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 (МФТИ) / 2021»

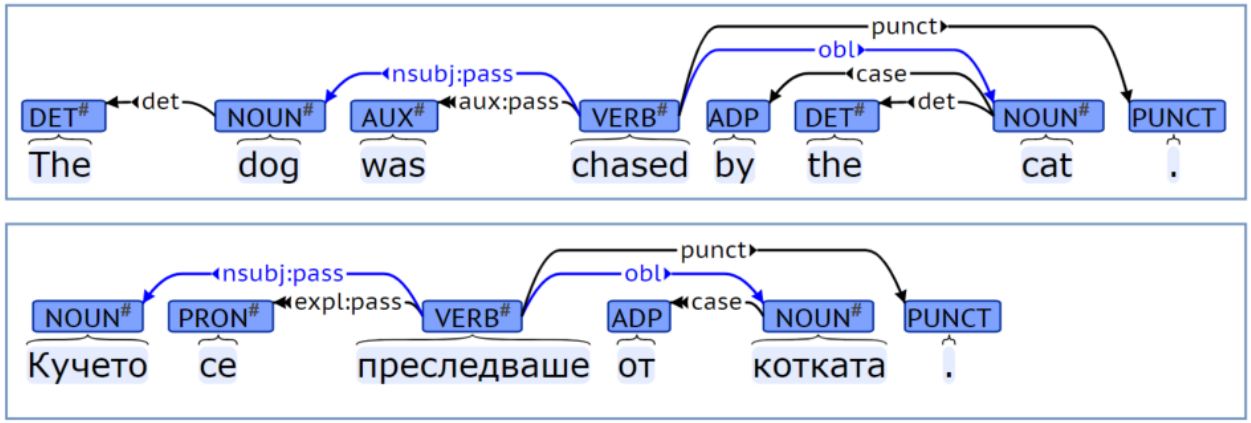
29 сентября 2021

Содержание

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

Разметка 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.

Разметка именованных сущностей (Named Entity Recognition)

Named entity — объект (сущность) реального мира, имеющий *наименование* и относящийся к определённой *категории*.

Примеры категорий:

- 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`

Разметка семантических ролей (semantic role labeling)

Задача: найти в предложении *актанты* — именные группы, обозначающие участников ситуации и их *семантические роли*

- агенс: одушевлённый инициатор и контролёр действия
- пациенс: участник, на которого направлено действие
- бенефактив: участник, получающий пользу или вред
- адресат: получатель сообщения (может быть бенефактивом)
- инструмент: посредством чего осуществляется действие
- экспериенцер: носитель чувств и восприятий
- стимул: источник восприятий
- источник: исходный пункт движения
- цель: конечный пункт движения

C.J.Fillmore. The Case for Case. Universals in Linguistic Theory. 1968.

D.Jurafsky, J.Martin. Speech and Language Processing. Chapter 20. 2019.

