Автоматическое выделение именованных сущностей в коллекциях текстовых документов

Хайруллин Ринат

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В.А. Серебряков

Москва,
2018 г.

Задача выделение именованных сущностей

Что такое именованная сущность?

Именованная сущность *– n-грамма в тексте, для которой определен класс. Классы:*

- ***Новостная тематика:*** *имена персон, названия организаций и геолокаций ...*
- ***Биологическая тематика:*** *названия протеинов, клеток ...*

[*Barack Obama*] *arrived this afternoon in* [*Washington, D.C.*]. [*President Obama*]*'s wife* [*Michelle*] *accompanied him* — **PERSON** **LOCATION**

[*TNF alpha*] *is produced chiefly by activated* [*macrophages*] — **PROTEIN** **CELL**

Подзадачи

1. *Выделение n-грамм в тексте.*
2. *Определение класса $y \in Y$ для каждой выделенной n-граммы. Y – некоторое заданное конечное множество классов.*

Постановка задачи

Требуется

Предложить алгоритм автоматического распознавания именованных сущностей в корпусе текстов.

Проблемы сущетсвующих алгоритмов

- *Требуется большой объем обучающей выборки.*
- *Допускается лексическая многозначность именованных сущностей.*

Предлагается

- *построить словарь n-грамм* $\mathcal{Q}$,
- *определять класс* $y \in Y = \{$***Персона***, ***Организация***, ***Геолокация***$\}$, *только для вхождений элементов словаря* $\mathcal{Q}$ *в текст (множество* $\mathcal{M}$*),*
- *для моделирования классов будем решать задачу трансдуктивного обучения на множестве объектов* $\mathcal{M}$.

Построение словаря n-грамм

$Y = \{\textbf{Персона}, \textbf{Организация}, \textbf{Геолокация}\}$

Можество $\mathcal{Q}$: все n-граммы удовлетворяющие следующим свойствам:

1. символьное представление:
   - все слова n-граммы начинаются с заглавной буквы,
   - кроме не более 2 слов подряд, длинной не более 3 символов,
2. $\rho(\{w_{d,k_1}, \ldots, w_{d,k_t}\}) > \alpha$, $\rho(\cdot)$ – значимость N-граммы.
3. последовательность частей речи соответствует виду ([причастие]$\{0,1\}$[прилагательное]$\{0,2\}$[существительное]$+$)

Примеры:
Салман ибн Абдул-Азиз Аль Сауд, *Объединенные арабские эмираты*.

