Математические модели в морфологии

Условные случайные поля.

Алексей Андреевич Сорокин

курс по выбору, ОТИПЛ МГУ,
осенний семестр 2017–2018 учебного года
31 октября 2017 г.

Недостатки марковских моделей

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки.
- Сложнее выделить шаблоны согласования (вместо согласования по роду-числу-падежу — 3∗2∗6 отдельных согласований по каждому набору граммем).

Недостатки марковских моделей

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки.
- Сложнее выделить шаблоны согласования (вместо согласования по роду-числу-падежу — $3*2*6$ отдельных согласований по каждому набору граммем).
- Ограниченная память (чаще всего состояния — биграммы меток, триграмм уже слишком много).
- Следствие: не учитываются дистантные зависимости.

Недостатки марковских моделей

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки.
- Сложнее выделить шаблоны согласования (вместо согласования по роду-числу-падежу — $3*2*6$ отдельных согласований по каждому набору граммем).
- Ограниченная память (чаще всего состояния — биграммы меток, триграмм уже слишком много).
- Следствие: не учитываются дистантные зависимости.
- t_n зависит только от w_n, t_{n-1}, t_{n-2}, но не от w_{n-1}.

Недостатки марковских моделей

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки.
- Сложнее выделить шаблоны согласования (вместо согласования по роду-числу-падежу — 3∗2∗6 отдельных согласований по каждому набору граммем).
- Ограниченная память (чаще всего состояния — биграммы меток, триграмм уже слишком много).
- Следствие: не учитываются дистантные зависимости.
- t_n зависит только от w_n, t_{n-1}, t_{n-2}, но не от w_{n-1}.
- Однако лексемы влияют на морфологические показатели соседних:

обманути *друга* vs *соврати* *другу*
case=Gen case=Dat

Условные случайные поля: определение

- Пусть требуется найти последовательность скрытых состояний $\mathbf{q}_{1,n}$ по видимой последовательности $\mathbf{w}_{1,n}$.
- В морфологии q_i — морфологические метки (энграммы меток).
- По формуле условной вероятности:

$$p(q_1 \dots q_n|\mathbf{w}) = p(q_1|\mathbf{w})p(q_2|\mathbf{w}, q_1) \dots p(q_n|\mathbf{w}, q_1, \dots, q_{n-1})$$

Условные случайные поля: определение

- Пусть требуется найти последовательность скрытых состояний $\mathbf{q}_{1,n}$ по видимой последовательности $\mathbf{w}_{1,n}$.
- В морфологии q_i — морфологические метки (энграммы меток).
- По формуле условной вероятности:

$$p(q_1 \dots q_n|\mathbf{w}) = p(q_1|\mathbf{w})p(q_2|\mathbf{w}, q_1) \dots p(q_n|\mathbf{w}, q_1, \dots, q_{n-1})$$

- В условных случайных полях:

$$p(q_t|\mathbf{w}, q_1, \dots, q_{t-1}) \sim \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

Условные случайные поля: определение

- Пусть требуется найти последовательность скрытых состояний $\mathbf{q}_{1,n}$ по видимой последовательности $\mathbf{w}_{1,n}$.
- В морфологии q_i — морфологические метки (энграммы меток).
- По формуле условной вероятности:

$$p(q_1 \dots q_n|\mathbf{w}) = p(q_1|\mathbf{w})p(q_2|\mathbf{w}, q_1) \dots p(q_n|\mathbf{w}, q_1, \dots, q_{n-1})$$

- В условных случайных полях:

$$p(q_t|\mathbf{w}, q_1, \dots, q_{t-1}) \sim \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

- Локальная вероятность — функция текущего и предыдущего состояния, а также текущей позиции во входной последовательности.

Условные случайные поля: интуиция

- Условные случайные поля пытаются “угадать” следующее состояние по формуле:

$$p(q_t|\mathbf{w}, q_1, \ldots, q_{t-1}) \sim \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

Условные случайные поля: интуиция

- Условные случайные поля пытаются “угадать” следующее состояние по формуле:

$$p(q_t|\mathbf{w}, q_1, \ldots, q_{t-1}) \sim \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

- Выбор состояния q_t эквивалентен (q'_t — другое состояние):

$$\begin{aligned} p(q_t|\mathbf{w}, q_1, \ldots, q_{t-1}) &\geqslant p(q'_t|\mathbf{w}, q_1, \ldots, q_{t-1}) \\ \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right) &\geqslant \exp\left(\sum_k \theta_k f_k(q'_t, q_{t-1}, \mathbf{w}, t)\right) \\ \sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t) &\geqslant \sum_k \theta_k f_k(q'_t, q_{t-1}, \mathbf{w}, t) \\ \sum_k \theta_k (f_k(q_t, q_{t-1}, \mathbf{w}, t) &- f_k(q'_t, q_{t-1}, \mathbf{w}, t)) \geqslant 0 \end{aligned}$$

Условные случайные поля: интуиция

- Условные случайные поля пытаются “угадать” следующее состояние по формуле:

