Вероятностные иерархические векторные представления слов

Петр Алексеевич Остроухов

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем
Научный руководитель: Воронцов К.В.

17 апреля 2019 г.

План

- Существующие подходы
- Мотивация
  - Skip-Gram, Skip-Gram Negative Sampling
  - Тематическое моделирование и вероятностные эмбеддинги
  - Схожесть двух подходов
  - Преимущества вероятностных эмбеддингов
- Новизна
- Цель эксперимента и задачи
- Рассматриваемые задачи
  - Классификация
  - Word-similarity
- Предварительные результаты
- Дальнейшие улучшения
- Список литературы

Существующие подходы

- Частотные модели:
  - **tf-idf**:

$$\text{tf-idf} = \frac{n_{wd}}{\sum_{w'} n_{w'd}} \times \log \frac{|D|}{|\{d_i \in D | w \in d_i\}|}$$

- Нейросетевые модели:
  - **Skip-Gram Negative Sampling** [1]
  - **GloVe** [2]
  - ...
- **Вероятностные эмбеддинги слов** [3]

Мотивация [Skip-Gram] [1]

Описание

Модель обучает эмбеддинги путем предсказания локального контекста для каждого слова в корпусе. Вероятность слова u из локального контекста слова v:

$$p(u|v) = \frac{\exp \sum_t \phi_{ut}\theta_{tv}}{\sum_{w \in W} \exp \sum_t \phi_{wt}\theta_{tv}},$$

где $\Phi^{|W|\times|T|} = (\phi_{ut})$, $\Theta^{|T|\times|W|} = (\theta_{tv})$ — матрицы эмбеддингов.

Критерий

$$\mathcal{L} = \sum_{v \in W} \sum_{u \in W} n_{uv} \log p(u|v) \to \max_{\Phi,\Theta}.$$

Skip-Gram Negative Sampling (SGNS)

Описание

SGNS моделирует вероятность совстречаемости пары слов (u, v). Модель обучается на словах из локального контекста (v), а также на случайно семплированных словах из корпуса $(\bar{v})$.

Критерий

$$\mathcal{L} = \sum_{v \in W} \sum_{u \in W} n_{uv} \log \sigma\Big(\sum_t \phi_{ut}\theta_{tv}\Big) + k\mathsf{E}_{\bar{v}} \log \sigma\Big(-\sum_t \phi_{ut}\theta_{t\bar{v}}\Big) \to \max_{\Phi,\Theta}.$$

Модель эффективно обучается с помощью SGD.

Тематическое моделирование [4]

Описание

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

Критерий

$$\mathcal{L}(\Phi;\Theta) = \sum_{w \in W} \sum_{d \in D_p} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_{i} \tau_i R_i(\Phi,\Theta) \to \max_{\Phi;\Theta}$$

$$\sum_{w \in W} \phi_{wt} = \{0,1\},\ \phi_{wt} \geq 0,\ \sum_{t \in T} \theta_{td} = \{0,1\},\ \theta_{td} \geq 0.$$

Задача решается EM-алгоритмом или его онлайн модификацией для больших корпусов.

Вероятностные тематические эмбеддинги [3]

Описание

Рассмотрим модификацию тематической модели, предсказывающую слово u в локальном контексте слова v:

$$p(u|v) = \sum_{t \in T} p(u|t)p(t|v) = \sum_{t \in T} \phi_{ut}\theta_{tv}$$

$p(u|v)$ теперь описывает совстречаемости слов (!). $\Theta^{|T|\times|W|}$ — матрица вероятностей тем для слов (псевдодокументов).

Критерий

Log-likelihood maximization, как и в тематической модели.

Сходства двух описанных моделей

PWE	data type	$F_{uv} = \frac{n_{uv}}{n_v} = \hat{p}(u\|v)$
	objective	$\sum_{v \in W} n_v \, \mathrm{KL}\left(\hat{p}(u\|v) \,\middle\|\, \langle \phi_u \theta_v \rangle\right) \to \min_{\Phi,\Theta}$
	constrains	$\phi_{ut} > 0, \quad \sum_u \phi_{ut} = 1; \quad \theta_{tv} > 0, \quad \sum_t \theta_{tv} = 1$
	technique	EM-algorithm (online by F columns)
SGNS	data type	$F_{uv} = \log \frac{n_{uv} n}{n_u n_v} - \log k$
	objective	$\sum_{u \in W} \sum_{v \in W} n_{uv} \log \sigma\left(\langle \phi_u \theta_v \rangle\right) + k \, \mathbb{E}_{\bar{v}} \log \sigma\left(-\langle \phi_u \theta_v \rangle\right) \to \max_{\Phi,\Theta}$
	constrains	No constraints
	technique	SGD (online by corpus)

