

# Математические методы анализа текстов Задачи разметки, условные случайные поля (CRF)

К. В. Воронцов, М. А. Апишев, А. С. Попов

Этот курс доступен на странице вики-ресурса  
<http://www.MachineLearning.ru/wiki>  
«Математические методы анализа текстов  
(курс лекций) / осень 2019»

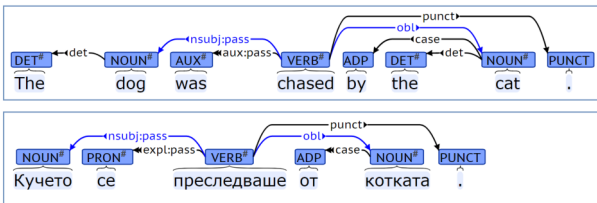
25 сентября 2019

- 1 Задачи обучения с учителем для разметки текста**
  - Примеры задач разметки и сегментации
  - Лог-линейная модель разметки
  - Линейный CRF: формальная постановка задачи
- 2 Обучение линейного CRF**
  - Алгоритм Витерби
  - Вычисление градиента
  - Алгоритм вперёд–назад
- 3 Краткий обзор модификаций и обобщений CRF**
  - Регуляризация и отбор признаков
  - CRF с частичным обучением
  - Обобщения CFR

## Примеры задач разметки и сегментации

- распознавание частей речи (part of speech tagging, POS)
- неглубокий синтаксический разбор (shallow syntax parsing)
- распознавание именованных сущностей (named entity, NER)
- выделение семантических ролей (semantic role labeling)
- анализ тональности заданной сущности (sentiment analysis)
- выделение текстовых полей данных (slot filling)
- выделение полей в библиографических записях
- сегментация научных или юридических текстов
- поиск кореференций и разрешение анафор
- поиск и разрешение эллипсиса (гэппинга)
- перевод речевого сигнала в текст
- перевод музыкального сигнала в нотную запись
- выделение генов в нуклеотидных последовательностях

## Пример частеречной и синтаксической разметки



Теги частей речи (не все, и могут зависеть от языка):

NOUN	noun	существительное	INTJ	interjection	междометие
PROPN	proper noun	имя собственное	ADP	adposition	предлог
ADJ	adjective	прилагательное	CONJ	conjunction	союз
VERB	verb	глагол	PART	particle	частица
ADV	adverb	наречие	PUNCT	punctuation	знак пунктуации
PRON	pronoun	местоимение	SYM	symbol	символ
NUM	numeral	числительное	X	other	иное

<http://universaldependencies.org/>

## Пример выделения частей речи русского языка методом CRF

Часть речи	Отн. частота ЧР, %	Точность, %	Полнота, %	F1, %
Существительное	30.42	96.03	96.98	96.50
Прилагательное	9.40	92.45	92.16	92.30
Глагол	9.12	98.32	98.86	98.59
Причастие	0.76	82.37	82.58	82.48
Деепричастие	0.24	94.80	90.11	92.40
Наречие	4.17	96.43	96.07	96.25
Предлог	9.83	99.39	99.61	99.50
Союз	5.92	99.40	99.54	99.47
Числительное (как слово)	0.64	90.27	89.22	89.74
Числительное (как цифра)	1.56	92.80	94.78	93.78
Личное местоимение	1.20	99.31	99.84	99.57
Другие местоимения	3.65	98.89	98.68	98.78
Сокращение	0.35	96.69	82.23	88.88
Знак препинания	17.54	99.97	99.88	99.93
Остальное	4.66	84.68	79.35	81.93

А. Ю. Антонова, А. Н. Соловьев. Метод условных случайных полей в задачах обработки русскоязычных текстов. Диалог, 2013

## Задача распознавания именованных сущностей

Сущность всегда имеет категорию. Множество употребляемых категорий зависит от предметной области и жанра текста

- персона, организация, локация, дата-время
- профессия, должность, звание
- ссылка на нормативно-правовой акт
- артикул, изделие, производственный процесс
- заболевание, симптом, метод лечения,
- лекарственный препарат, химическое вещество
- биологический вид
- астрономический объект

---

Stanford Named Entity Recognizer:

<http://www-nlp.stanford.edu/software/CRF-NER.shtml>

## Задача разметки библиографических записей

Основные поля метаданных:

- автор(ы), название, издание, журнал, конференция,
- редактор, издательство, страна, город,
- страницы, номер, том, год, месяц,
- сайт, DOI, аннотация, ...