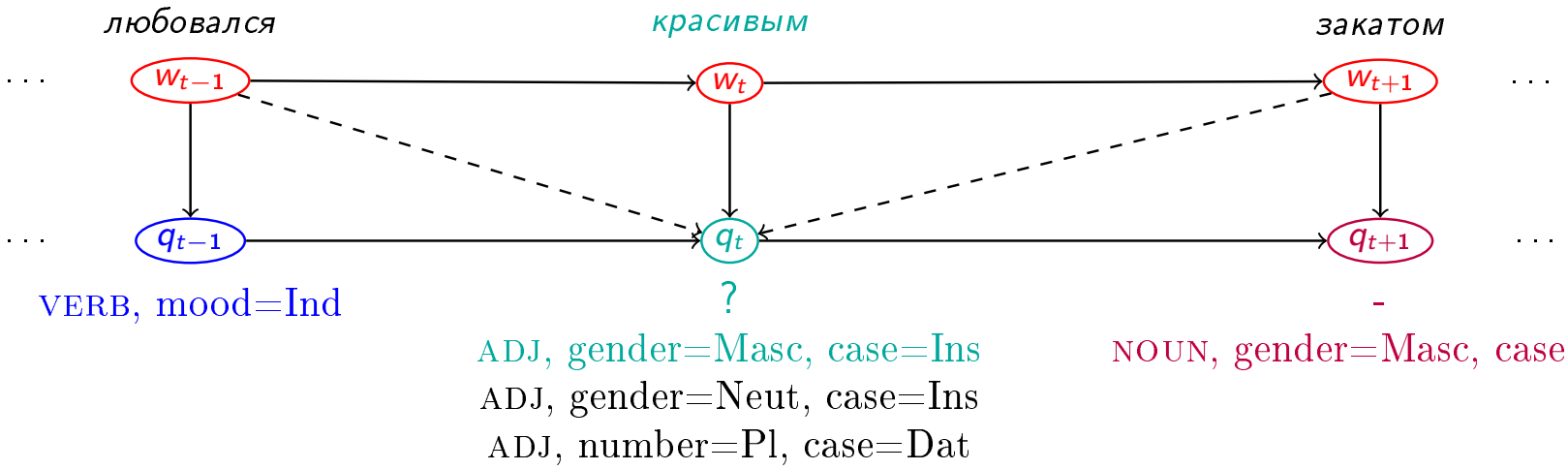
$$p(q_t|\mathbf{w}, q_1, \ldots, q_{t-1}) \sim \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

- Выбор состояния q_t эквивалентен (q'_t — другое состояние):

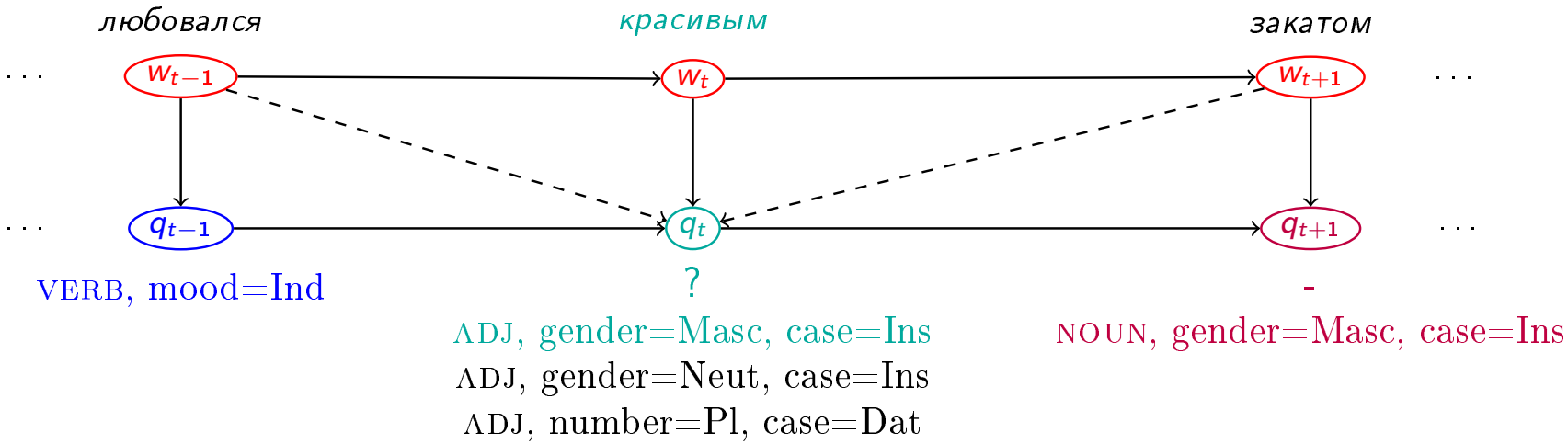
$$\begin{aligned} p(q_t|\mathbf{w}, q_1, \ldots, q_{t-1}) &\geqslant p(q'_t|\mathbf{w}, q_1, \ldots, q_{t-1}) \\ \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right) &\geqslant \exp\left(\sum_k \theta_k f_k(q'_t, q_{t-1}, \mathbf{w}, t)\right) \\ \sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t) &\geqslant \sum_k \theta_k f_k(q'_t, q_{t-1}, \mathbf{w}, t) \\ \sum_k \theta_k (f_k(q_t, q_{t-1}, \mathbf{w}, t) &- f_k(q'_t, q_{t-1}, \mathbf{w}, t)) \geqslant 0 \end{aligned}$$

- То есть следующее состояние выбирается линейным классификатором с весами θ_k.