Преимущества вероятностных эмбеддингов

- **Интерпретируемость**: каждый вектор является компонентом матрицы [*слова*; *темы*]
- **Иерархичность**: категоризация происходит автоматически благодаря иерархии тем
- **Мультимодальность**:
  - Модальность категорий повысит интерпретируемость тем
  - Модальности разных языков позволят осуществлять кросс-язычный поиск

Цель эксперимента и подзадачи

Цель эксперимента

Построение иерархических тематических векторных эмбеддингов слов для нескольких модальностей, способных решать несколько задач NLP на уровне существующих аналогов (multi-task learning).

Подзадачи

- Построение мультимодальной иерархической тематической модели над корпусом Википедии.
- Взяв в качестве эмбеддингов слов пересчитанные по формуле Байеса строки полученной матрицы Φ:

$$emb(u) = p(t|u) = \frac{\phi(u|t)p(t)}{p(u)} = \frac{n_{ut}}{n_u},$$

проверить качество их работы на приведенных далее задачах и сравнить с существующими аналогами.

Новизна

- Тематические иерархические эмбеддинги слов
- Интерпретируемые и разреженные компоненты
- Предобученные на большой коллекции текстов с несколькими модальностями
- Способные эффективно решать несколько задач (Multi-task learning)

Рассматриваемые задачи

Классификация документов

- **Дано**: обучающая выборка $(d_i, y_i)_{i=1}^{l}$, $d_i \in D$ — множество документов, $y_i \in Y$ — множество меток классов.
- **Найти**: отображение $f : D \to Y$.
- **Критерий**: Accuracy, F-мера.
- **Датасеты**: Large Movie Review Dataset

Word-similarity

- **Дано**: множество слов W.
- **Найти**: для некоторого $w \in W$ слово $w' \in W \setminus \{w\}$:

$$w' = \arg \min_{w_s \in W \setminus \{w\}} \rho(f(w), f(w_s)).$$

- **Критерий**: корреляция с человеческими оценками.
- **Датасеты**: MEN, SIMLEX999, WS353.

Предварительные результаты

	tau_0	stop_words	num_words	epochs	emb_size	matrix	imdb	MEN	SIMLEX999	WS353
TF-IDF	-	-	-	-	-	-	0.871	-	-	-
GloVe	-	-	-	-	300	-	0.835	0.737	0.371	0.543
ARTM-offline	0	50	200K	5	100	theta	0.648	0.456	0.136	0.257
	0	50	200K	5	300	theta	0.639	0.473	0.172	0.318
	0	50	200K	5	500	theta	0.627	0.478	0.158	0.341
	0	50	200K	5	100	phi_bayes	0.622	0.389	0.140	0.229
	0	50	200K	5	300	phi_bayes	0.615	0.415	0.152	0.266
	0	50	200K	5	500	phi_bayes	0.605	0.418	0.122	0.267
ARTM-online	0	0	100K	1 (10)	400	phi_bayes	0.700	0.607	0.219	0.508
	0	0	300K	1 (10)	400	phi_bayes	0.701	0.594	0.208	0.512

Дальнейшие улучшения

- Модификация предобработки документов в задаче классификации (пропускать документ через построенную тематическую модель, таким образом получая тематический вектор документа)
- Подбор гиперпараметров: коэффициенты регуляризаторов и веса модальностей (слова и категории статей)
- Проверка нетематических моделей на неинтерпретируемость
- Добавление иерархий

Список литературы

1. T. Mikolov et al. *Distributed Representations of Words and Phrases and their Compositionality.* 2013
2. J. Pennington, R. Socher, C. D. Manning *GloVe: Global Vectors for Word Representation.* 2017
3. A. Potapenko, A. Popov, K. Vorontsov *Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks.* 2017
4. K. Vorontsov, A. Potapenko *Additive regularization of topic models.* 2015