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

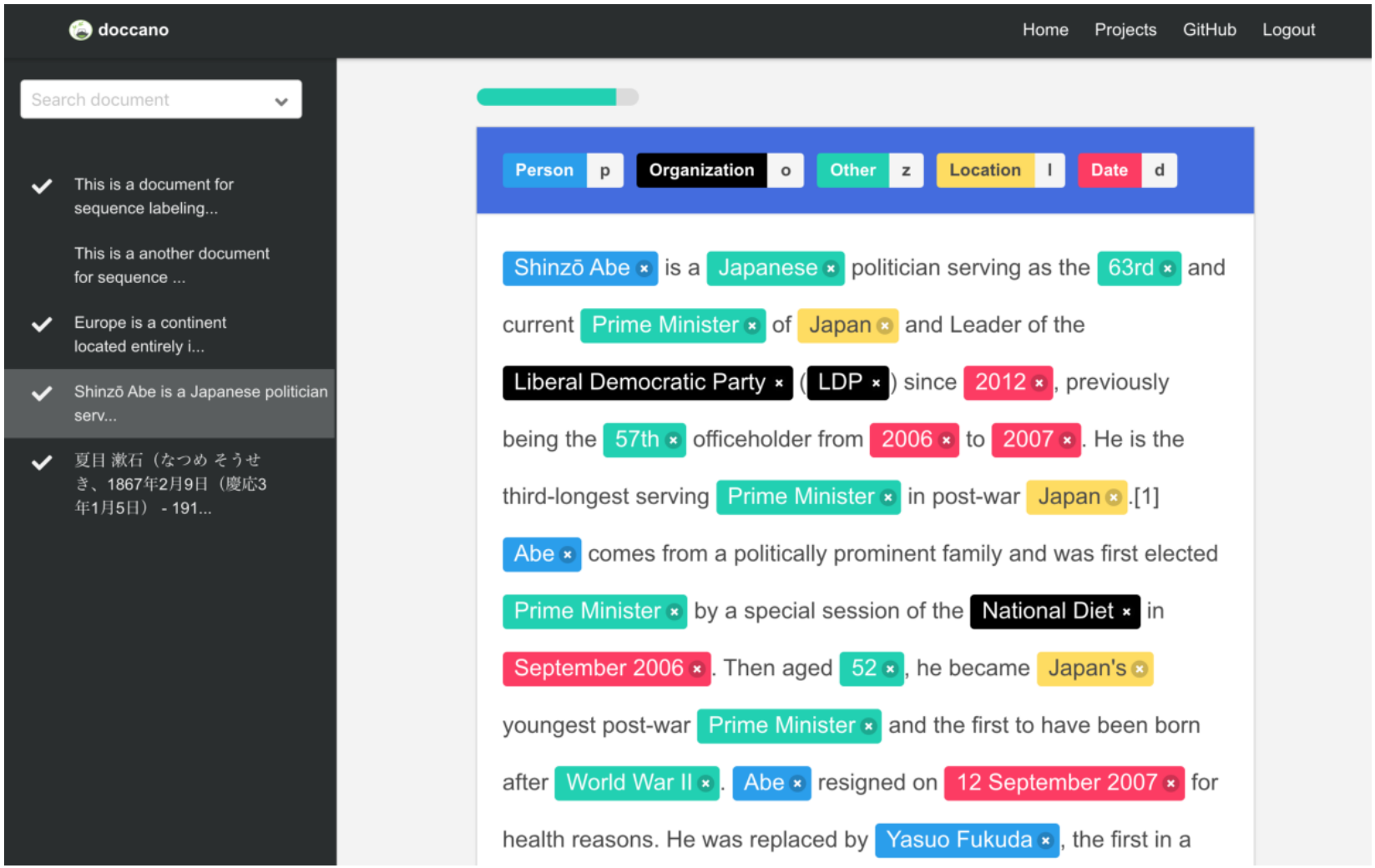
Пример задачи распознавания именованных сущностей:

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

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



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

Линейная предсказательная модель разметки

Пусть D — множество размеченных последовательностей (x, y),
$x = (x_1, \ldots, x_\ell)$ — последовательность объектов из X,
$y = (y_1, \ldots, y_\ell)$ — 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 $w \in \mathbb{R}^n$ оценивает целиком набор меток 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:

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

Формирование признаков f_j

Признак $f_j(u, v, x, i)$ — это некоторая полезная информация для предсказания условной вероятности $\mathrm{P}(y_i = v | y_{i-1} = u)$

- $f_j(u, v, x, i)$ может смотреть на весь x, не только на x_i
- $f_j(u, v, x, i)$ может смотреть только на метки y_{i-1}, y_i (*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
- при любой длине ℓ последовательности $(x, y) = (x_i, y_i)_{i=1}^{\ell}$ размерность $F(x, y)$ фиксирована и равна n

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

Признаки могут выражать наши гипотезы, от чего зависит y_i:

- $y_i = \text{ADVERB}$ 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 линейной вероятностной модели разметки

Аналог многоклассовой логистической регрессии:

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

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

Задача 1. Максимизация правдоподобия выборки D:

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

Задача 2. Оптимизация разметки y для x при известном w:

$$\ln p(y|x;w) \to \max_{y \in Y^\ell}$$

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

Вычисление оптимальной разметки (Задача 2)

Оптимизируемый критерий является *парно-сепарабельным* по y:

$$\ln p(y|x;w) + C = \sum_{j=1}^{n} w_j F_j(x,y) = \sum_{i=1}^{\ell} G_i[y_{i-1}, y_i] \to \max_{y \in Y^\ell},$$

где $G_i[u,v] = \sum_{j=1}^{n} w_j f_j(u,v,x,i)$ — матрицы $Y \times Y$, $i = 1,\ldots,\ell$.