Условные случайные поля: интуиция

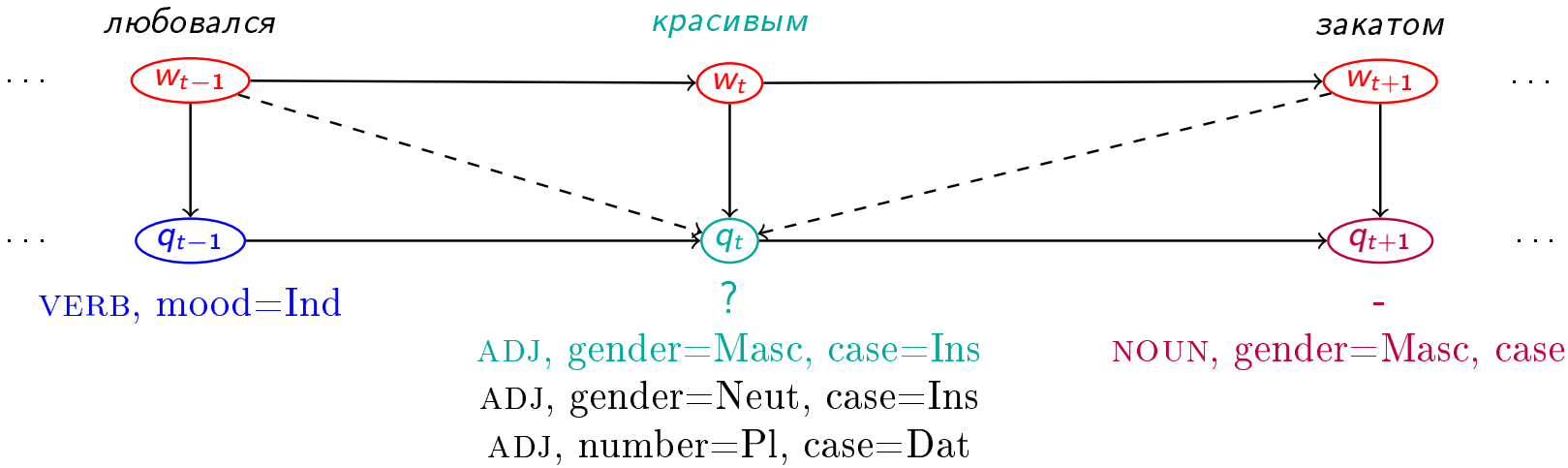


Условные случайные поля: интуиция



- Полезная информация:
  - $l_{t-1} =$ *любоваться* управляет творительным падежом.
  - Для w_{t+1} однозначный разбор NOUN, gender=Masc, для w_t однозначный разбор ADJ $\Leftarrow \text{gender}(w_t) = \text{Masc}$.
  - Для w_{t+1} однозначный разбор NOUN, gender=Masc, для w_t однозначный разбор ADJ $\Leftarrow \text{gender}(w_t) = \text{Masc}$.

Условные случайные поля: интуиция



- Полезная информация:
  - $l_{t-1} =$ *любоваться* управляет творительным падежом.
  - Для w_{t+1} однозначный разбор NOUN, gender=Masc, для w_t однозначный разбор ADJ $\Leftarrow \text{gender}(w_t) = \text{Masc}$.
  - Для w_{t+1} однозначный разбор NOUN, gender=Masc, для w_t однозначный разбор ADJ $\Leftarrow \text{gender}(w_t) = \text{Masc}$.
- Это нужно отразить в признаках.

Условные случайные поля: определение

- Условные случайные поля (conditional random fields, CRF)— модель вероятности $p(\mathbf{y}|\mathbf{w})$ вида:

$$p(\mathbf{q}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})} \prod_{t=1}^{n} \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

Условные случайные поля: определение

- Условные случайные поля (conditional random fields, CRF)— модель вероятности $p(\mathbf{y}|\mathbf{w})$ вида:

$$p(\mathbf{q}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})} \prod_{t=1}^{n} \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

- $Z(\mathbf{w})$ — нормировочная константа.

Условные случайные поля: определение

- Условные случайные поля (conditional random fields, CRF)— модель вероятности $p(\mathbf{y}|\mathbf{w})$ вида:

$$p(\mathbf{q}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})} \prod_{t=1}^{n} \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

- $Z(\mathbf{w})$ — нормировочная константа.
- $f_k(q_t, q_{t-1}, \mathbf{w}, t)$ — признаки.
- Например

$$f_{117}(q_t, q_{t-1}, \mathbf{w}, t) = \begin{cases} 1, & q_t \in \text{NOUN},\ q_{t-1} \in \text{ADJ}, \\ & \text{они согласованы по роду, числу и падежу.} \\ 0, & \text{иначе.} \end{cases}$$

$$f_{4568}(q_t, q_{t-1}, \mathbf{w}, t) = [\![w_t = \textit{что}, q_t \in \text{PRON}]\!]$$

Условные случайные поля: определение

- Условные случайные поля (conditional random fields, CRF)— модель вероятности $p(\mathbf{y}|\mathbf{w})$ вида:

$$p(\mathbf{q}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})} \prod_{t=1}^{n} \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

- $Z(\mathbf{w})$ — нормировочная константа.
- $f_k(q_t, q_{t-1}, \mathbf{w}, t)$ — признаки.
- Например

$$f_{117}(q_t, q_{t-1}, \mathbf{w}, t) = \begin{cases} 1, & q_t \in \text{NOUN},\ q_{t-1} \in \text{ADJ}, \\ & \text{они согласованы по роду, числу и падежу.} \\ 0, & \text{иначе.} \end{cases}$$

$$f_{4568}(q_t, q_{t-1}, \mathbf{w}, t) = [\![w_t = \textit{что}, q_t \in \text{PRON}]\!]$$

- θ_k — веса, настраиваемые по обучающей выборке.

Марковские модели как условные случайные поля

- Марковские модели – частный случай CRF:

$$p(\mathbf{q}|\mathbf{w}) = \prod_{t=1}^{n} a_{q_{t-1},q_t} b_{q_t,w_t}$$

- Представим $a_{q_{t-1},q_t} b_{q_t,w_t}$ в виде $\exp \sum_k \theta_k f_k(q_t, q_{t-1}, w_t)$.

Марковские модели как условные случайные поля

- Марковские модели – частный случай CRF:

$$p(\mathbf{q}|\mathbf{w}) = \prod_{t=1}^{n} a_{q_{t-1},q_t} b_{q_t,w_t}$$

- Представим $a_{q_{t-1},q_t} b_{q_t,w_t}$ в виде $\exp \sum_k \theta_k f_k(q_t, q_{t-1}, w_t)$.
- Обозначим $\alpha_{ij} = \log a_{ij}, \beta_{ij} = \log b_{ij}$.

