Выделение типов именованных сущностей используя контекстную совстречаемость слов

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.

Общий вид постановки задачи

Дано:

$\mathcal{D} = \{d_1, \ldots, d_D\}$ – копрус текстов,
$\mathcal{V} = \{w_1, \ldots, w_N\}$ – представление копруса $\mathcal{D}$ в виде последовательности слов,
$\mathcal{T} = \{t_1, \ldots, t_T\}$ – множество типов именованных сущностей.

Задача:
Определить тип $t \in \mathcal{T}$ для каждого слова $w \in \mathcal{V}$.

Проблемы существующих алгоритмов:

- отбор признаков важнее выбора модели,
- неустойчивость к смене предметной области,
- требуется большой объем обучающей выборки.

Цель работы

Задача

Построить алгоритм выделения именованных сущностей, обладающий следующими свойствами:

1. *независимость от предметной области,*
2. *использование универсальных признаков,*
3. *использование частично размеченной выборки.*

Способ решения

- ***независимость от предметной области**:*
  выделение значимых слов и фраз в тексте, которые могут быть именованными сущностями,
- ***универсальные признаки**:*
  1. *выделение слов и фраз, которые задают отношения между словами и фразами в предложении,*
  2. *кластеризация выделенных отношений на основе контекста,*
- ***использование частичной разметки выборки***

Постановка задачи

Дано: $\mathcal{D}, \quad (\Psi, \mathcal{T}_\Psi, \sigma), \quad \mathcal{T} \subset \mathcal{T}_\Psi$

$\phi = \{w_{d,i}, \ldots, w_{d,i+k}\}$ $k \geq 0$ – значимая фраза, если $\rho(\phi) > \alpha$
$\rho(\cdot)$ – некоторая функция значимости, α – порог значимости

Задача:

- построить множество значимых фраз $\mathcal{Q}$

  $\mathcal{M}_d(q) = \{i \mid \phi_i = q\} \quad i \in \{\tilde{n}_{d-1} + 1, \ldots, \tilde{n}_d\}$

  $\mathcal{M} = \bigcup_{d \in \mathcal{D}} \bigcup_{q \in \mathcal{Q}} \mathcal{M}_d(q)$
- построить множество отношений $\mathcal{R}$

  $\mathcal{L}_d(r) = \{i \mid \phi_i = r\} \quad i \in \{\tilde{n}_{d-1} + 1, \ldots, \tilde{n}_d\}$

  $\mathcal{L} = \bigcup_{d \in \mathcal{D}} \bigcup_{r \in \mathcal{R}} \mathcal{L}_d(r)$
- выделить подмножество $\mathcal{Q}_L = \{q \in \mathcal{Q} \mid q \in \Psi, \mathcal{T}(q) \cap \mathcal{T} = \{t\}\}$
- для всех $q \in \mathcal{Q}_U = \mathcal{Q} \setminus \mathcal{Q}_L$ в корпусе $\tilde{\mathcal{D}} = sort(\mathcal{M} \bigcup \mathcal{L})$

  $t(q_i) = \underset{j \in \mathcal{T}}{\operatorname{argmax}} \mathbf{y}(q_i)_j \quad \mathbf{y}(q_i) \in [0,1]^{|\mathcal{T}|} \quad i \in \mathcal{M}(q) \quad j \in \{1, \ldots, |\mathcal{T}|\}$

Пример графового представления

```
... [в]:PR [РФ]:EM [Воробьев]:EM [был назначен]:RP ...
... [Халиков]:EM [был назначен на]:RP должность торгпреда [РФ]:EM
[в]:RP [Пакистане]:EM ...
... [Арам Габрелянов]:EM [в]:RP феврале [был назначен]:RP
[председателем совета директоров]:EM ...

... [пишет в]:RP среду [газета Известия]:EM ...
... [в]:RP среду [газета Известия]:EM [написала]:RP что ....
... указанная информация впервые [была распространена]:RP [газетой
Известия]:EM ...
```

