

Intra-Text Coherence as a Measure of Topic Models' Interpretability

Vasiliy Alekseev, Victor Bulatov, Konstantin Vorontsov

24rd International Conference on Computational Linguistics and
Intellectual Technologies

1 June 2018



Topic, Its Interpretability & Coherence

Topic is a set of words that often occur together in text.

Interpretability of the topic means that a human is able to explain the meaning behind its set of words. However, such human assessment is expensive.

Well Interpreted Topic (Most Frequent Terms)

actor, play, musical, premiere, parterre, spectator, producer, audience, backstage, orchestra

Badly Interpreted Topic (Most Frequent Terms)

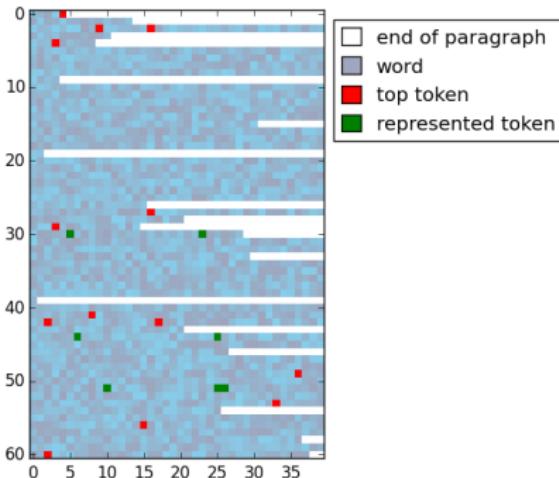
express, epigraph, foggy, result, image, right, loan, debt, bankrupt, interest

Coherence is a commonly used method for estimating the interpretability, which measures how often 10 most probable terms of the topic occur in close proximity within text.

Purpose of the Study

Problem

Existing methods of calculating topic's coherence are based on the analysis of its most frequent words' co-occurrences. However, the proportion of text covered by these top-words is not controlled in any way.



Solution

Evaluate coherence as an average thematic proximity of words closely located in text.

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

Table of Contents

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

Conventional Signs

- W – dictionary, D – documents, T – topics
- n_{dw} – number of occurrences of word w in the document d
- ν_{wd} – frequency of word's w occurrences in document d
- $\varphi_{wt} \equiv p(w | t)$ – probability that word w refers to topic t
- $\theta_{td} \equiv p(t | d)$ – probability that topic t refers to document d
- $w \equiv p(t | w)$ – corresponding to word w vector
- Matrices $\Phi \equiv (\varphi_{wt})_{W \times T}$, $\Theta \equiv (\theta_{td})_{T \times D}$
- "Bag of words" hypothesis: it doesn't matter how words are ordered in documents
- Hypothesis of conditional independence:
 $p(w | d, t) \equiv p(w | t)$

Topic model through topics describes occurrences of words in documents

$$p(w | d) = \sum_T p(w | t)p(t | d)$$

Matrix decomposition problem $(\nu_{wd})_{W \times D} = \Phi \Theta$

ARTM¹: Regularized log likelihood $p(T, D, W)$ maximization

$$\sum_{d \in D} \sum_{w \in W_d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

Proper regularizer $R(\Phi, \Theta)$ may help with

- smoothing a background topic, spreading its probability among general vocabulary words
- sparsing a topic, thickening its probability in a small number of subject area words
- decorrelation of topics

¹Vorontsov K. et al. Bigartm: Open source library for regularized multimodal topic modeling of large collections, 2015

Original Dataset

Approximately 2000 *monotopic* articles from PostNauka² on 19 topics.

Topic	First Top-Word	Second Top-Word	Third Top-Word
1: математика	математика (0.016)	задача (0.008)	декарт (0.008)
2: технологии	технология (0.015)	робот (0.012)	сеть (0.010)
3: физика	частица (0.027)	электрон (0.015)	кварт (0.015)
4: химия	химия (0.021)	молекула (0.019)	материал (0.016)
5: земля	земля (0.029)	планета (0.028)	атмосфера (0.012)
6: астрономия	звезда (0.039)	галактика (0.031)	вселенная (0.019)
7: биология	клетка (0.027)	организм (0.011)	мозг (0.010)
8: медицина	пациент (0.016)	препарат (0.012)	заболевание (0.012)
9: психология	психология (0.009)	мозг (0.009)	психолог (0.008)
10: экономика	экономика (0.016)	страна (0.010)	цена (0.008)
11: история	история (0.010)	историк (0.007)	власть (0.006)
12: политика	государство (0.014)	политика (0.012)	политический (0.011)
13: социология	социология (0.013)	социолог (0.009)	социальный (0.008)
14: культура	культура (0.015)	фильм (0.007)	искусство (0.006)
15: образование	университет (0.021)	образование (0.014)	школа (0.013)
16: язык	язык (0.077)	слово (0.037)	словарь (0.011)
17: философия	философия (0.018)	философ (0.013)	философский (0.008)
18: религия	святилище (0.010)	религия (0.007)	царь (0.006)
19: россия	россия (0.028)	страна (0.009)	русский (0.009)