Марковские модели как условные случайные поля

- Марковские модели – частный случай CRF:

$$p(\mathbf{q}|\mathbf{w}) = \prod_{t=1}^{n} a_{q_{t-1},q_t} b_{q_t,w_t}$$

- Представим $a_{q_{t-1},q_t} b_{q_t,w_t}$ в виде $\exp \sum_k \theta_k f_k(q_t, q_{t-1}, w_t)$.
- Обозначим $\alpha_{ij} = \log a_{ij}, \beta_{ij} = \log b_{ij}$.

$$a_{q_{t-1},q_t} b_{q_t,w_t} = \exp\left(\alpha_{q_{t-1},q_t} + \beta_{q_t,w_t}\right)$$

Марковские модели как условные случайные поля

- Марковские модели – частный случай CRF:

$$p(\mathbf{q}|\mathbf{w}) = \prod_{t=1}^{n} a_{q_{t-1},q_t} b_{q_t,w_t}$$

- Представим $a_{q_{t-1},q_t} b_{q_t,w_t}$ в виде $\exp \sum_k \theta_k f_k(q_t, q_{t-1}, w_t)$.
- Обозначим $\alpha_{ij} = \log a_{ij}, \beta_{ij} = \log b_{ij}$.

$$\begin{aligned} a_{q_{t-1},q_t} b_{q_t,w_t} &= \exp\left(\alpha_{q_{t-1},q_t} + \beta_{q_t,w_t}\right) \\ &= \exp\left(\sum_{i,j} \alpha_{ij} [\![q_{t-1} = i, q_t = j]\!] + \sum_{j,k} \beta_{jk} [\![q_t = j, w_t = k]\!]\right) \end{aligned}$$

Марковские модели как условные случайные поля

- Марковские модели – частный случай CRF:

$$p(\mathbf{q}|\mathbf{w}) = \prod_{t=1}^{n} a_{q_{t-1},q_t} b_{q_t,w_t}$$

- Представим $a_{q_{t-1},q_t} b_{q_t,w_t}$ в виде $\exp \sum_k \theta_k f_k(q_t, q_{t-1}, w_t)$.
- Обозначим $\alpha_{ij} = \log a_{ij}, \beta_{ij} = \log b_{ij}$.

$$\begin{aligned} a_{q_{t-1},q_t} b_{q_t,w_t} &= \exp\left(\alpha_{q_{t-1},q_t} + \beta_{q_t,w_t}\right) \\ &= \exp\left(\sum_{i,j} \alpha_{ij} [\![q_{t-1} = i, q_t = j]\!] + \sum_{j,k} \beta_{jk} [\![q_t = j, w_t = k]\!]\right) \\ &= \exp\left(\sum_{i,j} \alpha_{ij} f_{ij}(q_{t-1}, q_t) + \sum_{j,k} \beta_{jk} \beta_{jk} g_{jk}(q_t, w_t)\right) \end{aligned}$$

- Здесь $f_{ij}(q_{t-1}, q_t) = [\![q_{t-1} = i, q_t = j]\!]$ — индикатор биграммы состояний.
- $g_{jk}(q_t, w_t)$ — индикатор выбора метки q_t для w_t.

Марковские модели как условные случайные поля

- Марковские модели – частный случай CRF:

$$p(\mathbf{q}|\mathbf{w}) = \prod_{t=1}^{n} a_{q_{t-1},q_t} b_{q_t,w_t}$$

- Представим $a_{q_{t-1},q_t} b_{q_t,w_t}$ в виде $\exp \sum_k \theta_k f_k(q_t, q_{t-1}, w_t)$.
- Обозначим $\alpha_{ij} = \log a_{ij}, \beta_{ij} = \log b_{ij}$.

$$\begin{aligned} a_{q_{t-1},q_t} b_{q_t,w_t} &= \exp\left(\alpha_{q_{t-1},q_t} + \beta_{q_t,w_t}\right) \\ &= \exp\left(\sum_{i,j} \alpha_{ij} [\![q_{t-1} = i, q_t = j]\!] + \sum_{j,k} \beta_{jk} [\![q_t = j, w_t = k]\!]\right) \\ &= \exp\left(\sum_{i,j} \alpha_{ij} f_{ij}(q_{t-1}, q_t) + \sum_{j,k} \beta_{jk} \beta_{jk} g_{jk}(q_t, w_t)\right) \end{aligned}$$

- Здесь $f_{ij}(q_{t-1}, q_t) = [\![q_{t-1} = i, q_t = j]\!]$ — индикатор биграммы состояний.
- $g_{jk}(q_t, w_t)$ — индикатор выбора метки q_t для w_t.
- В произвольной CRF возможны признаки более сложной природы.

Признаки в условных случайных полях

- Слово в текущей позиции и соседние слова (чаще всего $w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$).

Признаки в условных случайных полях

- Слово в текущей позиции и соседние слова (чаще всего $w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$).
- Текущая и предыдущая морфологическая метка q_t, q_{t-1}.

Признаки в условных случайных полях

- Слово в текущей позиции и соседние слова (чаще всего $w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$).
- Текущая и предыдущая морфологическая метка q_t, q_{t-1}.
- Отдельные признаки морфологических меток (например, падеж метки q_t).

Признаки в условных случайных полях

- Слово в текущей позиции и соседние слова (чаще всего $w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$).
- Текущая и предыдущая морфологическая метка q_t, q_{t-1}.
- Отдельные признаки морфологических меток (например, падеж метки q_t).
- Комбинации признаков морфологических меток (например, $q_{t-1} \in \text{ADJ}, q_t \in \text{NOUN}, \text{CASE}(q_{t-1}) = \text{CASE}(q_t)$).

Признаки в условных случайных полях

- Слово в текущей позиции и соседние слова (чаще всего $w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$).
- Текущая и предыдущая морфологическая метка q_t, q_{t-1}.
- Отдельные признаки морфологических меток (например, падеж метки q_t).
- Комбинации признаков морфологических меток (например, $q_{t-1} \in \text{ADJ}, q_t \in \text{NOUN}, \text{CASE}(q_{t-1}) = \text{CASE}(q_t)$).
- Комбинации лексических и морфологических признаков (например, $[\![w_t = \textit{что}, q_t \in \text{PRON}]\!]$).