Проблема вариативности библиографических записей:

- David Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation. JMLR, 2003.
- D.Blei, A.Ng, M.Jordan. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. V.3. Pp.993–1022.
- Blei, David M. and Ng, Andrew Y. and Jordan, Michael I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research. JMLR.org. Vol.3, P.993–1022.

## Нотация BIOES (begin-inside-outside-end-single)

Для выделения групп слов используются метки с префиксами:

- B–Begin, I–Inside, O–outside — упрощённая BIO-нотация
- E–end, S–single

Пример задачи распознавания именованных сущностей:

Карл	Фридрих	Иероним	фон	Мюнхгаузен	родился	в	Боденвердере
B-PER	I-PER	I-PER	I-PER	E-PER	OUT	OUT	S-LOC

Пример задачи определения семантических ролей:

B_ACT	I_ACT	I_ACT	O	B_NUM_PER	O	B_LOC	I_LOC
Book	a	table	for	3	in	Domino's	pizza



## Пример инструмента разметки текста

The screenshot displays the doccano web application. On the left is a sidebar with a search bar and a list of documents. The main area shows a document with a text editor and a toolbar. The toolbar includes buttons for 'Person' (blue), 'Organization' (black), 'Other' (green), 'Location' (yellow), and 'Date' (red). The text in the editor is annotated with colored boxes corresponding to these categories. For example, 'Shinzō Abe' is highlighted in blue, 'Japanese' in green, '63rd' in red, and 'Prime Minister' in green. The text reads: 'Shinzō Abe is a Japanese politician serving as the 63rd and current Prime Minister of Japan and Leader of the Liberal Democratic Party (LDP) since 2012, previously being the 57th officeholder from 2006 to 2007. He is the third-longest serving Prime Minister in post-war Japan. [1] Abe comes from a politically prominent family and was first elected Prime Minister by a special session of the National Diet in September 2006. Then aged 52, he became Japan's youngest post-war Prime Minister and the first to have been born after World War II. Abe resigned on 12 September 2007 for health reasons. He was replaced by Yasuo Fukuda, the first in a

Text annotation for Human. <https://doccano.herokuapp.com>

## Линейная предсказательная модель разметки

Пусть  $D$  — множество размеченных последовательностей  $(x, y)$ ,  
 $x = (x_1, \dots, x_\ell)$  — последовательность объектов из  $X$ ,  
 $y = (y_1, \dots, y_\ell)$  — последовательность меток из  $Y$ .

**Например**, данные  $D$  — все предложения коллекции текстов;  
в предложении  $(x, y) \in D$  слову  $x_i$  соответствует метка  $y_i$ .

**Линейная модель** с параметром  $w \in \mathbb{R}^n$  предсказывает  
все метки  $y$  для последовательности  $x$  (structured prediction):

$$\langle w, F(x, y) \rangle = \sum_{j=1}^n w_j F_j(x, y)$$

Признаки  $F_j$  складываются из признаков отдельных объектов:

$$F_j(x, y) = \sum_{i=1}^{\ell} f_j(y_{i-1}, y_i, x, i), \quad j = 1, \dots, n$$

## Замечания о формировании признаков $f_j$

$f_j(y', y, x, i)$  — это информация о последовательности  $x$ , полезная для предсказания метки  $y_i = y$  в позиции  $i$ , когда в предыдущей позиции  $(i - 1)$  находится метка  $y_{i-1} = y'$ .

- $f_j(\bullet, i)$  может зависеть от всего  $x$ , не только от  $x_i$
- $f_j(\bullet, i)$  не может зависеть от других меток, кроме  $y_{i-1}$  (*марковское свойство*, упрощающее вывод, см. далее)
- часто используются бинарные  $f_j$ , но это не обязательно
- часто используются разреженные признаки
- число признаков  $n$  может достигать десятков тысяч
- если  $w_j = 0$ , то признак  $f_j$  не информативен
- последовательности  $(x, y)$  могут иметь любые длины  $\ell$ , но размерность  $F(x, y)$  фиксирована и равна  $n$

## Примеры признаков $f_j(y_{i-1}, y_i, x, i)$ для POS-теггинга

- $y_i = \text{ADVERB}$  и слово  $x_i$  оканчивается на «-ly»  
Если  $w_j > 0$ , то такие слова действительно часто оказываются наречиями
- $i = 1$  и  $y_i = \text{VERB}$  и предложение оканчивается знаком «?»  
Если  $w_j > 0$ , то первое слово в вопросительных предложениях действительно часто оказывается глаголом
- $y_{i-1} = \text{ADJECTIVE}$  и  $y_i = \text{NOUN}$   
Если  $w_j > 0$ , то существительные действительно часто следуют за прилагательным
- $y_i = \text{PREPOSITION}$  и  $y_{i-1} = \text{PREPOSITION}$   
Если  $w_j < 0$ , то перед предлогом действительно редко находится другой предлог