Three top-words of each of 19 topics

²<https://postnauka.ru>

Table of Contents

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

- Newman³ $|_t = \frac{1}{\binom{k}{2}} \sum_{i=1}^{k-1} \sum_{j=i+1}^k \ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$

- Mimno⁴ $|_t = \frac{1}{\binom{k}{2}} \sum_{i=1}^{k-1} \sum_{j=i+1}^k \ln \frac{D(w_i, w_j) + 1}{D(w_i)}$

- k – number of top-words of topic t used to evaluate coherence
- $p(w_i)$, $p(w_i, w_j)$ – probability to find word w_i and two words w_i, w_j in a context window of given size
- $D(w_i)$, $D(w_i, w_j)$ – number of documents containing word w_i and two words w_i, w_j in a context window of given size

³Newman et al. Automatic Evaluation of Topic Coherence, 2010

⁴Mimno et al. Optimizing Semantic Coherence in Topic Models, 2011

Drawbacks of Top-Tokens Based Approach

Top token-based coherences may ignore more than 98% of words of the documents' collection.

	PostNauka, %	Wikipedia, %
Minimum	0.016	0.0065
Median	0.048	0.029
Mean	0.062	0.036
Maximum	0.28	0.11
Total	1.2	1.7

The proportion of corpus contributing to the co-occurrence counts of top 10 most frequent words for each topic

Drawbacks of Top-Tokens Based Approach

A single top token "частиц" out of the first 10 ones is seen.
The wide range of less strong topical words is ignored by the top-tokens based coherences.

Напротив, если предположить существование суперсимметрии, то введение новых **частиц** приводит как раз к такому объединению. Оказывается, что суперсимметрия не только обеспечивает объединение взаимодействий, но и стабилизирует объединённую теорию, в которой присутствуют два совершенно разных масштаба: масштаб масс обычных **частиц** (порядка 100 масс протона) и масштаб великого объединения (порядка 10^{16} масс протона). Последний масштаб уже близок к так называемому планковскому масштабу, равному обратной ньютоновской константе тяготения, что составляет порядка 10^{19} масс протона. На этом масштабе мы ожидаем проявление эффектов квантовой гравитации. В этом моменте нас ожидает приятный сюрприз. Дело в том, что гравитация всегда стояла несколько особняком по отношению к остальным взаимодействиям. Переносчик гравитации, гравитон, имеет спин 2, в то время как переносчики остальных взаимодействий имеют спин 1. Однако суперсимметрия перемешивает спины.

first top words of topic 3: физика with top 10 in bold: **частица, электрон, кварк, атом, энергия, вселенная, фотон, физика, физик, эксперимент**, масса, теория, свет, симметрия, протон, эйнштейн, нейтрино, вещество, квантовый, ускоритель, детектор, волна, эффект, свойство, спин, гравитация, материя, адрон, поль, частота