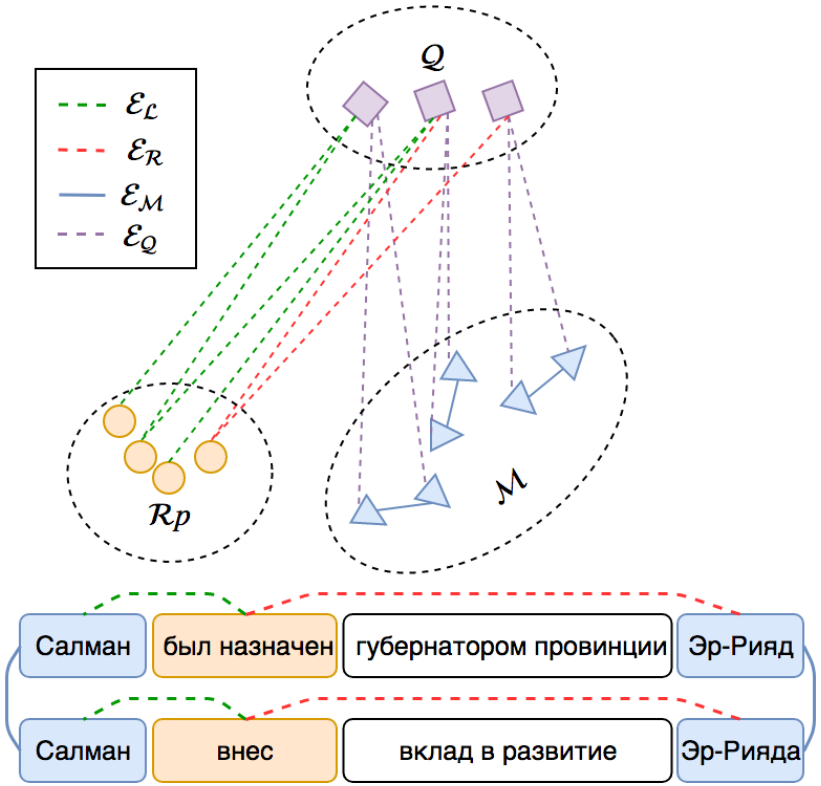
Precicion	Recall	F1
0.87	0.92	0.88

Связью r - будем называть n-грамму, последовательность частей речи, которой соответствует виду ([предлог]$|$[глагол] $+$ [предлог]$\{0,1\}$)
Можество $\mathcal{R}p$: все *связи* r.

Примеры: *был назначен на*, *происходит в*.

Представление корпуса текста в виде графа

- $\mathcal{Q}$ – множество выделенных n-грамм,
- $\mathcal{R}p$ – множество связей,
- $\mathcal{M}$ – множество словопозиций в тексте n-грамм из $\mathcal{Q}$.
- двудольные графы:
  - $\mathcal{G}_{\mathcal{Q}} = (\mathcal{M} \sqcup \mathcal{Q}, \mathcal{E}_{\mathcal{Q}})$
  - $\mathcal{G}_{L} = (\mathcal{M} \sqcup \mathcal{R}p, \mathcal{E}_{left})$
  - $\mathcal{G}_{R} = (\mathcal{M} \sqcup \mathcal{R}p, \mathcal{E}_{right})$
- *Knn*-граф: $\mathcal{W}_{\mathcal{M}} = (\mathcal{M}, \mathcal{E}_{\mathcal{M}}, f)$
- двудольные графы: $\mathcal{W}_{\{L,R\}} = (\mathcal{Q} \sqcup \mathcal{R}p, \mathcal{E}_{\{\mathcal{L},\mathcal{R}\}}, \nu)$. $\mathcal{W}_{\{L,R\}} = \mathcal{G}_{\mathcal{Q}}^{T}\mathcal{G}_{\{L,R\}}$



Задача распознавания именованных сущностей

Дано:
$\mathcal{M}_0 = \{(m, y)\}, y \in Y, |Y| = T$

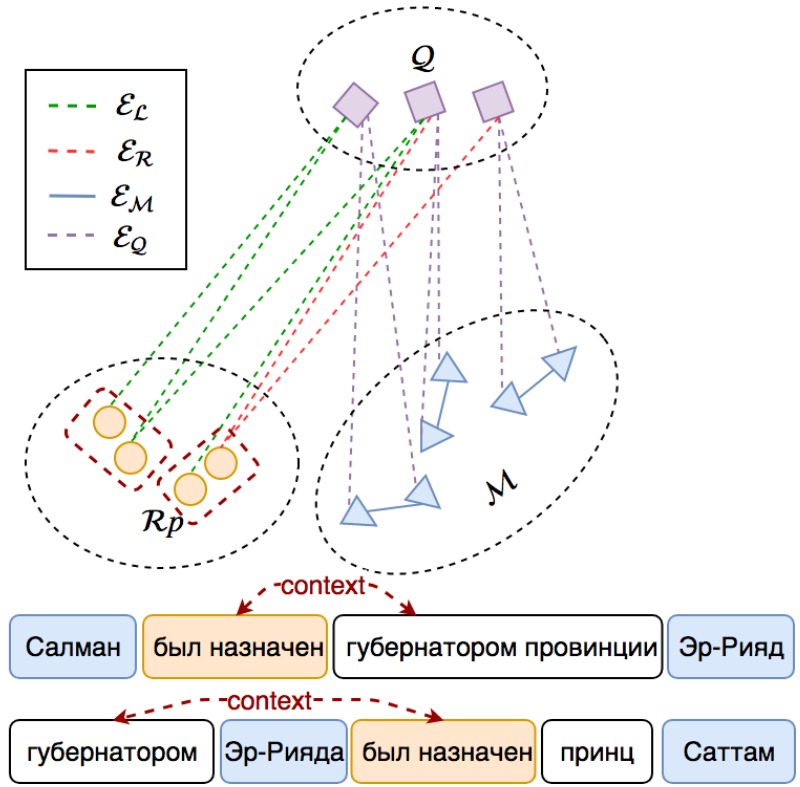
$\mathcal{G}_Q, \mathcal{G}_L, \mathcal{G}_R, \mathcal{W}_\mathcal{M}, \mathcal{W}_\mathcal{L}, \mathcal{W}_R$