Представление данных в виде графа

$|\mathcal{M}| = m \quad |\mathcal{Q}| = n \quad |\mathcal{R}| = k$

$\mathcal{G}_{\mathcal{Q}} = \left[\mathbb{1}_{\{\phi_i = q_j\}}\right] \in \mathbb{R}^{m \times n}$ – матрица соответствий $\mathcal{M}$ и $\mathcal{Q}$

$\mathcal{G}_{\mathcal{Z}} = \left[\mathbb{1}_{\{\phi_{j+\mathbf{1}} = q_i, i,j \in s\}}\right] \in \mathbb{R}^{m \times k}$ – матрица соответсвий слева/справа $\mathcal{M}$ и $\mathcal{R}$ в предложении s, где $\mathcal{Z} \in \{L, R\}$

$\mathcal{W}_{\mathcal{Z}} = \mathcal{G}_{\mathcal{Q}}^T \mathcal{G}_{\mathcal{Z}} \in \mathbb{R}^{n \times k}$

$\mathcal{W}_{\mathcal{M}} \in \mathbb{R}^{n \times n}$ – матрица соотвествий между значимыми фразами

Для каждого $\phi_i \in \mathcal{M}$ определим вектор весов q_j-х, которые встречаются с ϕ_i в одном предложении s

$$\mathbf{f}_{\phi_i} = \{\nu_s(q_j) \log(\frac{|\mathcal{D}|}{\nu_{\mathcal{D}}(q_j)}) \mid \phi_i, q_j \in s\} \in \mathbb{R}^n$$

$$W_{\mathcal{M},i,j} = \begin{cases} \exp(\frac{-\|\mathbf{f}_{\phi_i} - \mathbf{f}_{\phi_j}\|}{t}) & \text{если } \mathbf{f}^{(i)} \in N_k(\mathbf{f}^{(j)}) \text{ или } \mathbf{f}^{(j)} \in N_k(\mathbf{f}^{(i)}) \\ & \text{и } \phi_i = \phi_j \\ 0 & \text{иначе} \end{cases}$$

Гипотезы

Hypothesis (H1)

Если фраза ϕ часто встречается слева/справа от отношения r в предложениях, то ϕ и r должны относится к одному типу t.

Hypothesis (H2)

Если в контекстах двух фраз ϕ_i и ϕ_j много общих слов и $\phi_i = \phi_j$, то ϕ_i и ϕ_j должны относится к одному типу t.

Hypothesis (H3)

Два отношения r_i и r_j относятся к одному кластеру если у них много общих контекстных слов.

Hypothesis (H4)

Два отношения r_i и r_j относятся к одному типу t, если они принадлежат одному кластеру.

Оптимизация

$\phi_i \in \mathcal{M}$: $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{m \times T}$, $q_i \in \mathcal{Q}$: $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{n \times T}$, $r_i \in \mathcal{R}$: $\boldsymbol{\mathcal{P}}_{\mathcal{Z}} \in \mathbb{R}^{k \times T}$

$\alpha, \gamma, \mu \in [0, 1]$

$$\mathcal{F}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}}_L, \boldsymbol{\mathcal{P}}_R) = \sum_{\mathcal{Z} \in \{L,R\}} \sum_{i=1}^{n} \sum_{j=1}^{k} W_{\mathcal{Z},i,j} \| \frac{\boldsymbol{\mathcal{C}}_i}{\sqrt{\mathbf{D}_{\mathcal{Q},\mathcal{Z},i,i}}} - \frac{\boldsymbol{\mathcal{P}}_{\mathcal{Z},j}}{\sqrt{\mathbf{D}_{\mathcal{R},\mathcal{Z},j,j}}} \|_2^2 \quad (1)$$

