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»

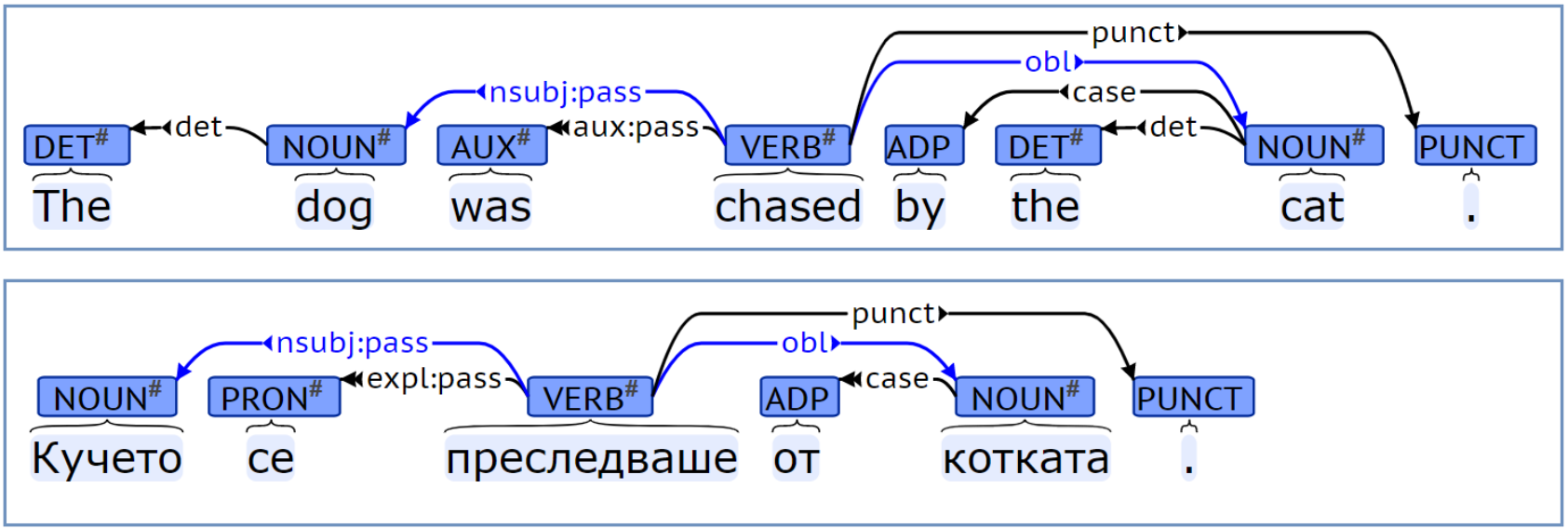
24 сентября 2019

Содержание

Примеры задач разметки и сегментации

- распознавание частей речи, членов предложения (part of speech tagging, POS)
- неглубокий синтаксический разбор (chunking, shallow syntax parsing)
- распознавание именованных сущностей (named entity recognition, NER)
- анализ тональности заданной сущности
- выделение текстовых полей данных (slot filling)
- выделение полей в библиографических записях
- сегментация научных или юридических текстов
- перевод речевого сигнала в текст
- перевод музыкального сигнала в нотную запись
- выделение генов в нуклеотидных последовательностях

Примеры частеречной и синтаксической разметки



http://universaldependencies.org/

Примеры тегов частей речи

Open class words	
ADJ	adjective
ADV	adverb
INTJ	interjection
NOUN	noun
PROPN	proper noun
VERB	verb

Other	
PUNCT	punctuation
SYM	symbol
X	other

Closed class words	
ADP	adposition
AUX	auxiliary verb
CCONJ	coordinating conjunction
DET	determiner
NUM	numeral
PART	particle
PRON	pronoun
SCONJ	subordinating conjunction

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
- название биологического вида
- заболевание, симптом, лекарственный препарат
- химическое вещество
- астрономический объект