Задача: Для каждого $m \in \mathcal{M} \setminus \mathcal{M}_0$ определить тип $y \in Y$

Индикаторы классов на множестве вершин графа:

$$\mathcal{M} : \mathcal{Y} \in \mathbb{R}^{m \times T},$$
$$\mathcal{Q} : \mathcal{C} \in \mathbb{R}^{n \times T},$$
$$\mathcal{R}p : \mathcal{P}_{\{\mathcal{L},\mathcal{R}\}} \in \mathbb{R}^{\ell \times T}$$

Решение: $y(m_i) = \text{argmax}_j \mathcal{Y}_i$



$$\mathcal{O} = \underbrace{\Omega_{\gamma,\mu}\left(\mathcal{Y}, \mathcal{Y}_0, \mathcal{C}, \mathcal{P}_{\{\mathcal{L},\mathcal{R}\}}\right)}_{\text{веса индикаторных матриц}} + \underbrace{\mathcal{L}_\alpha\left(\{\mathbf{F}_v, \mathbf{U}_v, \mathbf{V}_v, \beta_v\}, \mathbf{U}^*\right)}_{\text{кластеризация связей как MultiviewNMF задача}}$$

$$\mathbf{F}_v \in \left\{\mathcal{P}_{\{\mathcal{L},\mathcal{R}\}}, \mathbf{F}_{context}, \mathbf{F}_{characters}\right\}$$

Эксперимент

Цели эксперимента:

1. Изучение зависимости качества распознавания от размера начальной разметки.
2. Изучение зависимости качества распознавания от числа кластеров на множестве связей $\mathcal{R}p$.

Данные:

- Размеченные корпусы текстов **FactRuEval**[1] и **LABINFORM**[2], классы **Персона**, **Организация**, **Геолокация**.
  - размер корпуса $\sim$ 300000 слов,
  - словарь именованных сущностей:
    - **Персоны**: $\sim$ 6000 n-грамм, **Организации**: $\sim$ 4000 n-грамм, **Геолокации**: $\sim$ 2000 n-грамм,
  - именованных сущностей в корпусе:
    - **Персоны**: $\sim$ 12630 n-грамм, **Организации**: $\sim$ 10514 n-грамм, **Геолокации**: $\sim$ 8078 n-грамм,

Метрики качества: **Precicsion**, **Recall**, **F1 score**.