Признаки в условных случайных полях

- Слово в текущей позиции и соседние слова (чаще всего $w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$).
- Текущая и предыдущая морфологическая метка q_t, q_{t-1}.
- Отдельные признаки морфологических меток (например, падеж метки q_t).
- Комбинации признаков морфологических меток (например, $q_{t-1} \in \text{ADJ}, q_t \in \text{NOUN}, \text{CASE}(q_{t-1}) = \text{CASE}(q_t)$).
- Комбинации лексических и морфологических признаков (например, $[\![w_t = \textit{что}, q_t \in \text{PRON}]\!]$).
- Признаки слова w_t (капитализация, суффиксы, принадлежность к замкнутым классам...).

Признаки в условных случайных полях

- Слово в текущей позиции и соседние слова (чаще всего $w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}$).
- Текущая и предыдущая морфологическая метка q_t, q_{t-1}.
- Отдельные признаки морфологических меток (например, падеж метки q_t).
- Комбинации признаков морфологических меток (например, $q_{t-1} \in \text{ADJ}, q_t \in \text{NOUN}, \text{CASE}(q_{t-1}) = \text{CASE}(q_t)$).
- Комбинации лексических и морфологических признаков (например, $[\![w_t = \textit{что}, q_t \in \text{PRON}]\!]$).
- Признаки слова w_t (капитализация, суффиксы, принадлежность к замкнутым классам...).
- Признаки в основном бинарные.
- Для категориальных признаков (текущее слово) — 0/1 кодирование (для словаря размера D — D бинарных признаков).

Глобальные признаки

- Формулу:

$$p(\mathbf{q}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})} \prod_{t=1}^{n} \exp\left(\sum_{k} \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

Глобальные признаки

- Формулу:

$$p(\mathbf{q}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})} \prod_{t=1}^{n} \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

- Можно переписать в виде:

$$\begin{aligned} \log p(\mathbf{q}|\mathbf{w}) &= -\log Z(\mathbf{w}) + \sum_{k}^{K} \theta_k F_k(\mathbf{q}, \mathbf{w}) \\ F_k(\mathbf{q}, \mathbf{w}) &= \sum_{t=1}^{n} f_k(q_t, q_{t-1}, \mathbf{w}, t) \end{aligned}$$

Глобальные признаки

- Формулу:

$$p(\mathbf{q}|\mathbf{w}) = \frac{1}{Z(\mathbf{w})} \prod_{t=1}^{n} \exp\left(\sum_k \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)\right)$$

- Можно переписать в виде:

$$\begin{aligned} \log p(\mathbf{q}|\mathbf{w}) &= -\log Z(\mathbf{w}) + \sum_k^K \theta_k F_k(\mathbf{q}, \mathbf{w}) \\ F_k(\mathbf{q}, \mathbf{w}) &= \sum_{t=1}^{n} f_k(q_t, q_{t-1}, \mathbf{w}, t) \end{aligned}$$

- F_k — “глобальная версия” признака f_k.
- Например, если $f_k(q_t, q_{t-1}, \mathbf{w}, t) = [\![q_t \in \mathrm{NOUN}]\!]$, то F_k считает число существительных.

Обучение условных случайных полей

- Как определить параметры θ_k в

$$\log p(\mathbf{q}|\mathbf{w}) = -\log Z(\mathbf{w}) + \sum_{k}^{n} \theta_k F_k(\mathbf{q}, \mathbf{w})$$

Обучение условных случайных полей

- Как определить параметры θ_k в

$$\log p(\mathbf{q}|\mathbf{w}) = -\log Z(\mathbf{w}) + \sum_k^n \theta_k F_k(\mathbf{q}, \mathbf{w})$$

- Основная идея: при обучении правильный разбор $\mathbf{q}$ должен ранжироваться выше, чем неправильный $\mathbf{q}'$:

$$\begin{aligned} p(\mathbf{q}|\mathbf{w}) &\geqslant p(\mathbf{q}'|\mathbf{w}) \\ \sum_k^n \theta_k F_k(\mathbf{q}, \mathbf{w}) &\geqslant \sum_k^n \theta_k F_k(\mathbf{q}', \mathbf{w}) \end{aligned}$$

Обучение условных случайных полей

- Как определить параметры θ_k в

$$\log p(\mathbf{q}|\mathbf{w}) = -\log Z(\mathbf{w}) + \sum_k^n \theta_k F_k(\mathbf{q}, \mathbf{w})$$

- Основная идея: при обучении правильный разбор $\mathbf{q}$ должен ранжироваться выше, чем неправильный $\mathbf{q}'$:

$$\begin{aligned} p(\mathbf{q}|\mathbf{w}) &\geqslant p(\mathbf{q}'|\mathbf{w}) \\ \sum_k^n \theta_k F_k(\mathbf{q}, \mathbf{w}) &\geqslant \sum_k^n \theta_k F_k(\mathbf{q}', \mathbf{w}) \end{aligned}$$

- Это равносильно

$$\sum_k^n \theta_k (F_k(\mathbf{q}, \mathbf{w}) - F_k(\mathbf{q}', \mathbf{w})) \geqslant 0$$

Сведение к линейному классификатору

- Если рассматривать θ_k как веса классификатора, то условие

$$\sum_k^n \theta_k (F_k(\mathbf{q}, \mathbf{w}) - F_k(\mathbf{q}', \mathbf{w})) \geqslant 0$$

равносильно тому,что вектор $(\mathbf{F}(\mathbf{q}, \mathbf{w}) - \mathbf{F}'(\mathbf{q}', \mathbf{w}))$ относится к положительному классу.