$$\Omega_{\gamma,\mu}(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}}_L, \boldsymbol{\mathcal{P}}_R) = \frac{1}{2} \|\boldsymbol{\mathcal{Y}} - f(\boldsymbol{\mathcal{G}}_{\mathcal{Q}} \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{G}}_L \boldsymbol{\mathcal{P}}_L, \boldsymbol{\mathcal{G}}_R \boldsymbol{\mathcal{P}}_R)\|_F^2 + \frac{\gamma}{2} \sum_{i,j=1}^{n} W_{\mathcal{M},i,j} \| \frac{\boldsymbol{\mathcal{Y}}_i}{\sqrt{\mathbf{D}_{\mathcal{M},i,i}}} - \frac{\boldsymbol{\mathcal{Y}}_j}{\sqrt{\mathbf{D}_{\mathcal{M},j,j}}} \|_2^2 + \frac{\mu}{2} \|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{Y}}_0\|_F^2 \quad (2)$$

$$\mathcal{L}_\alpha(\boldsymbol{\mathcal{P}}_L, \boldsymbol{\mathcal{P}}_R, \{\mathbf{U}_v, \mathbf{V}_v, \beta_v\}, \mathbf{V}^*) = \sum_{v=0}^{d} \beta_v (\|\mathbf{F}_v^T - \mathbf{U}_v \mathbf{V}_v^T\|_F^2 + \alpha \|\mathbf{V}_v \mathbf{Q}_v - \mathbf{V}^*\|_F^2) \quad (3)$$

$\mathbf{D}_{\mathcal{Q},\mathcal{Z}} = \sum_{i=1}^{n} W_{\mathcal{Z},i,j}$, $\mathbf{D}_{\mathcal{R},\mathcal{Z}} = \sum_{j=1}^{k} W_{\mathcal{Z},i,j}$, $\boldsymbol{\mathcal{S}}_{\mathcal{Z}} = \mathbf{D}_{\mathcal{Q},\mathcal{Z}}^{-\frac{1}{2}} \boldsymbol{\mathcal{W}}_{\mathcal{Z}} \mathbf{D}_{\mathcal{R},\mathcal{Z}}^{-\frac{1}{2}}$

Аналогично определяется $\boldsymbol{\mathcal{S}}_{\mathcal{M}}$

Оптимизация

$$\min_{\mathcal{Y},\mathcal{C},\mathcal{P}_L,\mathcal{P}_R,\{\mathbf{U}_v,\mathbf{V}_v,\beta_v\},\mathbf{V}^*} \quad (\mathcal{F} + \Omega_{\gamma,\mu} + \mathcal{L}_\alpha)$$
$$\text{s.t.} \quad \{\mathbf{U}_v, \mathbf{V}_v\}, \mathbf{V}^* \geq 0, \quad \sum_{v=1}^{d} \exp(-\beta_v) = 1 \tag{4}$$

Решение: Определим тип $t(q_i) = \text{argmax}\ \mathcal{Y}_i$

Оценки параметров:

$$\mathcal{Y} = \left[(1+\gamma+\mu)\mathbf{1} - \gamma\mathcal{S}_{\mathcal{M}}\right]^{-1}\left[f(\mathcal{G}_{\mathcal{Q}}\mathcal{C}, \mathcal{G}_L\mathcal{P}_L, \mathcal{G}_R\mathcal{P}_R)^q + \mu\mathcal{Y}_0^q\right] \tag{5}$$

$$\mathcal{C} = \frac{1}{2}\left[\mathcal{S}_L\mathcal{P}_L + \mathcal{S}_R\mathcal{P}_R + \mathcal{G}_{\mathcal{Q}}^T(\mathcal{Y} - \mathcal{G}_L\mathcal{P}_L - \mathcal{G}_R\mathcal{P}_R)\right] \tag{6}$$

$$\mathcal{P}_L = \left[(1+\beta_0)\mathbf{1} + \mathcal{G}_L^T\mathcal{G}_L\right]^{-1}\left[\mathcal{S}_L^T\mathcal{C} + \mathcal{G}_L^T\left[\mathcal{Y} - \mathcal{G}_Q\mathcal{C} - \mathcal{G}_R\mathcal{P}_R\right] + \beta_0\mathbf{V}_0\mathbf{U}_0^T\right] \tag{7}$$

$$\mathcal{P}_R = \left[(1+\beta_1)\mathbf{1} + \mathcal{G}_R^T\mathcal{G}_R\right]^{-1}\left[\mathcal{S}_R^T\mathcal{C} + \mathcal{G}_R^T\left[\mathcal{Y} - \mathcal{G}_Q\mathcal{C} - \mathcal{G}_L\mathcal{P}_L\right] + \beta_1\mathbf{V}_1\mathbf{U}_1^T\right] \tag{8}$$

