Семинар 3: немного про эмбеддинги, Neural Nets + Topic Modeling, и про интерпретируемость моделей

Алексеев Василий

МФТИ, группа 874

23 июня 2020

Содержание

Про что?

Задача

Представить слова двух языков в одном векторном пространстве.

Постановка задачи

- $X \in \mathbb{R}^{n \times d}$ — эмбеддинги n слов в исходном языке
- $Z \in \mathbb{R}^{n \times d}$ — эмбеддинги n слов в целевом языке
- $W\colon\ \mathbb{R}^d \to \mathbb{R}^d$ — цель

Pratik Jawanpuria et al. *Geometry-aware Domain Adaptation for Unsupervised Alignment of Word Embeddings* //arXiv preprint arXiv:2004.08243. – 2020.

Обучение с учителем

$$\begin{cases} \|XW - YZ\|^2_{Fro} \to min_{W \in \mathbb{R}^{d \times d}} \\ W^T W = 1 \end{cases}$$

где $Y \in \{0, 1\}^n$ — матрица соответствия (alignment matrix): элемент Y_{ij} равен 1, если j-ое слово в целевом языке есть перевод i-го слова из исходного языка, и 0 в противном случае.

Если же словарь соответствия Y — тоже *не известен*, то

- можно обучать W и Y совместно
- можно сначала найти и зафиксировать Y, потом искать W. Например, Y можно искать[1] как *дважды стохастическую матрицу из* $\mathbb{DS}^n$

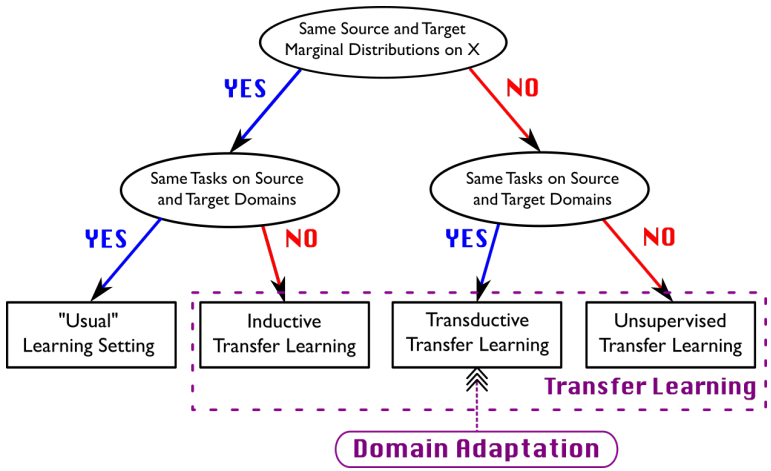
$$-\mathrm{Sp}(Y^T \cdot XX^T \cdot Y \cdot ZZ^T) \to min_{Y \in \mathbb{DS}^n}$$

[1] David Alvarez-Melis D., and Tommi S. Jaakkola *Gromov-wasserstein alignment of word embedding spaces* //arXiv preprint arXiv:1809.00013. – 2018.

Суть работы

- Учёт информации второго порядка о связи слов в языках.
- Если Y делает выравнивание между языками X и Z, то Y^T должна выравнивать Z и X.
- Alignment learning problem $\rightarrow$ Domain adaptation problem[1]

$$\|Y^T \cdot XX^T \cdot Y - ZZ^T\|^2 + \|Y \cdot ZZ^T \cdot Y^T - XX^T\|^2 \rightarrow \min_{Y \in \mathbb{DS}^n}$$



https://en.wikipedia.org/wiki/Domain_adaptation

[1] Baochen Sun et al. *Return of frustratingly easy domain adaptation* //Thirtieth AAAI Conference on Artificial Intelligence. – 2016.

