

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.

Требуется

Предложить алгоритм автоматического распознавания именованных сущностей в корпусе текстов.

Проблемы существующих алгоритмов

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

Предлагается

- *построить словарь n -грамм \mathcal{Q} ,*
- *моделировать классы $y \in Y = \{\text{Персона, Организация, Геолокация}\}$, только для вхождений элементов словаря \mathcal{Q} в текст (множество M),*
- *зная метки классов на некотором подмножестве $M_0 \subset M$, построить алгоритм получения меток классов на неразмеченной части множества $\{m \in M \setminus M_0\}$.*

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

Можество Q : все n-граммы удовлетворяющие следующим свойствам:

- 1 символное представление:
 - все слова n-граммы начинаются с заглавной буквы,
 - кроме не более 2 слов подряд, длиной не более 3 символов,
- 2 $\rho(\{w_{d,k_1}, \dots, w_{d,k_i}\}) > \alpha$, $\rho(\cdot)$ – 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.

Precision	Recall	F1
0.87	0.92	0.88

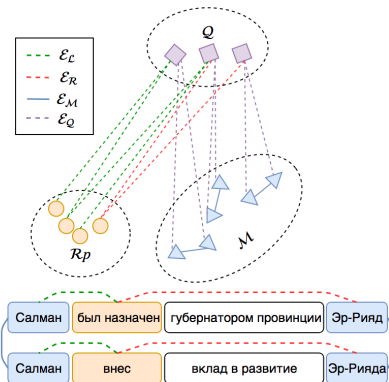
Связью r - будем называть n-грамму, последовательность частей речи, которой соответствует виду ([предлог][глагол] + [предлог]{0, 1})

Можество \mathcal{R}_p : все связи r .

Примеры: *был назначен на, происходит в.*

Представление корпуса текста в виде графа

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



Задача распознавания именованных сущностей

Дано:

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

$$\mathcal{G}_Q, \mathcal{G}_L, \mathcal{G}_R, \mathcal{W}_M, \mathcal{W}_L, \mathcal{W}_R$$

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

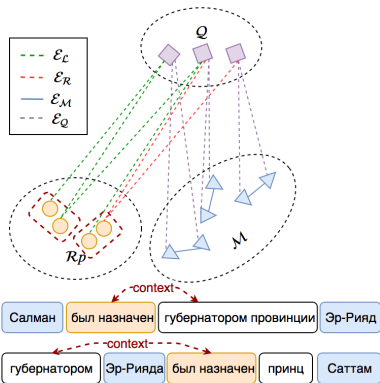
Индикаторы классов на множестве вершин графа:

$$M: \mathcal{Y} \in \mathbb{R}^{m \times T},$$

$$Q: \mathcal{C} \in \mathbb{R}^{n \times T},$$

$$\mathcal{R}_p: \mathcal{P}_{\{L, R\}} \in \mathbb{R}^{\ell \times T}$$

Решение: $y(m_i) = \arg\max_j \mathcal{Y}_i$



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

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

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

Метрики качества: Precision, Recall, F1 score.

¹<https://github.com/dialogue-evaluation/factRuEval-2016>

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

Зависимость качества распознавания, от размера начальной выборки

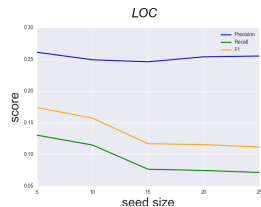
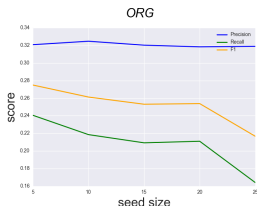
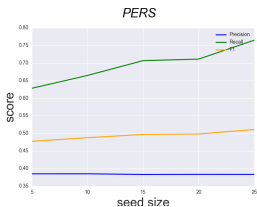


Figure: 10 кластеров

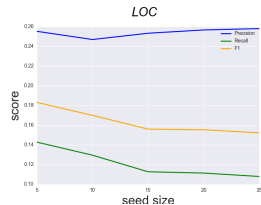
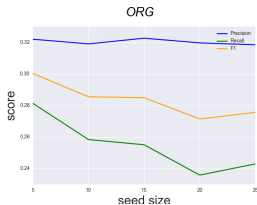
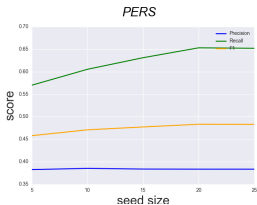
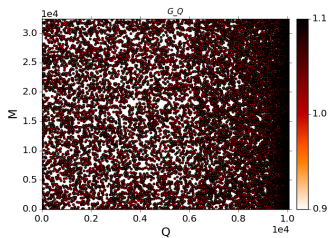
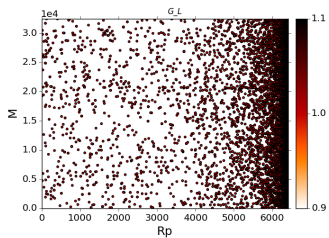


Figure: 100 кластеров

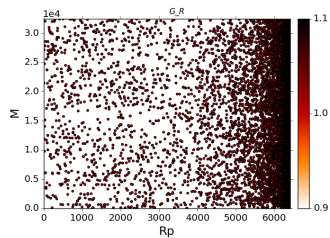
Матрицы графов



G_Q



G_L



G_R

Результаты на корпусе FactRuEval и LABINFORM

Алгоритм	Person			Location			Organization		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
<i>CLUS</i>	0.38	0.65	0.48	0.25	0.10	0.15	0.32	0.24	0.27
<i>NOCLUS</i>	0.39	0.54	0.45	0.25	0.16	0.2	0.32	0.32	0.32

- 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.
- Попытаться обобщить алгоритм для использования на корпусах текстов других тематик.

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

$$\begin{aligned} \Omega_{\gamma,\mu}(\mathbf{Y}, \mathbf{C}, \mathcal{P}_L, \mathcal{P}_R) &= \|\mathbf{Y} - (\mathbf{G}_Q \mathbf{C} + \mathbf{G}_L \mathcal{P}_L + \mathbf{G}_R \mathcal{P}_R)\|_F^2 + \mu \|\mathbf{Y} - \mathbf{Y}_0\|_F^2 \\ &+ \frac{\gamma}{2} \sum_{i,j}^{|\mathcal{M}|} W_{\mathcal{M},i,j} \left\| \frac{\mathbf{y}_i}{\sqrt{\mathbf{D}_{\mathcal{M},i,i}}} - \frac{\mathbf{y}_j}{\sqrt{\mathbf{D}_{\mathcal{M},j,j}}} \right\|_2^2 \\ &+ \sum_{Z \in \{L,R\}} \sum_i^{|\mathcal{Q}|} \sum_j^{|\mathcal{R}_p|} W_{Z,i,j} \left\| \frac{\mathbf{c}_i}{\sqrt{\mathbf{D}_{Z,i,i}^{\mathcal{Q}}}} - \frac{\mathcal{P}_{Z,j}}{\sqrt{\mathbf{D}_{Z,i,i}^{\mathcal{R}_p}}} \right\|_2^2 \end{aligned}$$

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

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

Задача минимизации

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

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