Кластеризация на множестве отношений

$|\mathcal{R}| = k$, $r_i \in \mathcal{R}$: $\boldsymbol{\mathcal{P}}_{\mathcal{Z}} \in \mathbb{R}^{k \times T}$, $\mathbf{F}_{\mathcal{N}_s} \in \mathbb{R}^{k \times n_s}$, $\mathbf{F}_{\mathcal{N}_c} \in \mathbb{R}^{k \times n_c}$
$\{\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3\} = \{\boldsymbol{\mathcal{P}}_L, \boldsymbol{\mathcal{P}}_R, \mathbf{F}_{\mathcal{N}_s}, \mathbf{F}_{\mathcal{N}_c}\}$ – признаковые представления
$\mathbf{U}_v \in \mathbb{R}^{N_v \times C}_{\geq 0}$, $\mathbf{V}_v \in \mathbb{R}^{k \times C}_{\geq 0}$, $\mathbf{V}^* \in \mathbb{R}^{k \times C}_{\geq 0}$ $\quad \mathbf{Q} \in \mathbb{R}^{C \times C}_{\geq 0}$ $\quad Q_{v,c,c} = \sum_{i=1}^{N_v} U_{v,i,c}$

$$\min_{\{\mathbf{U}_v, \mathbf{V}_v, \beta_v\}, \mathbf{V}^*} \quad \sum_{v=0}^{d} \beta_v \underbrace{(\|\mathbf{F}_v^T - \mathbf{U}_v \mathbf{V}_v^T\|_F^2 + \alpha \|\mathbf{V}_v \mathbf{Q}_v - \mathbf{V}^*\|_F^2)}_{\delta_v}$$

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

$$U_{i,k} \leftarrow U_{i,k} \frac{(\mathbf{FV})_{i,k} + \alpha \sum_{j=1}^{k} V_{j,k} V^*_{j,k}}{(\mathbf{U}\mathbf{V}^T\mathbf{V})_{i,k} + \alpha \sum_{l=1}^{N_v} U_{l,k} \sum_{j=1}^{k} V^2_{j,k}} \tag{9}$$

$$\mathbf{U} \leftarrow \mathbf{U}\mathbf{Q}^{-1} \quad \mathbf{V} \leftarrow \mathbf{V}\mathbf{Q} \tag{10}$$

$$V_{i,k} \leftarrow V_{i,k} \frac{(\mathbf{F}^T\mathbf{U})_{i,k} + \alpha V^*_{j,k}}{(\mathbf{V}\mathbf{U}^T\mathbf{U})_{j,k} + \alpha V_{j,k}} \tag{11}$$

$$\mathbf{V}^* = \frac{\sum_{v=1}^{d} \alpha \mathbf{V}_v \mathbf{Q}_v}{d\alpha} \quad \beta_v = -\log \frac{\delta_v}{\sum_{v=1}^{d} \delta_v} \tag{12}$$

$\mathbf{U}_0\mathbf{Q}_0^{-1} \approx \mathbf{U}_1\mathbf{Q}_1^{-1} \approx [p(t_i|c)]$, $\mathbf{Q}_0\mathbf{V}_0^T \approx \mathbf{Q}_1\mathbf{V}_1^T \approx [p(r_i, c)]$