Сведение к линейному классификатору

- Если рассматривать θ_k как веса классификатора, то условие

$$\sum_k^n \theta_k (F_k(\mathbf{q}, \mathbf{w}) - F_k(\mathbf{q}', \mathbf{w})) \geqslant 0$$

  равносильно тому,что вектор $(\mathbf{F}(\mathbf{q}, \mathbf{w}) - \mathbf{F}'(\mathbf{q}', \mathbf{w}))$ относится к положительному классу.
- Линейный классификатор можно обучать персептроном.

Сведение к линейному классификатору

- Если рассматривать θ_k как веса классификатора, то условие

$$\sum_k^n \theta_k (F_k(\mathbf{q}, \mathbf{w}) - F_k(\mathbf{q}', \mathbf{w})) \geqslant 0$$

  равносильно тому,что вектор $(\mathbf{F}(\mathbf{q}, \mathbf{w}) - \mathbf{F}'(\mathbf{q}', \mathbf{w}))$ относится к положительному классу.
- Линейный классификатор можно обучать персептроном.
- Алгоритм обновления при ошибке:

$$\begin{aligned} \theta' &= \theta + \eta \cdot (\mathbf{F}(\mathbf{q}, \mathbf{w}) - \mathbf{F}'(\mathbf{q}', \mathbf{w})) \\ \eta > 0 &— \text{темп обучения} \end{aligned}$$

Сведение к линейному классификатору

- Если рассматривать θ_k как веса классификатора, то условие

$$\sum_k^n \theta_k (F_k(\mathbf{q}, \mathbf{w}) - F_k(\mathbf{q}', \mathbf{w})) \geqslant 0$$

  равносильно тому,что вектор $(\mathbf{F}(\mathbf{q}, \mathbf{w}) - \mathbf{F}'(\mathbf{q}', \mathbf{w}))$ относится к положительному классу.
- Линейный классификатор можно обучать персептроном.
- Алгоритм обновления при ошибке:

$$\begin{aligned} \theta' &= \theta + \eta \cdot (\mathbf{F}(\mathbf{q}, \mathbf{w}) - \mathbf{F}'(\mathbf{q}', \mathbf{w})) \\ \eta > 0 &— \text{темп обучения} \end{aligned}$$

- Идея: если $F_k(\mathbf{q}, \mathbf{w}) > \mathbf{F}'_k(\mathbf{q}', \mathbf{w}))$, то F_k — хороший признак и надо увеличить его вес.

Персептрон для обучения CRF

Алгоритм обучения условных случайных полей.

- Инициализировать веса $\theta_k = 0,\ k = 1, \ldots, K$.

Персептрон для обучения CRF

Алгоритм обучения условных случайных полей.

- Инициализировать веса $\theta_k = 0,\ k = 1, \ldots, K$.
- В течение заданного числа итераций T:
  - Для всех пар $\mathbf{w}, \mathbf{q}$ из обучающей выборки:

Персептрон для обучения CRF

Алгоритм обучения условных случайных полей.

- Инициализировать веса $\theta_k = 0,\ k = 1, \ldots, K$.
- В течение заданного числа итераций T:
  - Для всех пар $\mathbf{w}, \mathbf{q}$ из обучающей выборки:
    - Найти наилучшую последовательность $\mathbf{q}'$ с точки зрения текущей модели.

Персептрон для обучения CRF

Алгоритм обучения условных случайных полей.

- Инициализировать веса $\theta_k = 0,\ k = 1, \ldots, K$.
- В течение заданного числа итераций T:
  - Для всех пар $\mathbf{w}, \mathbf{q}$ из обучающей выборки:
    - Найти наилучшую последовательность $\mathbf{q}'$ с точки зрения текущей модели.
    - Если $\mathbf{q}' \neq \mathbf{q}$, обновить веса:

$$\theta_k \mathrel{+}= \eta \cdot (F_k(\mathbf{q}, \mathbf{w}) - F_k(\mathbf{q}', \mathbf{w}))$$

Персептрон для обучения CRF

- Персептрон склонен к переобучению.
- Приёмы для улучшения сходимости персептрона:
  - Брать среднее значение весов со всех итераций, а не финальное.

Персептрон для обучения CRF

- Персептрон склонен к переобучению.
- Приёмы для улучшения сходимости персептрона:
  - Брать среднее значение весов со всех итераций, а не финальное.
  - Требовать “превосходства” правильной гипотезы с некоторым отступом:

$$\sum_k^n \theta_k F_k(\mathbf{q}, \mathbf{w}) \geqslant \sum_k^n \theta_k F_k(\mathbf{q}', \mathbf{w}) + \delta$$

Персептрон для обучения CRF

- Персептрон склонен к переобучению.
- Приёмы для улучшения сходимости персептрона:
  - Брать среднее значение весов со всех итераций, а не финальное.
  - Требовать “превосходства” правильной гипотезы с некоторым отступом:

$$\sum_k^n \theta_k F_k(\mathbf{q}, \mathbf{w}) \geqslant \sum_k^n \theta_k F_k(\mathbf{q}', \mathbf{w}) + \delta$$

- Также часто вначале определяют “грубые” метки (части речи), а потом их уточняют.

Определение наилучшей последовательности

- Морфологический разбор сводится к нахождению наилучшей последовательности состояний.
- Состояния нельзя находить полным перебором.