Table of Contents

0 Prologue

- Topic Modeling
- Original Dataset

1 Top-Tokens Based Coherences

- Newman, Mimno
- Drawbacks of Top-Tokens Based Approach

2 Intra-Text Coherences

3 Automatic Coherences' Quality Estimation

- Semisynthetic Dataset
- Segmentation Quality

4 Experiments

Meaning

Semantic proximity of closely located words in text

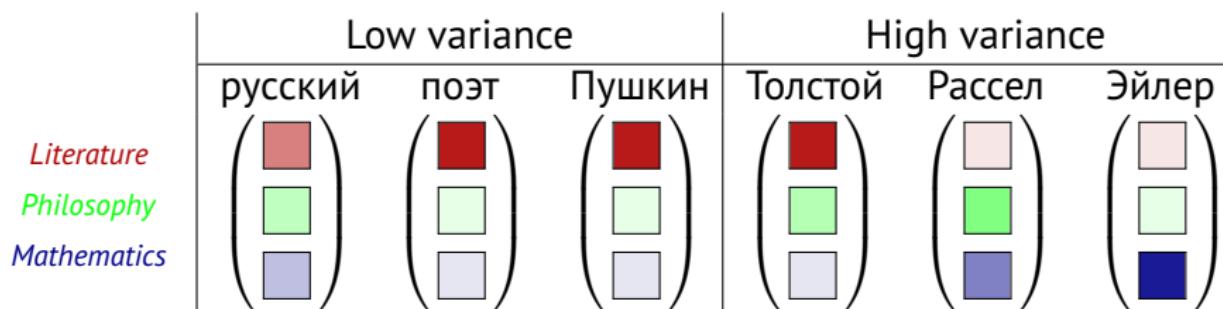
	Группа $\begin{pmatrix} \text{pink} \\ \text{green} \\ \text{blue} \end{pmatrix}$	астрономов $\begin{pmatrix} \text{red} \\ \text{light green} \\ \text{light blue} \end{pmatrix}$	обнаружила $\begin{pmatrix} \text{red} \\ \text{light green} \\ \text{light blue} \end{pmatrix}$	звезды $\begin{pmatrix} \text{red} \\ \text{light green} \\ \text{blue} \end{pmatrix}$
Astronomy				
Biology				
Music				

We compare every pair of vectors: e.g.

$$\begin{pmatrix} \text{pink} \\ \text{green} \\ \text{blue} \end{pmatrix} \text{ vs } \begin{pmatrix} \text{red} \\ \text{light green} \\ \text{light blue} \end{pmatrix}, \text{ etc.}$$

Meaning

How much the meanings of adjacent words differ, according to the topic model



Meaning

Average length of the topic in text

Formula

$$\text{TopLen} = \left\langle \max \left\{ n : \text{thr} + \sum_{j=i}^{i+n} \left(w_j[t] - \max_{\substack{1 \leq \tau \leq |T| \\ \tau \neq t}} w_j[\tau] \right) \geq 0 \right\} \right\rangle$$

t = "Чёрные дыры, thr ~ 0 – threshold

Группе $\underbrace{\text{астрономов}}_{w_1}$ удалось обнаружить $\underbrace{\text{звезды}}_{l_1=2}$, обращающуюся $\underbrace{\text{вокруг}}_{l_3=4}$ $\underbrace{\text{чёрной дыры}}_{l_2=2}$ на рекордно близком расстоянии.

Meaning

Estimation of how much the focus of a conversation drifts

Formula

$$\text{FoCon} = - \sum_{d \in D} \sum_{\substack{w, u \in W_d \\ \rho(w, u) = 1}} |\mathbf{w}[t] - \mathbf{u}[t]| + |\mathbf{w}[\tau] - \mathbf{u}[\tau]|$$

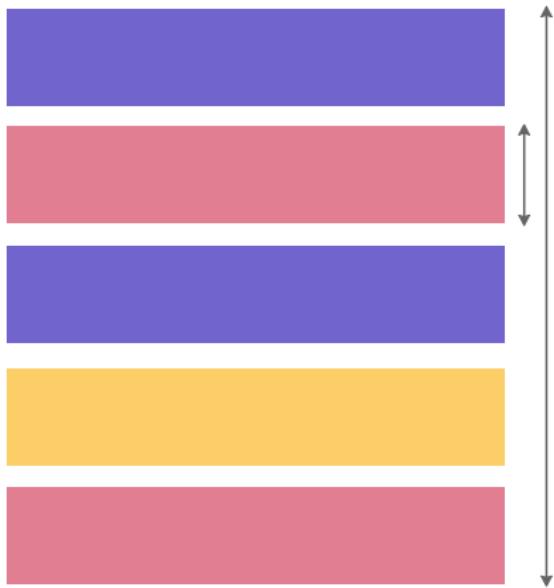
t, τ – maximal components of vectors \mathbf{w} and \mathbf{u} respectively

Table of Contents

- 0 Prologue
 - Topic Modeling
 - Original Dataset
- 1 Top-Tokens Based Coherences
 - Newman, Mimno
 - Drawbacks of Top-Tokens Based Approach
- 2 Intra-Text Coherences
- 3 Automatic Coherences' Quality Estimation
 - Semisynthetic Dataset
 - Segmentation Quality
- 4 Experiments