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

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

Задача распадается на одномерные задачи по y_k, $k = \ell,\ldots,1$:

$$\sum_{i=1}^{\ell} G_i[y_{i-1}, y_i] = U[k, y_k] + G_{k+1}[y_k, y_{k+1}] + \cancel{\sum_{i=k+2}^{\ell} G_i[y_{i-1}, y_i]} \to \max_{y_k}$$

Алгоритм Ви́терби (динамическое программирование)

Прямой ход: рекуррентное вычисление матрицы U:

$$U[0,v] := 0;$$
$$U[k,v] := \max_{u \in Y} \big(U[k-1,u] + G_k[u,v]\big), \quad k = 1,\ldots,\ell,\ v \in Y.$$

Обратный ход: вычисление оптимальной разметки $y \in Y^\ell$:

$$y_\ell := \arg\max_{u \in Y} U[\ell,u];$$
$$y_k := \arg\max_{u \in Y} \big(U[k,u] + G_{k+1}[u,y_{k+1}]\big), \quad k = \ell-1,\ldots,1.$$

Алгоритм Ви́терби находит оптимальное решение (Viterbi path):

- 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.

Максимизация правдоподобия (Задача 1)

Оптимизация логарифма правдоподобия по вектору весов w

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

методом стохастического градиента (Stochastic Gradient, SG):
обновляем w после градиентного шага по каждому слагаемому

Вход: выборка D, темп обучения h;
Выход: вектор весов w;
инициализировать веса w_j, $j = 1, \ldots, n$;
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. 1951.

Вычисление градиента

Градиент одного слагаемого log-правдоподобия по w:

$$\frac{\partial}{\partial w_j} \ln p(y|x;w) = F_j(x,y) - \frac{\partial}{\partial w_j} \ln Z(x,w) =$$

$$= F_j(x,y) - \frac{1}{Z(x,w)} \frac{\partial}{\partial w_j} Z(x,w);$$

$$\frac{\partial}{\partial w_j} Z(x,w) = \frac{\partial}{\partial w_j} \sum_{u \in Y^\ell} \exp \sum_{k=1}^{n} w_k F_k(x,u) =$$

$$= \sum_{u \in Y^\ell} F_j(x,u) \exp \sum_{k=1}^{n} w_k F_k(x,u);$$

$$\frac{\partial}{\partial w_j} \ln p(y|x;w) = F_j(x,y) - \sum_{u \in Y^\ell} F_j(x,u) p(u|x;w).$$

Упрощение вычислений благодаря марковскому свойству

Подставим в градиент выражение F_j через f_j:

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

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

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

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

$N(u)$ — *ненормированная вероятность* $u = (u_1, \ldots, u_\ell) \in Y^\ell$.

Два семейства векторов: вперёд и назад

Определим векторы ненормированных вероятностей для начальных $(u_1, \ldots, u_k)$ и конечных $(v_k, \ldots, v_\ell)$ фрагментов.

Начальные фрагменты, завершающиеся меткой v в позиции k:

$$\alpha_k[v] = \sum_{u_1 \ldots u_{k-1}} \exp\Big(\sum_{i=1}^{k-1} G_i[u_{i-1}, u_i] + G_k[u_{k-1}, v]\Big), \qquad v \in Y$$

Конечные фрагменты, начинающиеся меткой u в позиции k:

$$\beta_k[u] = \sum_{v_{k+1} \ldots v_\ell} \exp\Big(G_{k+1}[u, v_{k+1}] + \sum_{i=k+2}^{\ell} G_i[v_{i-1}, v_i]\Big), \quad u \in Y$$

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), $v \in Y$:

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

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

Назад-векторы (backward vectors), $u \in Y$:

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

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:

- $Z(x,w) = \sum_{v \in Y} \alpha_\ell[v] = \sum_{u \in Y} \beta_1[u]$
- $Z(x,w) = \sum_{u \in Y} \alpha_k[u]\,\beta_k[u]$ для любого $k = 1, \ldots, \ell$
- $p(y_i = u | x; w) = \dfrac{\alpha_i[u]\,\beta_i[u]}{Z(x,w)}$
- $p(y_{i-1} = u, y_i = v | x; w) = \dfrac{\alpha_{i-1}[u]\,\beta_i[v]\,\exp G_i[u,v]}{Z(x,w)}$

Отсюда получается выражение для градиента:

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

Собираем всё воедино: основной цикл алгоритма SG

повторять

выбрать последовательность (x, y) из D;

$G_i[u, v] := \sum_{j=1}^{n} w_j f_j(u, v, x, i)$ для $i = 1..\ell,\ u, v \in Y$;

$\alpha_i[v] := \sum_{u \in Y} \alpha_{i-1}[u] \exp G_i[u, v]$ для $i = 1..\ell,\ v \in Y$;

$\beta_i[u] := \sum_{v \in Y} \beta_{i+1}[v] \exp G_{i+1}[u, v]$ для $i = \ell..1,\ u \in Y$;