Определение наилучшей последовательности

- Морфологический разбор сводится к нахождению наилучшей последовательности состояний.
- Состояния нельзя находить полным перебором.
- Логарифм вероятности выходной последовательности:

$$\log p(\mathbf{q}|\mathbf{w}) = -\log Z(\mathbf{w}) + \sum_{t=1}^{n} \sum_{k} \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)$$

- Логарифм вероятности начала последовательности:

$$\log p(\mathbf{q}_{1,m}|\mathbf{w}) = -\log Z(\mathbf{w}) + \sum_{t=1}^{m} \sum_{k} \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)$$

Определение наилучшей последовательности

- Морфологический разбор сводится к нахождению наилучшей последовательности состояний.
- Состояния нельзя находить полным перебором.
- Логарифм вероятности выходной последовательности:

$$\log p(\mathbf{q}|\mathbf{w}) = -\log Z(\mathbf{w}) + \sum_{t=1}^{n} \sum_{k} \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)$$

- Логарифм вероятности начала последовательности:

$$\log p(\mathbf{q}_{1,m}|\mathbf{w}) = -\log Z(\mathbf{w}) + \sum_{t=1}^{m} \sum_{k} \theta_k f_k(q_t, q_{t-1}, \mathbf{w}, t)$$

- Формула пересчёта вероятностей:

$$\log p(\mathbf{q}_{1,m}|\mathbf{w}) = \log p(\mathbf{q}_{1,m-1}|\mathbf{w}) + \sum_{k} \theta_k f_k(q_m, q_{m-1}, \mathbf{w}, m)$$

Пересчёт вероятностей

- Формула пересчёта вероятностей:

$$\log p(\mathbf{q}_{1,m}|\mathbf{w}) = \log p(\mathbf{q}_{1,m-1}|\mathbf{w}) + \sum_k \theta_k f_k(q_m, q_{m-1}, \mathbf{w}, m)$$

Пересчёт вероятностей

- Формула пересчёта вероятностей:

$$\log p(\mathbf{q}_{1,m}|\mathbf{w}) = \log p(\mathbf{q}_{1,m-1}|\mathbf{w}) + \sum_k \theta_k f_k(q_m, q_{m-1}, \mathbf{w}, m)$$

- $L(q_{m-1}, q_m, m)$ — штраф за переход из q_{m-1} в q_m в момент m.
- Штраф не зависит от предыдущего пути (только от последнего состояния).

Пересчёт вероятностей

- Формула пересчёта вероятностей:

$$\log p(\mathbf{q}_{1,m}|\mathbf{w}) = \log p(\mathbf{q}_{1,m-1}|\mathbf{w}) + \sum_k \theta_k f_k(q_m, q_{m-1}, \mathbf{w}, m)$$

- $L(q_{m-1}, q_m, m)$ — штраф за переход из q_{m-1} в q_m в момент m.
- Штраф не зависит от предыдущего пути (только от последнего состояния).
- Можно запомнить последнее состояние и применить алгоритм Витерби (почти так же, как в марковских моделях).

Пересчёт вероятностей

- Формула пересчёта вероятностей:

$$\log p(\mathbf{q}_{1,m}|\mathbf{w}) = \log p(\mathbf{q}_{1,m-1}|\mathbf{w}) + \sum_k \theta_k f_k(q_m, q_{m-1}, \mathbf{w}, m)$$

- $L(q_{m-1}, q_m, m)$ — штраф за переход из q_{m-1} в q_m в момент m.
- Штраф не зависит от предыдущего пути (только от последнего состояния).
- Можно запомнить последнее состояние и применить алгоритм Витерби (почти так же, как в марковских моделях).
- Введём частичные штрафы $\alpha_{m,i}$:

$$\alpha_{m,i} = \max_{\mathbf{q_1}_{,m-1},q_m=i} \log p(\mathbf{q}_{1,m}|\mathbf{w})$$

Пересчёт вероятностей

- Формула пересчёта вероятностей:

$$\log p(\mathbf{q}_{1,m}|\mathbf{w}) = \log p(\mathbf{q}_{1,m-1}|\mathbf{w}) + \sum_k \theta_k f_k(q_m, q_{m-1}, \mathbf{w}, m)$$

- $L(q_{m-1}, q_m, m)$ — штраф за переход из q_{m-1} в q_m в момент m.
- Штраф не зависит от предыдущего пути (только от последнего состояния).
- Можно запомнить последнее состояние и применить алгоритм Витерби (почти так же, как в марковских моделях).
- Введём частичные штрафы $\alpha_{m,i}$:

$$\alpha_{m,i} = \max_{\mathbf{q_1}_{,m-1},q_m=i} \log p(\mathbf{q}_{1,m}|\mathbf{w})$$

- Формула пересчёта:

$$\begin{aligned} \alpha_{m+1,i} &= \max_j \alpha_{m,j} + L(j, i, m+1) \\ L(j, i, m+1) &= \sum_k \theta_k f_k(q_i, q_j, \mathbf{w}, m+1) \end{aligned}$$

Декодирование оптимальной последовательности

Нахождение оптимальной выходной последовательности:

- Вычислить частичные штрафы и обратные ссылки:

$$\begin{aligned}
\alpha_{0,j} &= [\![j = 0]\!], \\
\alpha_{m,i} &= \max_j \alpha_{m,j} + \sum_k \theta_k f_k(j, i, \mathbf{w}, m), \ m = 1, \ldots, n, \\
\delta_{m,i} &= \operatorname{argmax}_j \alpha_{m,j} + \sum_k \theta_k f_k(j, i, \mathbf{w}, m)
\end{aligned}$$

Декодирование оптимальной последовательности

Нахождение оптимальной выходной последовательности:

- Вычислить частичные штрафы и обратные ссылки:

$$\begin{aligned} \alpha_{0,j} &= [\![j = 0]\!], \\ \alpha_{m,i} &= \max_j \alpha_{m,j} + \sum_k \theta_k f_k(j, i, \mathbf{w}, m), \ m = 1, \ldots, n, \\ \delta_{m,i} &= \operatorname{argmax}_j \alpha_{m,j} + \sum_k \theta_k f_k(j, i, \mathbf{w}, m) \end{aligned}$$

- Восстановить последовательность по обратным ссылкам:

$$\begin{aligned} q(n) &= \operatorname{argmax}_i A_{n,i} \\ q(t-1) &= \delta_{t,q(t)} \end{aligned}$$

- По состояниям восстанавливается разбор.

Декодирование оптимальной последовательности

- Сложность декодирования: $L * M * R * K$.
  - L — длина последовательности,
  - M — максимальное число “активных” состояний,
  - R — количество способов продолжить состояние,
  - K — число признаков.