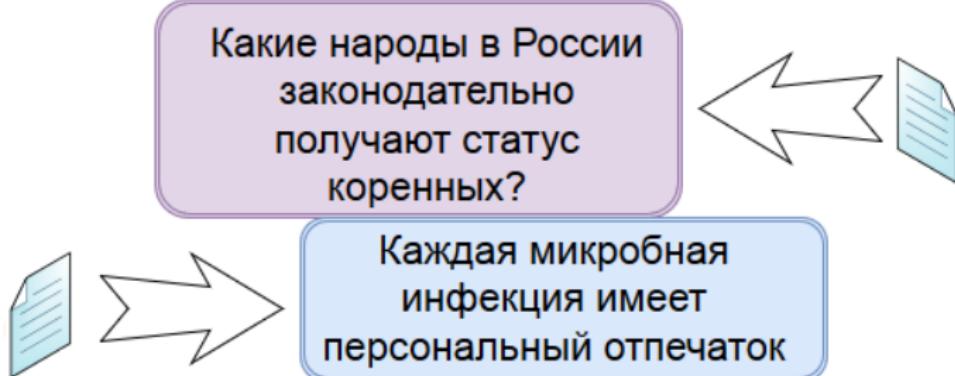
Dataset parameters

- document size
- thm – number of topics in a document
- sgm – number of words in a segment



The better the coherence is, the better it should describe the ability of a topic model to figure out the segmentation structure!

Example



Semi-synthetic document produced by combining a sociology segment and an medicine segment.

Soft

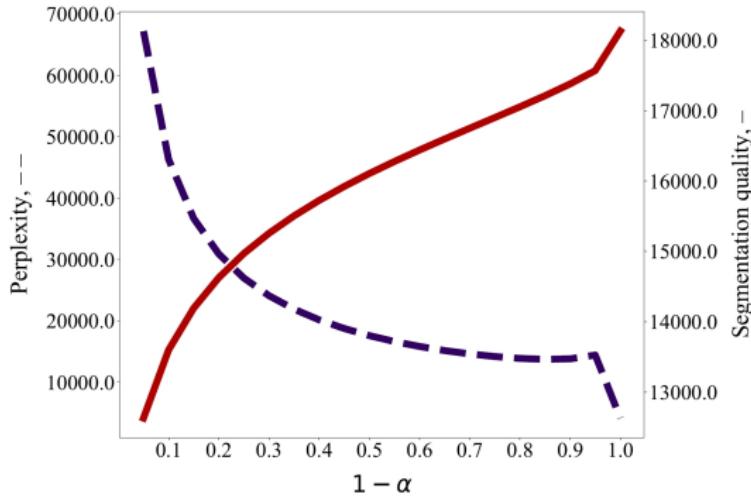
Sum among topics of sums $p(t | d, w)$ for each topic t over pairs $(d, w), d \in D, w \in W_d$

Strict

Number of coincidences of the topic $\arg \max_{\tau} p(\tau | d, w)$ predicted by the model for the word w in the document with the actual topic t of the segment

Segmentation Quality & Perplexity of Topic Model

- Perplexity: intrinsic quality criteria; the lower, the better.
- Range of topic models: $\Phi(\alpha) = \alpha\Phi_{bad} + (1 - \alpha)\Phi_{good}$.



Proposed segmentation quality estimation may be used as topic models' quality measure

Table of Contents

- 0 Prologue
 - Topic Modeling
 - Original Dataset
- 1 Top-Tokens Based Coherences
 - Newman, Mimno
 - Drawbacks of Top-Tokens Based Approach
- 2 Intra-Text Coherences
- 3 Automatic Coherences' Quality Estimation
 - Semisynthetic Dataset
 - Segmentation Quality
- 4 Experiments

Illustration of a Bad Model Segmenting Text

topic 16: язык

Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить "я это сделаю" или "это будет" а сказать "это возможно" или "я хочу это сделать" они говорят о будущем, но "попадают" в будущее непрямым путём.

topic 12: политика

И я посылаю деньги борцам за независимость Курдистана, участвуя в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных "гражданств". В литературе последних десяти лет бытуют такие выражения, как "гендерное гражданство" "экономическое гражданство". Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC L2	SC Cos	SC Var	TL	FC
5.54e3	1.10e4	-4.83	-3.12	-12.9	0.947	-37.0e3	2.87	-13.9e4
16.0e3	3.76e4	-3.65	-2.69	-3.70	0.700	-8.12e3	3.45	-5.44e4

- SQ (S), SQ (H) – Soft and Strict segmentation qualities
- N, M – Newman, Mimno
- SC, TL, FC – SemantiC, TopLen, FoCon

Illustration of the Good Model Segmenting Text

topic 16: язык

Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить "я это сделаю" или "это будет" а сказать "это возможно" или "я хочу это сделать" они говорят о будущем, но "попадают" в будущее непрямым путём.

