximity but the words are located in nearby but thematically different passages. The approach would extract a word set for each document that covers the query words as well as the main words of the documents passage that contains the query words. Low quality hits for the query are filtered out by setting a minimum threshold for the coherence of the extracted word set.

Second, coherence can be used to select advertising links that fits a web page. The approach is similar as above. A word set is extracted from a web page that contains the most describing words of the pages content. Advertising links are selected that maximize coherence of the union of the web page's word set with descriptive words of the respective ad.

Third, coherence can be used to improve automatic translations of web pages. When using a dictionary to translate a sentence from a web page into another language, usually ambiguous translations of single words remain. Coherence measures can be used select the most coherent combination among the ambiguous translations.

8. CONCLUSION

We proposed a unifying framework that span the space of all known coherence measures and allows to combine all main ideas in the context of coherence quantification. We evaluated 237 912 coherence measures on six different benchmarks for topic coherence—to the best of our knowledge, this is the largest number of benchmarks used for topic coherences so far. We introduced coherence measures from

[11] `http://topicexplorer.informatik.uni-halle.de`

coherence	topic
0.94	company sell corporation own acquire purchase buy business sale owner
0.91	age population household female family census live average median income
0.86	jewish israel jew israeli jerusalem rabbi hebrew palestinian palestine holocaust
0.85	ship navy naval fleet sail vessel crew sink sea submarine
0.83	guitar album track bass drum vocal vocals release personnel list
...	
0.35	number code type section key table block set example oth
0.34	group official website member site base oth form consist profile
0.34	part form become history change present create merge forme separate
0.29	know call name several refer oth hunter hunt thompson include
0.27	mark paul heart take read harrison follow become know include

Table 8: The five Wikipedia topics with highest and lowest coherences using best preforming measure C_V.

scientific philosophy that have been not used in the area of topic evaluation and—in the case of Fitelson's coherence—showed a good performance. Using our framework, we identified measures that clearly outperform all coherences proposed so far.

In the future, a next step will be a more detailed investigation on the sliding window and could bring up new probability estimation methods. Future work will include transfer of coherence to other applications beyond topic models, which can be done in a systematic way using the new unifying framework. Hence, our research agenda may lead to completely new insights of the behavior of topic models. Finally, our open framework enables researchers to measure the impact on a comparable basis for fields of application like text mining, information retrieval and the world wide web. Additionally, it will be possible to finally evaluate the reasons for specific behavior of statistics-driven models and uncover distinguishing properties of benchmarks.

9. ACKNOWLEDGMENTS

We want to thank Jey Han Lau and Jordan Boyd-Graber for making their data available. A special thanks goes to Nikos Aletras for publishing his data and several discussions about his approach. The work of the first author has been supported by the ESF and the Free State of Saxony. The work of the last author has been partially supported by the Klaus Tschira Foundation.

10. REFERENCES

[1] N. Aletras and M. Stevenson. Evaluating topic coherence using distributional semantics. In *Proc. of the 10th Int. Conf. on Computational Semantics (IWCS'13)*, pages 13–22, 2013.

[2] L. Alsumait, D. Barbará, J. Gentle, and C. Domeniconi. Topic significance ranking of lda generative models. In *Proc. of the Europ. Conf. on Machine Learning and Knowledge Disc. in Databases: Part I*, ECML PKDD '09, pages 67–82, 2009.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proc. of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen, 2009.

[5] L. Bovens and S. Hartmann. *Bayesian Epistemology*. Oxford University Press, 2003.

[6] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pages 288–296. 2009.

[7] I. Douven and W. Meijs. Measuring coherence. *Synthese*, 156(3):405–425, 2007.

[8] B. Fitelson. A probabilistic theory of coherence. *Analysis*, 63(279):194–199, 2003.

[9] A. Hinneburg, R. Preiss, and R. Schröder. Topicexplorer: Exploring document collections with topic models. In *ECML/PKDD (2)*, pages 838–841, 2012.

[10] J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proc. of the Europ. Chap. of the Assoc. for Comp. Ling.*, 2014.

[11] D. Mimno and D. Blei. Bayesian checking for topic models. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 227–237. 2011.

[12] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 262–272. 2011.

[13] 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*, pages 100–108. 2010.

[14] E. Olsson. What is the problem of coherence and truth? *The Jour. of Philosophy*, 99(5):246–272, 2002.

[15] M. Röder, M. Speicher, and R. Usbeck. Investigating quality raters' performance using interface evaluation methods. In *Informatik 2013*, pages 137–139. GI, 2013.

[16] F. Rosner, A. Hinneburg, M. Röder, M. Nettling, and A. Both. Evaluating topic coherence measures. *CoRR*, abs/1403.6397, 2014.

[17] T. Shogenji. Is coherence truth conducive? *Analysis*, 59(264):338–345, 1999.

[18] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. Exploring topic coherence over many models and many topics. In *Proc. of the 2012 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 952–961, 2012.