## Построение вероятностной линейной модели разметки

Аналог многоклассовой логистической регрессии:

$$p(y|x; w) = \frac{1}{Z(x, w)} \exp \langle w, F(x, y) \rangle, \quad y \in Y^\ell$$

где  $Z(x, w) = \sum_{y \in Y^\ell} \exp \langle w, F(x, y) \rangle$  — нормировочный множитель

**Принцип максимума правдоподобия:**

$$\sum_{(x,y) \in D} \ln p(y|x; w) \rightarrow \max_w$$

Оптимальная последовательность меток при известном  $w$ :

$$\hat{y} = \arg \max_{y \in Y^\ell} p(y|x; w)$$

Эффективное вычисление  $\arg \max$  и  $\sum$  по  $Y^\ell$  возможно благодаря марковскому свойству признаков  $f_j(y_{i-1}, y_i, x, i)$ .

## Вычисление оптимальной разметки

Оптимальная последовательность меток при известном  $w$ :

$$\begin{aligned}\hat{y} &= \arg \max_{y \in Y^\ell} \sum_{j=1}^n w_j F_j(x, y) = \arg \max_{y \in Y^\ell} \sum_{j=1}^n w_j \sum_{i=1}^{\ell} f_j(y_{i-1}, y_i, x, i) = \\ &= \arg \max_{y \in Y^\ell} \sum_{i=1}^{\ell} \sum_{j=1}^n w_j f_j(y_{i-1}, y_i, x, i)\end{aligned}$$

Определим  $Y \times Y$ -матрицы  $G_i[y', y]$ ,  $i = 1, \dots, \ell$ ,  $y', y \in Y$ :

$$G_i[y', y] = \sum_{j=1}^n w_j f_j(y', y, x, i)$$

Определим  $\ell \times Y$ -матрицу  $U[k, v]$ ,  $k = 1, \dots, \ell$ ,  $v \in Y$ :

$$U[k, v] = \max_{y_1 \dots y_{k-1}} \left( \sum_{i=1}^{k-1} G_i[y_{i-1}, y_i] + G_k[y_{k-1}, v] \right)$$

## Алгоритм Витерби (динамическое программирование)

**Прямой ход:** рекуррентное вычисление матрицы  $U$ :

$$U[0, v] := 0;$$

$$U[k, v] := \max_{u \in Y} (U[k-1, u] + G_k[u, v]), \quad k = 1, \dots, \ell, v \in Y.$$

**Обратный ход:** вычисление оптимальной разметки  $\hat{y} \in Y^\ell$ :

$$\hat{y}_\ell := \arg \max_{v \in Y} U[\ell, v];$$

$$\hat{y}_{k-1} := \arg \max_{u \in Y} (U[k-1, u] + G_k[u, \hat{y}_k]), \quad k = \ell, \dots, 2.$$

Алгоритм Витерби находит оптимальное решение:

- прямой ход:  $O(|Y|^2 \tilde{n} \ell)$ ,  $\tilde{n}$  — число ненулевых признаков
- обратный ход:  $O(|Y|^2 \ell)$

---

*Andrew Viterbi.* Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. 1967.

## Алгоритм SG (Stochastic Gradient) Роббинса–Монро

Максимизация логарифма правдоподобия:

$$\sum_{(x,y) \in D} \ln p(y|x; w) \rightarrow \max_w$$

**Идея ускорения сходимости:** обновлять вектор весов  $w$  после градиентного шага по каждому слагаемому

**Вход:** выборка  $D$ , темп обучения  $h$ ;

**Выход:** вектор весов  $w$ ;

инициализировать веса  $w_j$ ,  $j = 0, \dots, n$ ;

**повторять**

выбрать последовательность  $(x, y)$  из  $D$ ;

сделать градиентный шаг:  $w := w + h \nabla \ln p(y|x; w)$ ;

**пока** веса  $w$  не сойдутся;

---

*Robbins, H., Monro S. A stochastic approximation method // Annals of Mathematical Statistics, 1951, 22 (3), p. 400–407.*