topic 12: политика

И я посылаю деньги борцам за независимость Курдистана, участвую в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных "гражданств". В литературе последних десяти лет бытуют такие выражения, как "гендерное гражданство" "экономическое гражданство". Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC L2	SC Cos	SC Var	TL	FC
5.54e3	1.10e4	-4.83	-3.12	-12.9	0.947	-37.0e3	2.87	-13.9e4
16.0e3	3.76e4	-3.65	-2.69	-3.70	0.700	-8.12e3	3.45	-5.44e4

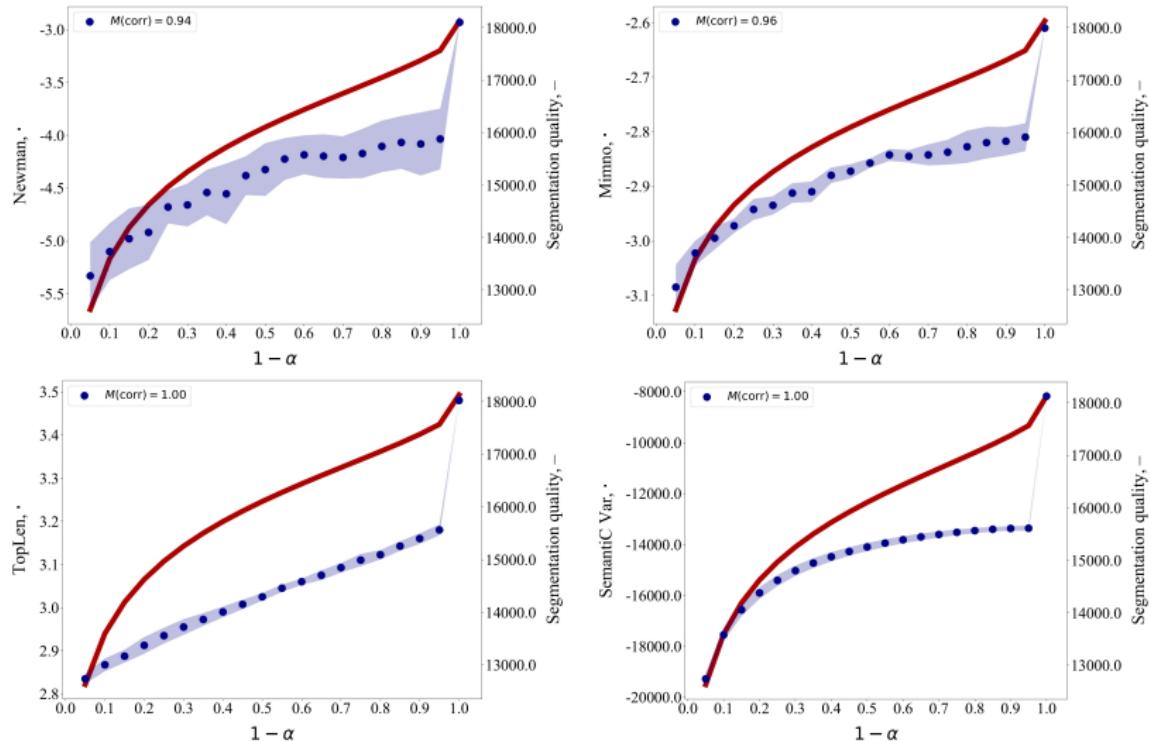
- SQ (S), SQ (H) – Soft and Strict segmentation qualities
- N, M – Newman, Mimmo
- SC, TL, FC – SemantiC, TopLen, FoCon

Spearman Correlations Between Coherences & Segmentation Quality

Coh	Corr	Coh	Corr	Coh	Corr
Newman	0.75	Newman	0.80	Newman	0.85
Mimno	0.96	Mimno	0.94	Mimno	0.97
SC L2	0.92	SC L2	0.70	SC L2	0.59
SC Cos	-0.97	SC Cos	-0.97	SC Cos	-0.96
SC Var	1.00	SC Var	1.00	SC Var	1.00
TopLen	1.00	TopLen	1.00	TopLen	1.00
FoCon	1.00	FoCon	1.00	FoCon	1.00

Results for datasets with sizes of segments: 50, 200 and 400 words – and with 5 topics in each document

Coherence Measures & Segmentation Quality as a Function of α (dataset sgm = 200, thm = 5)



- New methods of calculating coherence which take into account the whole text.
In the problem under consideration, the proposed coherences outperform top-tokens based ones.
- An automatic method for evaluating coherences functions based on their comparison with text segmentation quality.