Результаты

Получение оценок:

- Bilingual induction (BLI) task
- Cross-Domain Similarity Local Scaling (CSLS)
- Precision@1

Method	de-xx	en-xx	es-xx	fr-xx	it-xx	pt-xx	xx-de	xx-en	xx-es	xx-fr	xx-it	xx-pt	avg.
GW	62.6	77.4	**78.2**	**75.4**	**77.5**	77.2	62.6	75.9	**79.7**	**79.0**	**76.2**	74.9	74.7
MBA	**63.3**	**78.4**	**78.2**	75.3	77.0	**77.5**	**63.1**	**77.3**	79.4	78.7	**76.2**	**75.0**	**75.0**

Table 1: P@1 for BLI on six European languages: English, German, Spanish, French, Italian, and Portuguese. Here 'en-xx' refers to the average P@1 when English is the source language and others are target language. Similarly, 'xx-en' implies English as the target language and others as source language. Thus, 'avg.' shows P@1 averaged over all the thirty BLI results for each algorithm. The proposed algorithm MBA performs similar when the language pairs are closely related to each other.

Method	en-bg	en-cs	en-da	en-el	en-fi	en-hu	en-nl	en-pl	en-ru	
GW	22.8	42.1	54.4	21.5	37.7	43.7	72.9	49.1	36.1	
MBA	**38.1**	**46.8**	**56.1**	**40.0**	**40.4**	**46.1**	**73.8**	**50.4**	**37.5**	
Method	**bg-en**	**cs-en**	**da-en**	**el-en**	**fi-en**	**hu-en**	**nl-en**	**pl-en**	**ru-en**	**avg.**
GW	29.9	52.9	60.7	32.7	49.5	57.6	70.9	57.7	48.3	47.0
MBA	**50.0**	**57.7**	**62.3**	**54.4**	**54.4**	**61.0**	**71.0**	**60.5**	**54.1**	**53.0**

Table 2: P@1 for BLI on English and nine European languages: Bulgarian, Czech, Danish, Greek, Finnish, Hungarian, Dutch, Polish, and Russian. The 'avg.' shows P@1 averaged over all the eighteen BLI results. The proposed algorithm MBA outperforms GW when the bilingual mapping is learned between distant languages.

Для пар разных языков качество лучше, чем у baseline-решений.

Почему эта статья?

- Понравилась задача :)
- Пример того, как можно видоизменять оптимизируемый функционал, чтобы можно было получить численное решение.

Содержание

Задача

Постановка задачи

$\mathcal{F}_1, \ldots, \mathcal{F}_J$ — отображения из множества предложений $\mathbb{S}$ в множества векторов размерностей $d_1, \ldots, d_J$. Применить способы получения мета-эмбеддингов[1] слов для получения мета-эмбеддингов предложений $\mathcal{F} : s \mapsto \mathbb{R}^d$.

Мотивация

- Мета-эмбеддинги предложений могут показывать лучшее качество, чем отдельные эмбеддинги[2]
- Обычные эмбеддинги зависят от архитектуры, данных, задачи.

Nina Poerner et al. *Sentence Meta-Embeddings for Unsupervised Semantic Textual Similarity* //arXiv preprint arXiv:1911.03700. – 2019.

[1]Мета-эмбеддинги — объединения предобученных разными способами векторных представлений

[2]Wenpeng Yin et al. *Learning word meta-embeddings*, 2016.

Суть работы

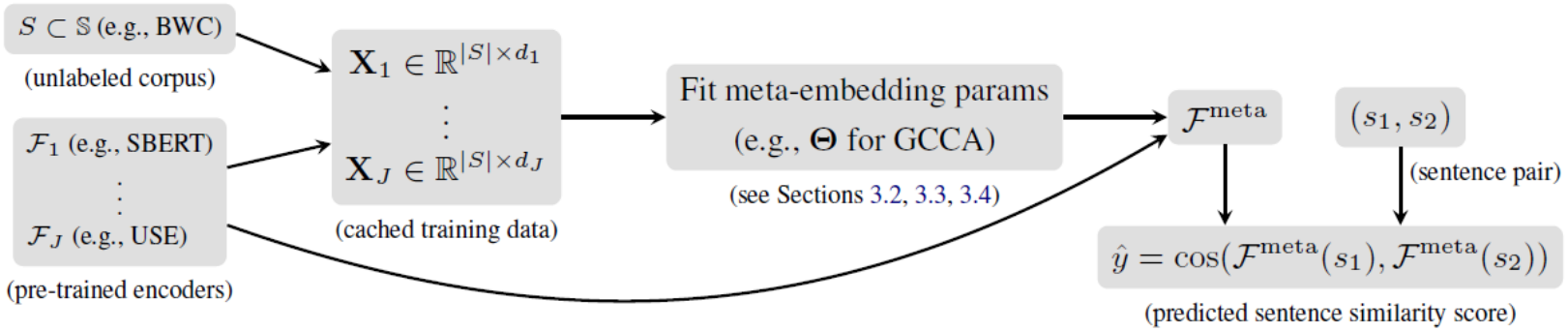


Figure 1: Schematic depiction: Trainable sentence meta-embeddings for unsupervised STS.