Задача выделения полей в библиографических записях

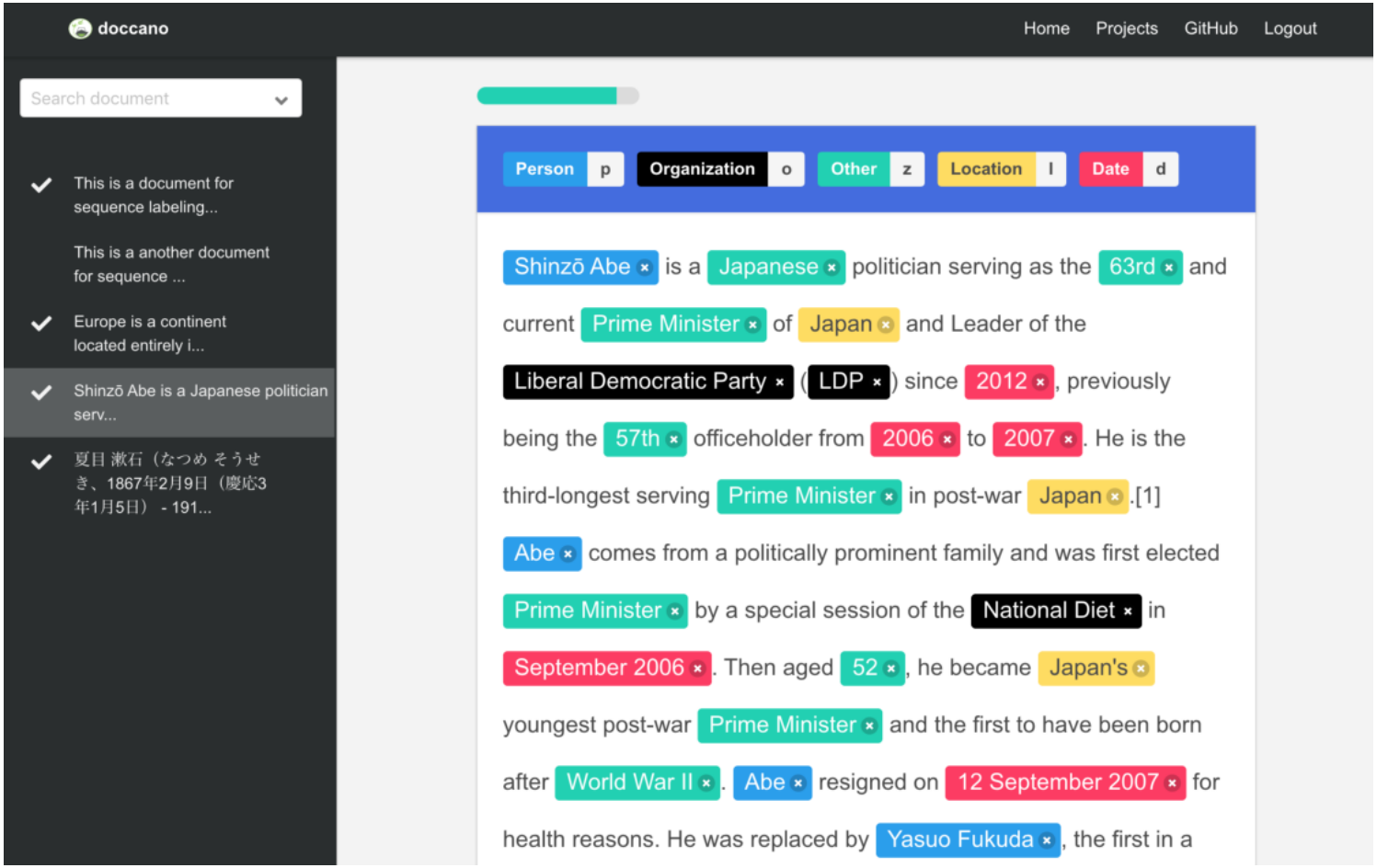
Разметка по полям метаданных:

- автор,
- название,
- месяц, год,
- издание,. . .

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.

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 $w \in \mathbb{R}^n$ 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:

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

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 ℓ, но размерность $F(x, y)$ фиксирована и равна n

Построение вероятностной линейной модели разметки

Аналог многоклассовой логистической регрессии:

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

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

Принцип максимума правдоподобия:

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

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^ℓ возможно благодаря марковскому свойству признаков $f_j(y_{i-1}, y_i, x, i)$.

Вычисление оптимальной разметки

Оптимальная последовательность меток при известном w:

$$\hat{y} = \arg\max_{y \in Y^\ell} p(y|x;w) = \arg\max_{y \in Y^\ell} \sum_{i=1}^{\ell} \underbrace{\sum_{j=1}^{n} w_j f_j(y_{i-1}, y_i, x, i)}_{G_i[y_{i-1}, y_i]}$$

Определим $\ell \times |Y|$-матрицу U:

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

Алгоритм Витерби. Рекуррентное вычисление U:

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

затем вычисление оптимальной $\hat{y}$ «обратным ходом»:

$$\hat{y}_\ell := \arg\max_v U[\ell, v];$$
$$\hat{y}_{k-1} := \arg\max_u \big(U[k-1,u] + G_k[u, \hat{y}_k] \big).$$

Свойства алгоритма Витерби

- алгоритм находит оптимальное решение
- прямой ход: за $O(|Y|^2 n\ell)$ операций
- обратный ход: за $O(|Y|^2 \ell)$ операций
- когда признаки сильно разрежены (преобладают нули), множитель n для прямого хода может сильно уменьшиться
- это алгоритм динамического программирования
- классический алгоритм Витерби предназначен для скрытых марковских моделей, данная версия совсем немного модифицирована

Алгоритм SG (Stochastic Gradient)

Максимизация логарифма правдоподобия:

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

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, $j = 0, \ldots, n$;
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 // Annals of Mathematical Statistics, 1951, 22 (3), p. 400–407.