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
  $\mathcal{Y}, \mathcal{C}, \mathcal{P}_L, \mathcal{P}_R \leftarrow Eq.(5-8)$
  for $v = 0$ to 3 **do**
    repeat
      $\mathbf{U} \leftarrow Eq.(9)$
      $\mathbf{U}, \mathbf{V} \leftarrow Eq.(10)$
      $\mathbf{V} \leftarrow Eq.(11)$
    until δ_v converges
  endfor
  $\mathbf{V}^*, \beta_v \leftarrow Eq.(12)$
until Objective in Eq.(4) converges

Матрицы признаков

Признаки:

$\mathcal{N}_s = \{w_1, \ldots, w_{n_s}\}$ – все уникальные слова входящие в $\mathcal{R}$

$\mathcal{N}_c = \{\tilde{w}_1, \ldots, \tilde{w}_{n_c}\}$ – все уникальные слова в окне из 10 слов для фраз отношений

Определим матрицы $\mathbf{F}_{\mathcal{N}_s} \in \mathbb{R}^{k \times n_s}, \mathbf{F}_{\mathcal{N}_c} \in \mathbb{R}^{k \times n_c}$, как:

$$F_{\mathcal{N}_s,i,j} = \begin{cases} 1 & \text{если } w_j \in r_i \\ 0 & \text{иначе} \end{cases}$$

$$F_{\mathcal{N}_c,i,j} = TFIDF(\tilde{w}_{i,j})$$

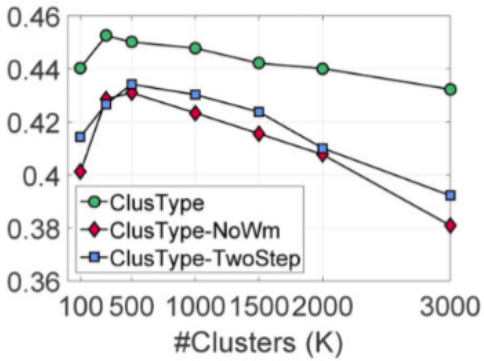
Эксперименты

$|\mathcal{V}| = 300000$, $|\mathcal{Q}| = 10000$, $|\mathcal{R}| = 6000$, $|\mathcal{M}| = 31000$.
$|\mathcal{M}_0| = 2000$

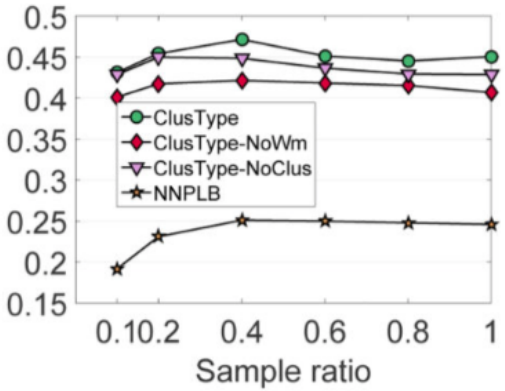
Table: Результаты на корпусе FactRuEval и LABINFORM

	Type	presision	recall	F1
No Clus	Pers	0.39	0.54	0.45
No Clus	Loc	0.25	0.16	0.2
No Clus	Org	0.32	0.3	0.31
Clus	Pers	0.38	0.65	0.48
Clus	Loc	0.25	0.1	0.15
Clus	Org	0.32	0.24	0.27

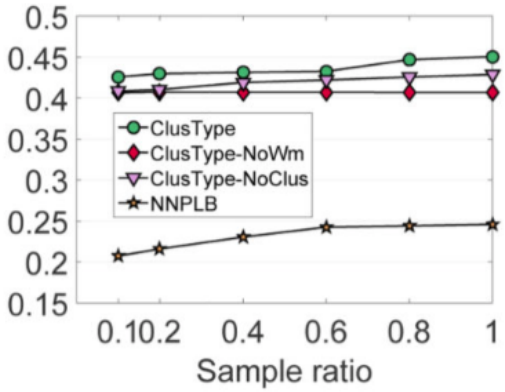
Эксперименты



(a) #Clusters



(b) Seed set size



(c) Corpus size