Разные способы объединения эмбеддингов:

- Конкатенация
- Конкатенация и SVD
- Линейная проекция
- Generalized canonical correlation analysis (GCCA)
- Автоенкодеры[1]

[1] Для каждого F_j обучать энкодер $\mathcal{E}_j : R^{d_j} \to R^d$ и декодер $\mathcal{D}_j : R^d \to R^{d_j}$. И тогда $\mathcal{F}^{ae}(s) = \sum_j \mathcal{E}_j(\mathcal{F}_j(s))$.

Результаты

Описание эксперимента:

- Semantic Textual Similarity (STS) task: множество троек $\{(s_{i1}, s_{i2}, y_i)\}$, состоящих из пары предложений и их оценки сходства $y_i \in \mathbb{R}$.
- Корреляции Спирмена и Пирсона между разметкой y и получаемыми оценками $\cos(\mathcal{F}(s_1), \mathcal{F}(s_2))$.

		dimensionality	STS12	STS13	STS14	STS15	STS16	STS-B
single:ParaNMT		$d = 600$	67.5/66.3	62.7/62.8	77.3/74.9	80.3/80.8	78.3/79.1	79.8/78.9
single:USE		$d = 512$	62.6/63.8	57.3/57.8	69.5/66.0	74.8/77.1	73.7/76.4	76.2/74.6
single:SBERT		$d = 1024$	66.9/66.8	63.2/64.8	74.2/74.3	77.3/78.3	72.8/75.7	76.2/79.2
single:ParaNMT – up-projection*		$d = 1024$	67.3/66.2	62.1/62.4	77.1/74.7	79.7/80.2	77.9/78.7	79.5/78.6
single:USE – up-projection*		$d = 1024$	62.4/63.7	57.0/57.5	69.4/65.9	74.7/77.1	73.6/76.3	76.0/74.5
meta:conc		$d = 2136$	72.7/71.3	68.4/68.6	81.0/79.0	84.1/**85.5**	**82.0/83.8**	82.8/83.4
meta:avg		$d = 1024$	72.5/71.2	68.1/68.3	80.8/78.8	83.7/85.1	81.9/83.6	82.5/83.2
meta:svd		$d = 1024$	71.9/70.8	68.3/68.3	80.6/78.6	83.8/85.1	81.6/83.6	83.4/83.8
meta:gcca	(hyperparams on dev set)	$d = 1024$	**72.8/71.6**	**69.6/69.4**	**81.7/79.5**	**84.2/85.5**	81.3/83.3	**83.9/84.4**
meta:ae	(hyperparams on dev set)	$d = 1024$	71.5/70.6	68.5/68.4	80.1/78.5	82.5/83.1	80.4/81.9	82.1/83.3
Ethayarajh (2018)		(unsupervised)	68.3/-	66.1/-	78.4/-	79.0/-	-/-	79.5/-
Wieting and Gimpel (2018)		(unsupervised)	68.0/-	62.8/-	77.5/-	80.3/-	78.3/-	79.9/-
Tang and de Sa (2019)		(unsupervised meta)	64.0/-	61.7/-	73.7/-	77.2/-	76.7/-	-
Hassan et al. (2019)†		(unsupervised meta)	67.7/-	64.6/-	75.6/-	80.3/-	79.3/-	77.7/-
Poerner and Schütze (2019)		(unsupervised meta)	-/-	-/-	-/-	-/-	-/-	80.4/-
Reimers and Gurevych (2019)		(sup. siamese SoTA)	-/-	-/-	-/-	-/-	-/-	-/86.2
Raffel et al. (2019)		(supervised SoTA)	-/-	-/-	-/-	-/-	-/-	93.1/92.8

Table 2: Results on STS12–16 and STS Benchmark test set. STS12–16: mean Pearson's $r \times 100$ / Spearman's $\rho \times 100$. STS Benchmark: overall Pearson's $r \times 100$ / Spearman's $\rho \times 100$. Evaluated by SentEval (Conneau and Kiela, 2018). **Boldface:** best in column (except supervised). Underlined: best single-source method. *Results for up-projections are averaged over 10 random seeds. †Unweighted average computed from Hassan et al. (2019, Table 8). There is no supervised SoTA on STS12–16, as they are unsupervised benchmarks.