## Вычисление градиента

Градиент одного слагаемого лог-правдоподобия по  $w$ :

$$\begin{aligned}\frac{\partial}{\partial w_j} \ln p(y|x; w) &= F_j(x, y) - \frac{\partial}{\partial w_j} \ln Z(x, w) = \\ &= F_j(x, y) - \frac{1}{Z(x, w)} \frac{\partial}{\partial w_j} Z(x, w); \\ \frac{\partial}{\partial w_j} Z(x, w) &= \frac{\partial}{\partial w_j} \sum_{y \in Y^\ell} \exp \sum_{k=1}^n w_k F_k(x, y) = \\ &= \sum_{y \in Y^\ell} F_j(x, y) \exp \sum_{k=1}^n w_k F_k(x, y); \\ \frac{\partial}{\partial w_j} \ln p(y|x; w) &= F_j(x, y) - \sum_{u \in Y^\ell} F_j(x, u) p(u|x; w).\end{aligned}$$

## Упрощение вычислений благодаря марковскому свойству

Подставим в градиент выражение  $F_j$  через  $f_j$ :

$$\begin{aligned}\sum_{u \in Y^\ell} p(u|x; w) F_j(x, u) &= \sum_{u \in Y^\ell} p(u|x; w) \sum_{i=1}^{\ell} f_j(u_{i-1}, u_i, x, i) = \\ &= \sum_{i=1}^{\ell} \sum_{u_{i-1} \in Y} \sum_{u_i \in Y} p(u_{i-1}, u_i|x; w) f_j(u_{i-1}, u_i, x, i).\end{aligned}$$

Осталось найти способ быстрого вычисления  $p(u_{i-1}, u_i|x; w)$ .

Вспомним выражение  $Z(x, w) = \sum_{y \in Y^\ell} \underbrace{\exp \sum_{i=1}^{\ell} G_i[y_{i-1}, y_i]}_{N(y)}$ .

Назовём  $N(y)$  *ненормированной вероятностью*  $y = (y_1, \dots, y_\ell)$ .

## Два семейства векторов: вперёд и назад

Определим векторы ненормированных вероятностей для начальных  $(y_1, \dots, y_k)$  и конечных  $(y_k, \dots, y_\ell)$  фрагментов.

Начальные фрагменты, завершающиеся меткой  $v$  в позиции  $k$ :

$$\alpha_k[v] = \sum_{y_1 \dots y_{k-1}} \exp\left(\sum_{i=1}^{k-1} G_i[y_{i-1}, y_i] + G_k[y_{k-1}, v]\right), \quad v \in Y$$

Конечные фрагменты, начинающиеся меткой  $v$  в позиции  $k$ :

$$\beta_k[u] = \sum_{y_{k+1} \dots y_\ell} \exp\left(G_{k+1}[u, y_{k+1}] + \sum_{i=k+2}^{\ell} G_i[y_{i-1}, y_i]\right), \quad v \in Y$$

Для них существуют эффективные рекуррентные формулы (аналогичные алгоритму Витерби, только  $\sum$  вместо  $\max$ )

## Рекуррентные формулы для вперёд-векторов и назад-векторов

**Вперёд-векторы** (forward vectors):

$$\alpha_k[v] = \sum_{u \in Y} \alpha_{k-1}[u] \exp G_k[u, v];$$
$$\alpha_0[v] = [v = \text{start}]$$

где  $y_0 = \text{start}$  — выделенная метка начала последовательности.

**Назад-векторы** (backward vectors):

$$\beta_k[u] = \sum_{v \in Y} \beta_{k+1}[v] \exp G_{k+1}[u, v];$$
$$\beta_{\ell+1}[u] = [u = \text{stop}]$$

где  $y_{\ell+1} = \text{stop}$  — выделенная метка конца последовательности.

## Полезные свойства вперёд-назад-векторов

Через  $\alpha_k[v]$ ,  $\beta_k[u]$  выражаются различные вероятности:

- $Z(x, w) = \sum_{v \in Y} \alpha_\ell[v]$
- $Z(x, w) = \sum_{u \in Y} \alpha_k[u] \beta_k[u]$  для любого  $k = 1, \dots, \ell$
- $p(y_k = u | x; w) = \frac{\alpha_k[u] \beta_k[u]}{Z(x, w)}$
- $p(y_k = u, y_{k+1} = v | x; w) = \frac{\alpha_k[u] \beta_{k+1}[v] \exp G_{k+1}[u, v]}{Z(x, w)}$

Отсюда получается выражение для градиента:

$$\frac{\partial \ln p(y|x; w)}{\partial w_j} = F_j(x, y) - \sum_{i=1}^{\ell} \sum_{y_{i-1} \in Y} \sum_{y_i \in Y} p(y_{i-1}, y_i | x; w) f_j(y_{i-1}, y_i, x, i)$$