$Z := \sum_{v \in Y} \alpha_\ell[v]$;

$p_i[u, v] := \frac{1}{Z} \alpha_{i-1}[u]\, \beta_i[v] \exp G_i[u, v]$ для $i = 1..\ell,\ u, v \in Y$;

$\nabla_j := F_j(x, y) - \sum_{i=1}^{\ell} \sum_{u,v \in Y} p_i[u, v] f_j(u, v, x, i)$ для $j = 1..n$;

градиентный шаг: $w := w(1 - \tau h) + h\nabla$;

пока веса w не сойдутся;

Максимизация регуляризованного правдоподобия

L_2-регуляризация для уменьшения переобучения:

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

L_1-регуляризация для отбора признаков с селективностью γ:

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

ElasticNet для менее агрессивного отбора признаков:

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

CRF — обобщение скрытых марковских моделей HMM

HMM (Hidden Markov Model) моделирует совместную плотность

$$p(x,y) = \prod_{i=1}^{\ell} p(x_i|y_i)\,p(y_i|y_{i-1}) = \exp \sum_{i=1}^{\ell} \underbrace{\ln p(x_i|y_i)}_{w_{x_i y_i}} + \underbrace{\ln p(y_i|y_{i-1})}_{w_{y_i y_{i-1}}} =$$

$$= \exp\Bigg(\sum_{i=1}^{\ell} \sum_{x\in X} \sum_{y\in Y} w_{xy} \underbrace{[y_i = y]\,[x_i = x]}_{f_{xy}} +$$

$$+ \sum_{y'\in Y} \sum_{y\in Y} w_{y'y} \underbrace{[y_{i-1} = y']\,[y_i = y]}_{f_{y'y}} \Bigg) =$$

$$= \frac{1}{Z} \exp\Bigg(\sum_{i=1}^{\ell} \sum_{j=1}^{n} w_j f_j(y_{i-1}, y_i, x, i) \Bigg)$$

CRF — обобщение скрытых марковских моделей HMM

HMM — генеративная модель совместной плотности

$$p(x,y) = \frac{1}{Z}\exp\left(\sum_{i=1}^{\ell}\sum_{j=1}^{n} w_j f_j(y_{i-1}, y_i, x, i)\right)$$

CRF — дискриминативная модель $p(y|x)$, обобщающая HMM:

- y_i зависит от всего x, а не только от x_i
- произвольные f_j, а не только индикаторы (и тогда $Z \neq 1$)
- произвольное число признаков n, а не $|X| \cdot |Y| + |Y|^2$
- произвольное множество X, а не только конечное
- для вывода $y_1, \ldots, y_\ell$ в HMM также используется Витерби
- для обучения в HMM чаще используется EM, чем SG

CRF с частичным обучением

Пусть наряду с D имеются неразмеченные данные $D' = \{x'\}$, $x' = (x'_1, \ldots, x'_\ell)$ — последовательность объектов $x'_i \in X$

Энтропийный регуляризатор:

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

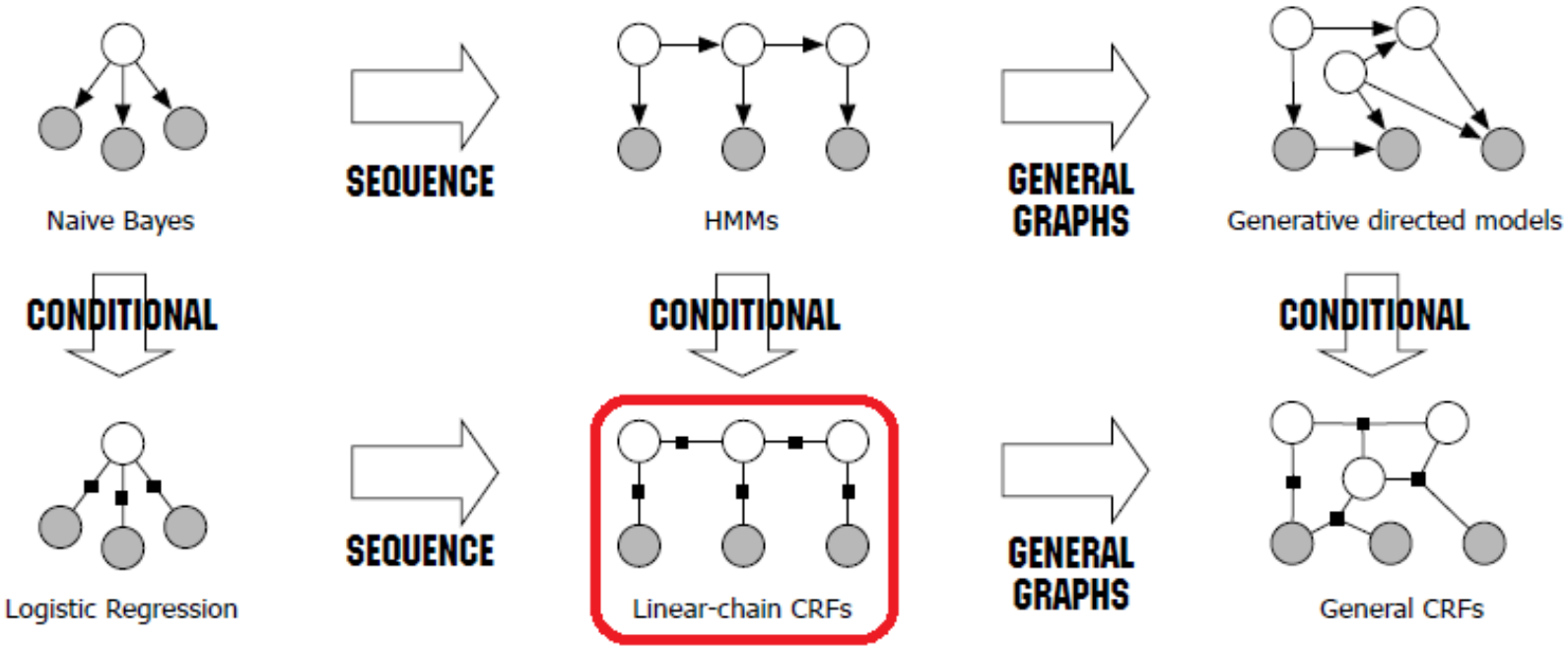
Минимизация энтропии уменьшает неопределённость, распределения $p(y|x';w)$ становятся сконцентрированными, менее похожими на равномерное распределение, повышается уверенность классификации неразмеченных x'.

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

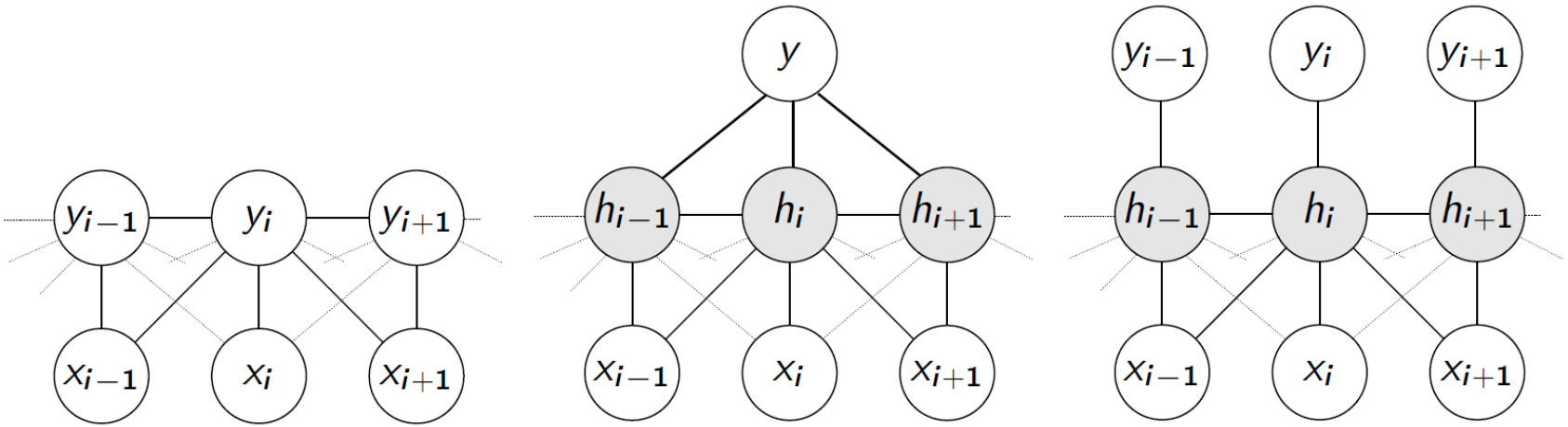
Quattoni, Wang, Morency, Collins, Darrell. Hidden conditional random fields. 2007.

- 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.



Литература

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.