---

[1] https://github.com/dialogue-evaluation/factRuEval-2016

[2] http://labinform.ru/pub/named_entities/descr_ne.htm

Зависимость качества распознавания, от размера начальной выборки

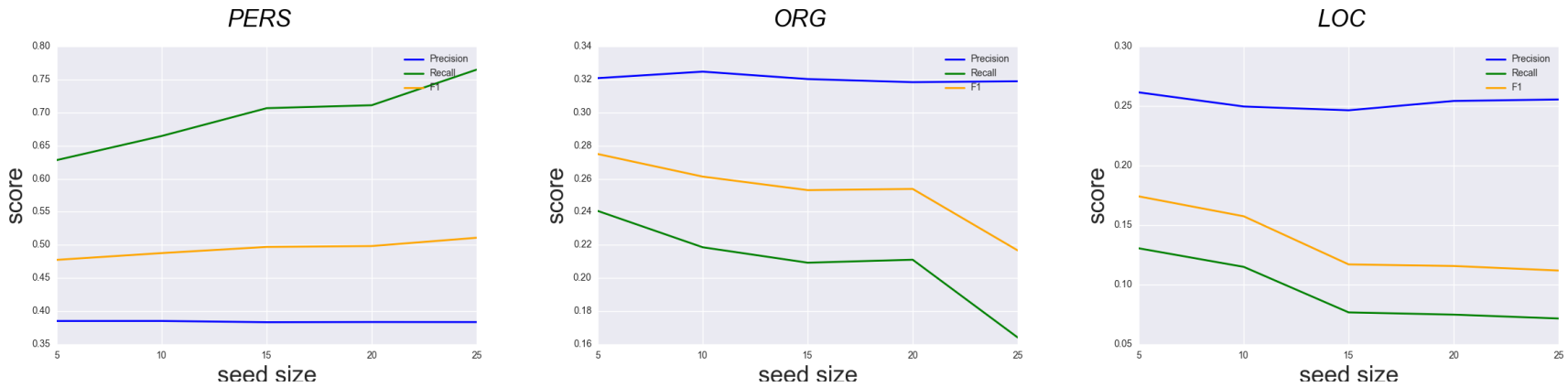


Figure: 10 кластеров

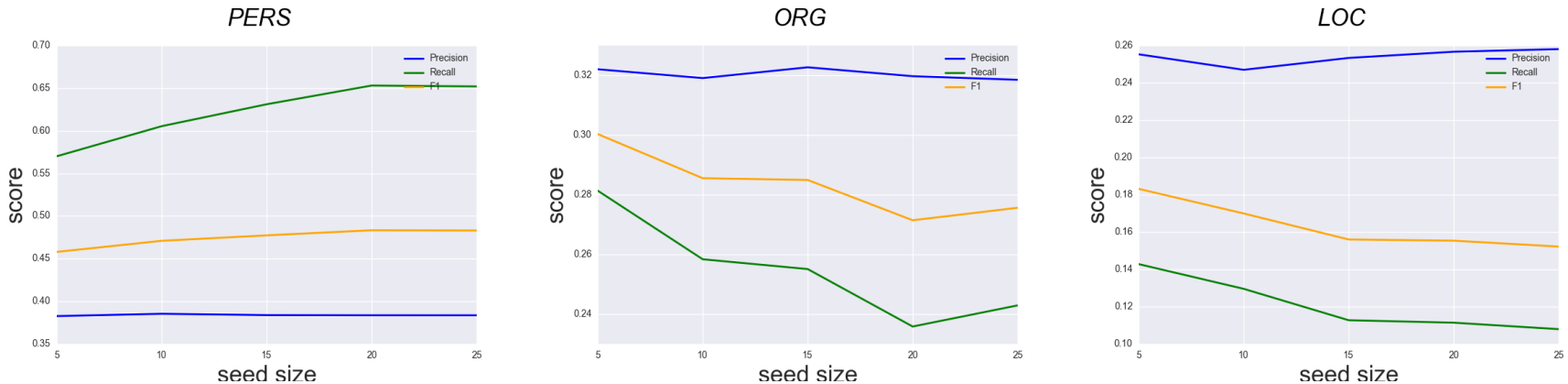


Figure: 100 кластеров

Результаты эксперимента

Результаты на корпусе FactRuEval и LABINFORM

Алгоритм	**Precicion**			**Recall**			**F1**		
	Person	Location	Organisation	Person	Location	Organisation	Person	Location	Organisation
CLUS	0.38	0.25	0.32	**0.65**	0.10	0.24	**0.48**	0.15	0.27
NOCLUS	0.39	0.25	0.32	0.54	**0.16**	**0.32**	0.45	**0.2**	**0.32**

Заключение

1. Предложен и реализован алгоритм трансдуктивного обучения для решения задачи распознавания именованных сущностей.
2. Проведен эксперимент на размеченном корпусе на русском языке:
   - Алгоритм показывает низкое качество распознавания на корпусе текстов малого объема.

План дальнейших работ

- Провести эксперименты на расширенном корпусе текстов, исследовать динамику изменения качества при увеличении выборки.
- Сравнить работу алгоритма с существующими решениями для русского языка.
- Сравнить работу алгоритма с существующими решениями для русского языка.