Вычисление градиента

Градиент одного слагаемого log-правдоподобия по w:

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

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

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

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

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

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

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

Два семейства векторов: вперёд и назад

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

Начальные фрагменты, завершающиеся меткой v в позиции k:

$$\alpha_k[v] = \sum_{y_1 \cdots y_{k-1}} \exp\Big(\sum_{i=1}^{k-1} G_i[y_{i-1}, y_i] + G_k[y_{k-1}, v]\Big)$$

Конечные фрагменты, начинающиеся меткой v в позиции k:

$$\beta_k[u] = \sum_{y_{k+1} \cdots y_\ell} \exp\Big(G_{k+1}[u, y_{k+1}] + \sum_{i=k+2}^{\ell} G_i[y_{i-1}, y_i]\Big)$$

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 $k = 1, \ldots, \ell$
- $p(y_k = u | x; w) = \dfrac{\alpha_k[u]\,\beta_k[u]}{Z(x,w)}$
- $p(y_k = u, y_{k+1} = v | x; w) = \dfrac{\alpha_k[u]\,\beta_{k+1}[v]\,\exp G_{k+1}[u,v]}{Z(x,w)}$

Отсюда получается выражение для градиента:

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

Максимизация регуляризованного правдоподобия

L_2-регуляризация для уменьшения переобучения:

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

L_1-регуляризация с отбором признаков:

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

ElasticNet с менее агрессивным отбором признаков:

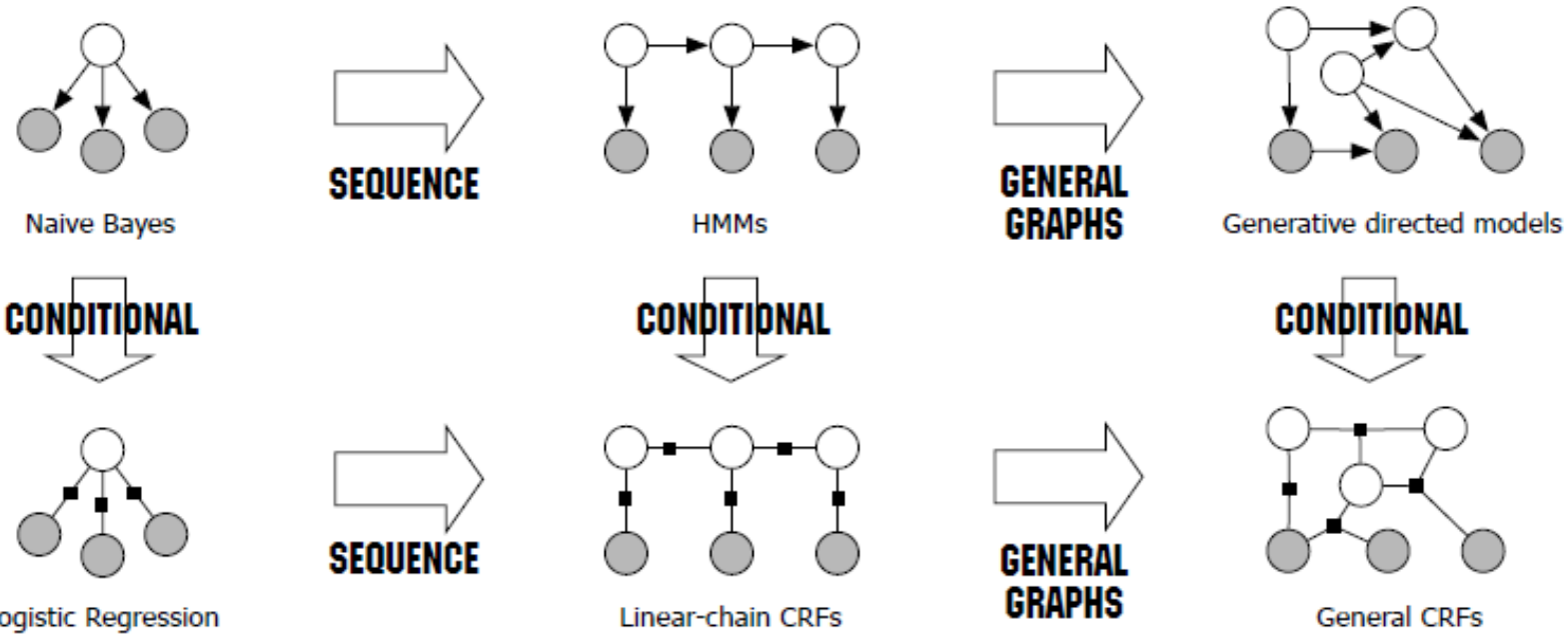
$$\sum_{(x,y)\in D} \ln p(y|x;w) + \gamma \sum_{j=1}^{n} |w_j| + \tau \sum_{j=1}^{n} w_j^2 \to \max_w$$