Почему эта статья?

- Мысль: любые эмбеддинги — не эталон, они зависят от данных и задачи.
- Эмбеддинги документов?

Содержание

Про что?

Задача

Предложить нейросетевую тематическую модель.

Мотивация

Существующие NN модели

- либо используют неподходящие априорные распределения на темах (например, нормальное распределение вместо Дирихле).
- либо не способны предсказывать темы для данного документа (!!)

Rui Wang et al. *Neural Topic Modeling with Bidirectional Adversarial Training* //arXiv preprint arXiv:2004.12331. – 2020.

Суть работы 1/2 (BAT)

Bidirectional Adversarial Topic (BAT) model: получение отображений между распределениями: тем в документах и слов в документах.

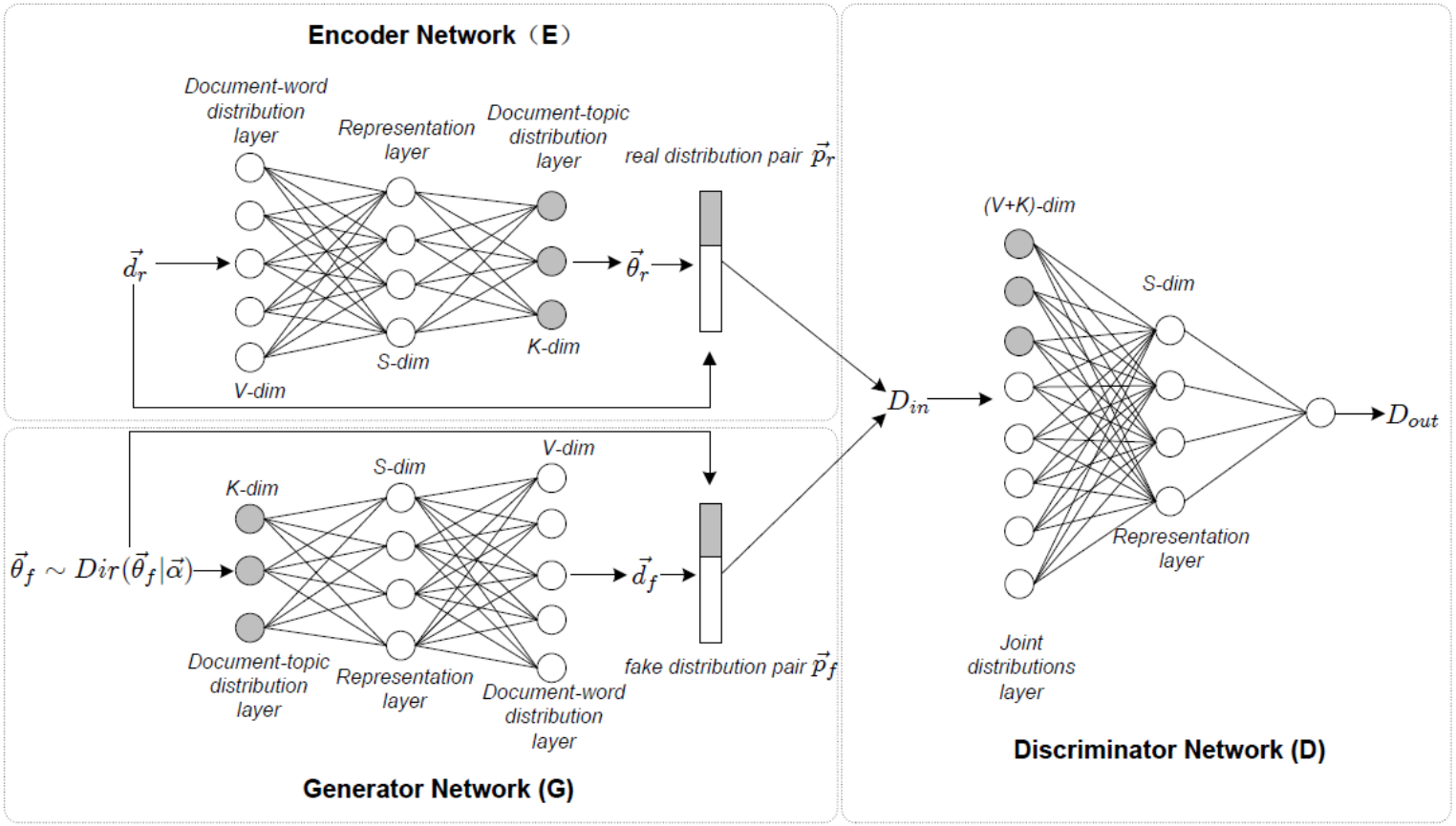
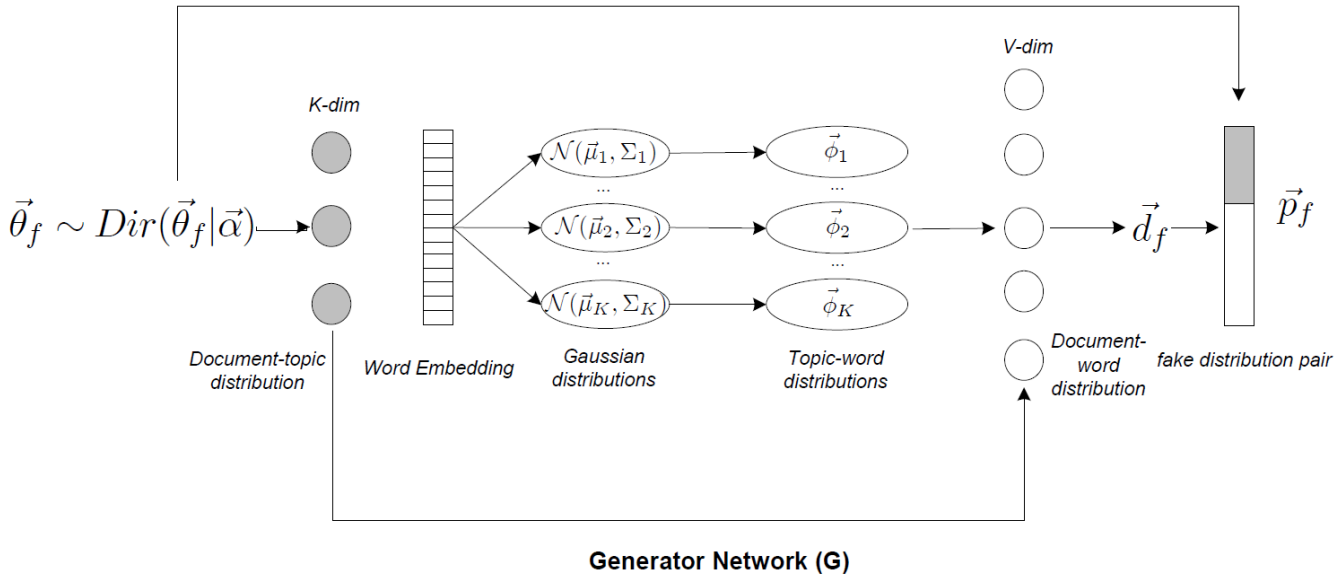


Figure 2: The framework of the Bidirectional Adversarial Topic (BAT) model.

Суть работы 2/2 (GBAT)

Bidirectional Adversarial Topic model with Gaussian (Gaussian-BAT): использование эмбеддингов.



$$p(\vec{e}_v \mid t) = \mathcal{N}(\vec{e}_v; \vec{\mu}_k, \Sigma_k) = \frac{\exp\left(-\frac{1}{2}(\vec{e}_v - \vec{\mu}_k)^T \Sigma_k^{-1}(\vec{e}_v - \vec{\mu}_k)\right)}{\sqrt{(2\pi)^{D_e}|\Sigma_k|}}$$

$$\varphi_{k,v} = \frac{p(\vec{e}_v \mid t)}{\sum_{u=1}^{V} p(\vec{e}_u \mid t)} \qquad \vec{d}_f = \sum_{k=1}^{K} \vec{\varphi}_k \cdot \theta_k$$

Результаты 1/3: Когерентности в зависимости от учитываемой части тем

- Тренируются модели на 20, 30, 50, 75, 100 тем.
- Когерентности считаются по топ 50, 70, 90, 100% тем.
- Результаты усредняются по моделям с разным числом тем.