## Максимизация регуляризованного правдоподобия

$L_2$ -регуляризация для уменьшения переобучения:

$$\sum_{(x,y) \in D} \ln p(y|x; w) + \tau \sum_{j=1}^n w_j^2 \rightarrow \max_w$$

$L_1$ -регуляризация с отбором признаков:

$$\sum_{(x,y) \in D} \ln p(y|x; w) + \gamma \sum_{j=1}^n |w_j| \rightarrow \max_w$$

ElasticNet с менее агрессивным отбором признаков:

$$\sum_{(x,y) \in D} \ln p(y|x; w) + \gamma \sum_{j=1}^n |w_j| + \tau \sum_{j=1}^n w_j^2 \rightarrow \max_w$$

## CRF с частичным обучением

Пусть наряду с  $D$  имеются неразмеченные данные  $U = \{u\}$ ,  
 $u = (u_1, \dots, u_\ell)$  — последовательность объектов  $u_i \in X$

Энтропийный регуляризатор:

$$\sum_{(x,y) \in D} \ln p(y|x; w) + \tau \sum_{u \in U} \sum_{y \in Y^\ell} p(y|u; w) \ln p(y|u; w) \rightarrow \max_w$$

Минимизация энтропии уменьшает неопределённость, распределения  $p(y|u; w)$  становятся сконцентрированными, менее похожими на равномерное распределение, повышается уверенность классификации неразмеченных  $u$ .

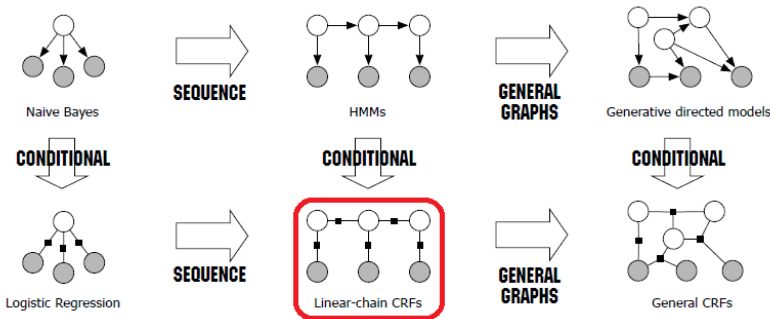
Вычисление градиента динамическим программированием так же эффективно, как для размеченных данных,  $O(|Y|^2 \tilde{n} \ell)$ .

---

*G. Mann, A. McCallum.* Efficient computation of entropy gradient for semi-supervised Conditional Random Fields. 2007.

## CRF — дискриминативная модель

CRF обобщает логистическую регрессию и скрытые марковские модели (Hidden Markov Model, HMM).



Генеративные модели:  $p(x, y; w)$

Дискриминативные модели:  $p(y|x; w)$ , не моделируется  $p(x)$

C.Sutton, A.McCallum. An introduction to Conditional Random Fields. 2011.



## Ещё несколько обобщений CRF

- HCRF: Hidden-state CRF

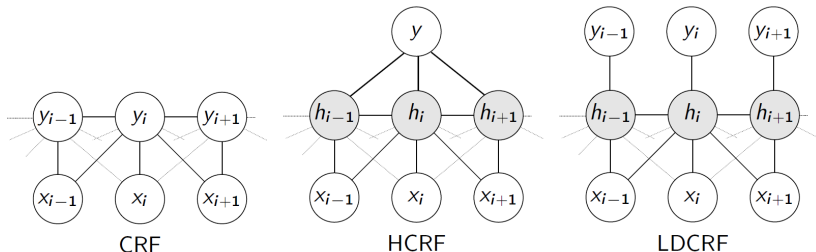
*Quattoni, Wang, Morency, Collins, Darrell.* Hidden conditional random fields. 2007.

- LDCRF: Latent-Dynamic CRF

*Sung, Jurafsky.* Hidden Conditional Random Fields for phone recognition. 2009.

- CCRF: Continuous CRF

*Qin, Liu.* Global ranking using continuous conditional random fields. 2008.





*Charles Elkan. (2012).*

*Log-linear models and Conditional Random Fields.*

— коротко и понятно объясняются все детали в формулах, 20 стр.



*Charles Sutton, Andrew McCallum. (2011).*

*An introduction to Conditional Random Fields.*

— прекрасный канонический обзор, но слишком детальный, 120 стр.



*John Lafferty, Andrew McCallum, Fernando Pereira. (2001).*

*Conditional Random Fields: probabilistic models for segmenting and labeling sequence data.*

— первая статья про CRF.