Функция потерь

$$\mathbf{D}_{\mathcal{M},i,i} = \sum_{j}^{|\mathcal{M}|} W_{\mathcal{M},i,j}, \quad \mathbf{D}^{\mathcal{Q}}_{Z,i,i} = \sum_{j}^{|\mathcal{R}p|} W_{Z,i,j}, \quad \mathbf{D}^{\mathcal{R}p}_{Z,j,j} = \sum_{i}^{|\mathcal{Q}|} W_{Z,i,j}$$

$$\Omega_{\gamma,\mu}\left(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}}_L, \boldsymbol{\mathcal{P}}_R\right) = \|\boldsymbol{\mathcal{Y}} - (\boldsymbol{\mathcal{G}}_Q\boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{G}}_L\boldsymbol{\mathcal{P}}_L + \boldsymbol{\mathcal{G}}_R\boldsymbol{\mathcal{P}}_R)\|_F^2 + \mu\|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{Y}}_0\|_F^2$$
$$+ \frac{\gamma}{2}\sum_{i,j}^{|\mathcal{M}|} W_{\mathcal{M},i,j}\left\|\frac{\boldsymbol{\mathcal{Y}}_i}{\sqrt{\mathbf{D}_{\mathcal{M},i,i}}} - \frac{\boldsymbol{\mathcal{Y}}_i}{\sqrt{\mathbf{D}_{\mathcal{M},j,j}}}\right\|_2^2$$
$$+ \sum_{Z\in\{L,R\}}\sum_{i}^{|\mathcal{Q}|}\sum_{j}^{|\mathcal{R}p|} W_{Z,i,j}\left\|\frac{\boldsymbol{\mathcal{C}}_i}{\sqrt{\mathbf{D}^{\mathcal{Q}}_{Z,i,i}}} - \frac{\boldsymbol{\mathcal{P}}_{Z,j}}{\sqrt{\mathbf{D}^{\mathcal{R}p}_{Z,i,i}}}\right\|_2^2$$

$$\mathbf{F}_v \in \left\{\boldsymbol{\mathcal{P}}_{\{\mathcal{L},\mathcal{R}\}}, \mathbf{F}_{context}, \mathbf{F}_{characters}\right\}$$

$$\mathcal{L}_\alpha\left(\{\mathbf{F}_v, \mathbf{U}_v, \mathbf{V}_v, \beta_v\}, \mathbf{U}^*\right) = \sum_v \left(\beta_v\|\mathbf{F}_v - \mathbf{U}_v\mathbf{V}_v^T\|_F^2 + \alpha\|\mathbf{U}_v\mathbf{H}_v - \mathbf{U}^*\|_F^2\right)$$

Задача минимизации

$$\min_{\mathcal{Y},\mathcal{C},\mathcal{P}_{\{\mathcal{L},\mathcal{R}\}},\{\mathbf{U}_v,\mathbf{V}_v,\beta_v\},\mathbf{V}^*} \underbrace{\Omega_{\gamma,\mu}\left(\mathcal{Y},\mathcal{Y}_0,\mathcal{C},\mathcal{P}_{\{\mathcal{L},\mathcal{R}\}}\right)}_{\text{веса индикаторных матриц}} + \underbrace{\mathcal{L}_\alpha\left(\{\mathbf{F}_v,\mathbf{U}_v,\mathbf{V}_v,\beta_v\},\mathbf{V}^*\right)}_{\text{кластеризация связей как MultiNMF задача}}$$

$$\text{s.t.} \{\mathbf{U}_v,\mathbf{V}_v\},\mathbf{V}^* \geq 0, \quad \sum_v \exp(-\beta_v) = 1$$

Алгоритм оптимизации

Algorithm 1 Optimization procedure

$\{\mathcal{Y}, \mathcal{C}, \mathcal{P}_L, \mathcal{P}_R\} \leftarrow \{\mathcal{Y}_0, \mathcal{G}_Q^T \mathcal{Y}_0, \mathcal{G}_L^T \mathcal{Y}_0, \mathcal{G}_R^T \mathcal{Y}_0\}$
$\{\mathbf{U}_v, \mathbf{V}_v, \beta_v\}, \mathbf{V}^* \leftarrow$ положительные числа
repeat
    обновляем $\mathcal{Y}, \mathcal{C}, \mathcal{P}_L, \mathcal{P}_R$
    for $v \in d$ **do**
        repeat
            обновляем $\mathbf{U}, \mathbf{V}$
        until δ_v converges
    endfor
    обновляем $\mathbf{V}^*, \beta_v$
until Objective in Eq.(4) converges