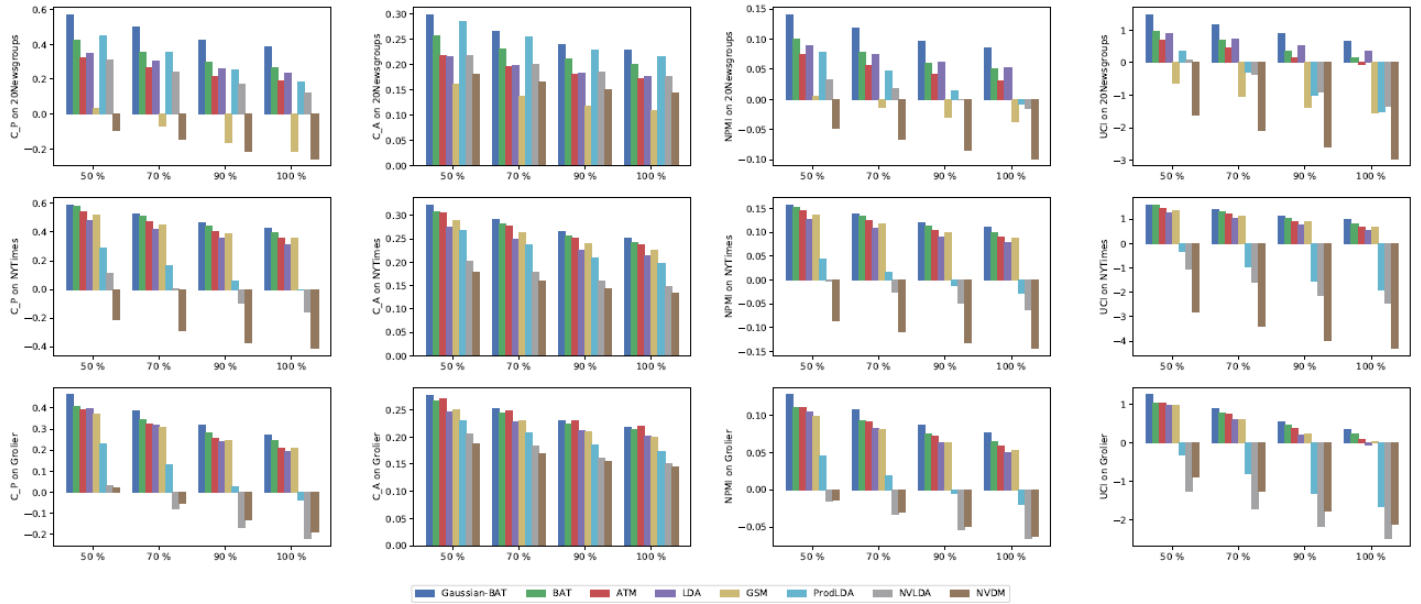


Figure 4: The comparison of average topic coherence vs. different topic proportion on three datasets.

Michael Röder et al. *Exploring the space of topic coherence measures* //Proceedings of the eighth ACM international conference on Web search and data mining. – 2015. – С. 399-408.