Декодирование оптимальной последовательности

- Сложность декодирования: $L * M * R * K$.
  - L — длина последовательности,
  - M — максимальное число “активных” состояний,
  - R — количество способов продолжить состояние,
  - K — число признаков.
- Если состояние — энграмма порядка m, то $M \sim R^m$, где R — максимальная степень неоднозначности слова.

Декодирование оптимальной последовательности

- Сложность декодирования: $L * M * R * K$.
  - L — длина последовательности,
  - M — максимальное число “активных” состояний,
  - R — количество способов продолжить состояние,
  - K — число признаков.
- Если состояние — энграмма порядка m, то $M \sim R^m$, где R — максимальная степень неоднозначности слова.
- Число признаков тоже растёт экспоненциально по m.

- Следствие: $m > 2$ нереализуемо на практике, с $m = 2$ проблемы для языков с развитой морфологией.

Декодирование оптимальной последовательности

- Сложность декодирования: $L * M * R * K$.
  - L — длина последовательности,
  - M — максимальное число “активных” состояний,
  - R — количество способов продолжить состояние,
  - K — число признаков.
- Если состояние — энграмма порядка m, то $M \sim R^m$, где R — максимальная степень неоднозначности слова.
- Число признаков тоже растёт экспоненциально по m.
- Сложность обучения: $T * N * C_0$, где C_0 — сложность декодирования, T — число эпох обучения, N — размер обучающей выборки.
- Следствие: $m > 2$ нереализуемо на практике, с $m = 2$ проблемы для языков с развитой морфологией.
- Часто CRF реализуют иерархически: сначала грубая классификация (части речи), потом более точная.

Недостатки CRF

- Преимущества CRF:
  - Большая гибкость по сравнению с марковскими моделями.
  - Локальные признаки произвольной природы (в том числе лексические).

Недостатки CRF

- Преимущества CRF:
  - Большая гибкость по сравнению с марковскими моделями.
  - Локальные признаки произвольной природы (в том числе лексические).
- Недостатки CRF:
  - Большие затраты (время и память) на обучение.
  - Большое количество признаков (в том числе избыточных).

Недостатки CRF

- Преимущества CRF:
  - Большая гибкость по сравнению с марковскими моделями.
  - Локальные признаки произвольной природы (в том числе лексические).
- Недостатки CRF:
  - Большие затраты (время и память) на обучение.
  - Большое количество признаков (в том числе избыточных).
  - Невозможность учитывать удалённый контекст.

Недостатки CRF

- Преимущества CRF:
  - Большая гибкость по сравнению с марковскими моделями.
  - Локальные признаки произвольной природы (в том числе лексические).
- Недостатки CRF:
  - Большие затраты (время и память) на обучение.
  - Большое количество признаков (в том числе избыточных).
  - Невозможность учитывать удалённый контекст.
- CRF применимы к произвольной задаче разметке последовательно распознавание именованных сущностей, разбиение на морфемы и т. д.

Недостатки CRF

- Преимущества CRF:
  - Большая гибкость по сравнению с марковскими моделями.
  - Локальные признаки произвольной природы (в том числе лексические).
- Недостатки CRF:
  - Большие затраты (время и память) на обучение.
  - Большое количество признаков (в том числе избыточных).
  - Невозможность учитывать удалённый контекст.
- CRF применимы к произвольной задаче разметке последовательно распознавание именованных сущностей, разбиение на морфемы и т. д.
- В последние годы уступают нейронным моделям.

Качество морфологического анализа

Старейший ресурс — Penn Treebank (английский):
$\approx$ 3 млн. слов (обучение) + 0.5 млн. слов настройка + 0.5 млн. слов тестирование.

Статья	Метод	Результат
Huang et al., 2015	нейронные сети (BiLSTM-CRF)	97.55
dos Santos, Zadrozny, 2015	нейронные сети (символьная модель)	97.32
Spoustova et al., 2009	CRF	97.23
Gimenez, Marquez, 2004	локальная класси-фикация (SVM)	97.16
Brants, 2000	HMM	96.46

Качество морфологического анализа

Результаты морфологического анализа для других языков (соревнование 2017 года по автоматическому синтаксическому анализу):

Язык	Результат
Арабский	82.08
Немецкий	77.53
Английский	91.50
Испанский	96.89
Французский	94.93
Венгерский	71.80
Русский	94.20
Чешский	91.63

Параметры систем не настраивались под конкретный язык!

Признаки в морфологическом анализе

- Наиболее эффективная из доступных в открытом доступе систем — система проекта UDPIPE (университет Праги, UFAL).

Признаки в морфологическом анализе

- Наиболее эффективная из доступных в открытом доступе систем — система проекта UDPIPE (университет Праги, UFAL).
- Основана на условных случайных полях.
- Обучается с помощью усреднённого персептрона.

Признаки в морфологическом анализе

- Наиболее эффективная из доступных в открытом доступе систем — система проекта UDPIPE (университет Праги, UFAL).
- Основана на условных случайных полях.
- Обучается с помощью усреднённого персептрона.
- Использует развёрнутые шаблоны признаков (изначально разрабатывалась для чешского).

Признаковое описание для английского языка

Context predicting whole tag	
Tags	Previous tag Previous two tags First letter of previous tag
Word forms	Current word form Previous word form Previous two word forms Following word form Following two word forms Last but one word form
Current word affixes	Prefixes of length 1-9 Suffixes of length 1-9
Current word features	Contains number Contains dash Contains upper case letter

Признаковое описание для чешского языка

Context predicting whole tag	
Tags	Previous tag Previous two tags First letter of previous tag
Word forms	Current word form Previous word form Previous two word forms Following word form Following two word forms Last but one word form
Current word affixes	Prefixes of length 1-9 Suffixes of length 1-9
Current word features	Contains number Contains dash Contains upper case letter