Примеры признаков для POS-теггинга

- $f_1(y_{i-1}, y_i, x, i) = 1$ если $y_i =$ ADVERB и слово x_i оканчивается на «-ly». Если $w_1 > 0$, то такие слова действительно часто оказываются наречиями.
- $f_2(y_{i-1}, y_i, x, i) = 1$ если $i = 1$, $y_i =$ VERB и предложение оканчивается знаком «?».
  Если $w_2 > 0$, то первое слово в вопросительных предложениях действительно часто оказывается глаголом.
- $f_3(y_{i-1}, y_i, x, i) = 1$ если $y_{i-1} =$ ADJECTIVE и $y_i =$ NOUN.
  Если $w_3 > 0$, то существительные действительно часто следуют за прилагательным.
- $f_4(y_{i-1}, y_i, x, i) = 1$ если $y_{i-1} = y_i =$PREPOSITION.
  Если $w_4 > 0$, то предлоги действительно крайне редко следуют друг за другом, поэтому второе слово вряд ли будет предлогом.

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.

Некоторые разновидности и обобщения CRF

- HCRF: Hidden-state CRF

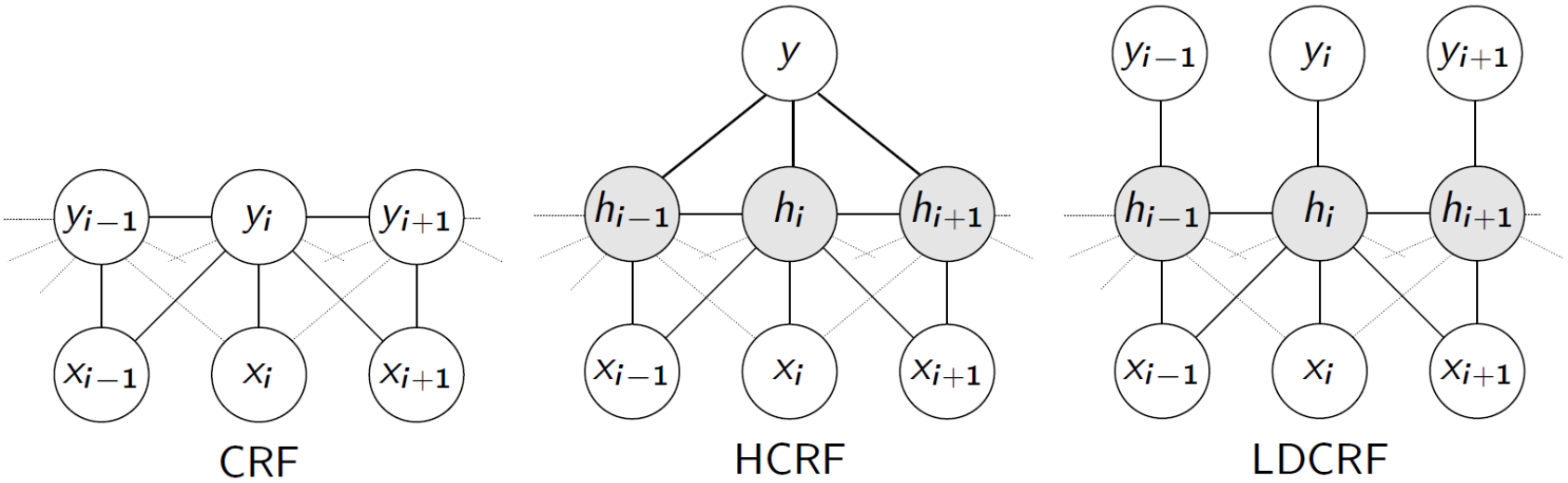
  Quattoni, Wang, Morency, Collins, Darrell. Hidden conditional random fields. 2007.

- LDCRF: Latent-Dynamic CRF

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

- CCRF: Continuous CRF

  Qin and Liu. Global ranking using continuous conditional random fields. 2008.



Литература


Charles Elkan. Log-linear models and Conditional Random Fields. 2012.


Charles Sutton, Andrew McCallum. An introduction to Conditional Random Fields. 2011.


John Lafferty, Andrew McCallum, Fernando Pereira. Conditional Random Fields: probabilistic models for segmenting and labeling sequence data. ICML, 2001.