Результаты 2/3: Когерентности в зависимости от числа тем в модели

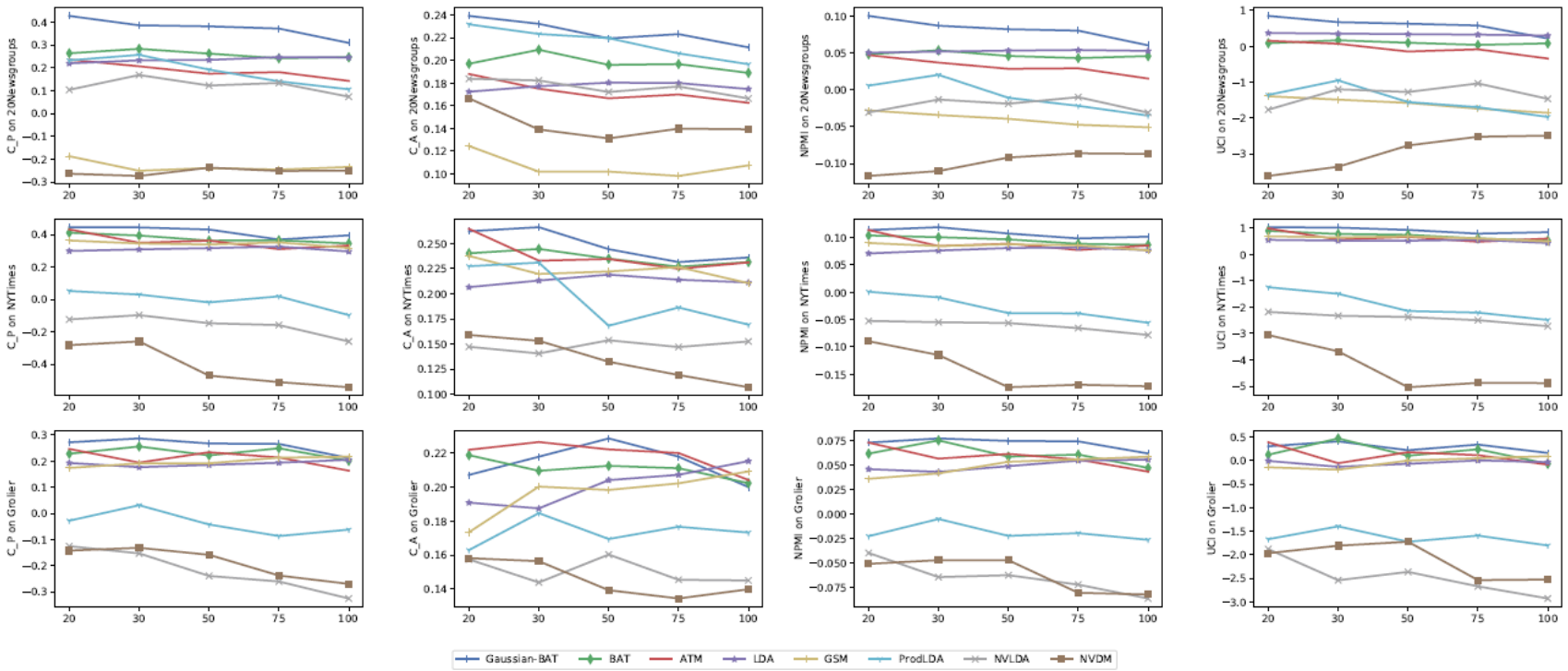


Figure 5: The comparison of average topic coherence vs. different topic number on 20Newsgroups, Grolier and NYTimes.

Результаты 3/3: Сравнение с известной разметкой

Описание эксперимента:

- Сопоставление разметок (известной и полученной) с помощью вергерского алгоритма[1]
- Оценка точности

Dataset	NVLDA	ProdLDA	LDA	BAT	G-BAT
20NG	33.31%	33.82%	35.36%	35.66%	**41.25%**

Table 4: Text clustering accuracy on 20Newsgroups (20NG). 'G-BAT' refers to 'Gaussian-BAT'. The best result is highlighted in bold.

[1] Harold W. Kuhn *The Hungarian method for the assignment problem* //Naval research logistics quarterly. – 1955. – Т. 2. – №. 1-2. – С. 83-97.

Почему эта статья?

- Нейросети + ТМ
- Учёт эмбеддингов слов
- Пример сравнения тематических моделей
- Два датасета, которые ещё ни разу не использовали

Содержание

Что есть интерпретируемость?

Примеры

- Байесовские сети
- Attention, heat-map

Предположения

- по модели: две модели дают одинаковые предсказания **iff** они используют одни и те же рассуждения,
- по предсказанию: модель выдаёт похожие предсказания на похожих входах **iff** её рассуждения похожи,
- по линейности: для рассуждений модели какие-то части входа важнее других.

Alon Jacovi, and Yoav Goldberg *Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?* //arXiv preprint arXiv:2004.03685. – 2020.

Почему эта статья?

- Чтобы задуматься
- Пример небольшого (по объёму) полностью теоретического исследования