Применение условных случайных полей в задачах обработки текстов на естественном языке

А. А. Романенко

Научный руководитель: К. В. Воронцов
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

10 июня 2014г.
Москва

План презентации

Задача разметки последовательности

Дано:

$\mathbf{x} = \{x_1, \ldots, x_T\}$ — наблюдаемые переменные (слова),

$\mathbf{y} = \{y_1, \ldots, y_T\}$ — скрытые переменные (метки слов).

$\forall t = 1, \ldots, T \; x_t \in \mathcal{V}, \; y_t \in \mathcal{S}$.

Множества $\mathcal{S}, \mathcal{V}$ — конечные.

Элементы последовательностей $\mathbf{x}$ и $\mathbf{y}$ не i.i.d.

Стадия обучения:

Размеченная выборка $D = \{(\mathbf{y}^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^{N}$

Найти алгоритм $a : \mathcal{V}^{|T|} \to \mathcal{S}^{|T|}$, максимизирующий некоторый критерий $Q(D, a)$.

Стадия тестирования:

Размеченная выборка $D' = \{\mathbf{x}^{(i)})\}_{i=1}^{M}$

Построить $\mathbf{y} = a(\mathbf{x})$ для любого $\mathbf{x} \in D'$

Примеры задач разметки последовательности в NLP

- Выделение именованных сущностей (NER)
- Выделение синтаксических групп (Chunking)
- Выделение временных выражений (TER)
- Определение частей речи слов (POS tagging)
- Полное снятие морфологической неоднозначности
- Разрешение анафоры

Пример: POS tagging

Пример определения частей речи для предложения

«Сорока жила на горе.»

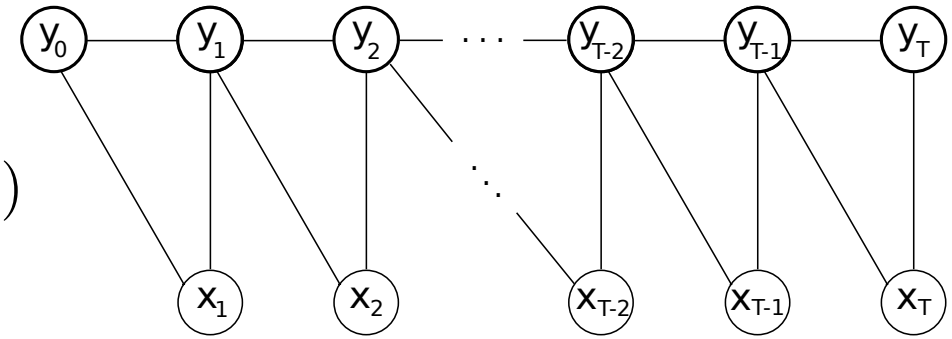
y	Сущ	Глаг	Пред	Сущ
x	Сорока	жила	на	горе
Возможные метки	Числ Сущ	Глаг Сущ	Пред Межд	Сущ

Линейная модель CRF

Linear-chain CRF

Линейная модель CRF — это разновидность марковской модели случайных полей, у которой множество скрытых переменных вытянуто в цепочку.

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_t \Psi_t(y_t, y_{t-1}, \mathbf{x}_t)$$

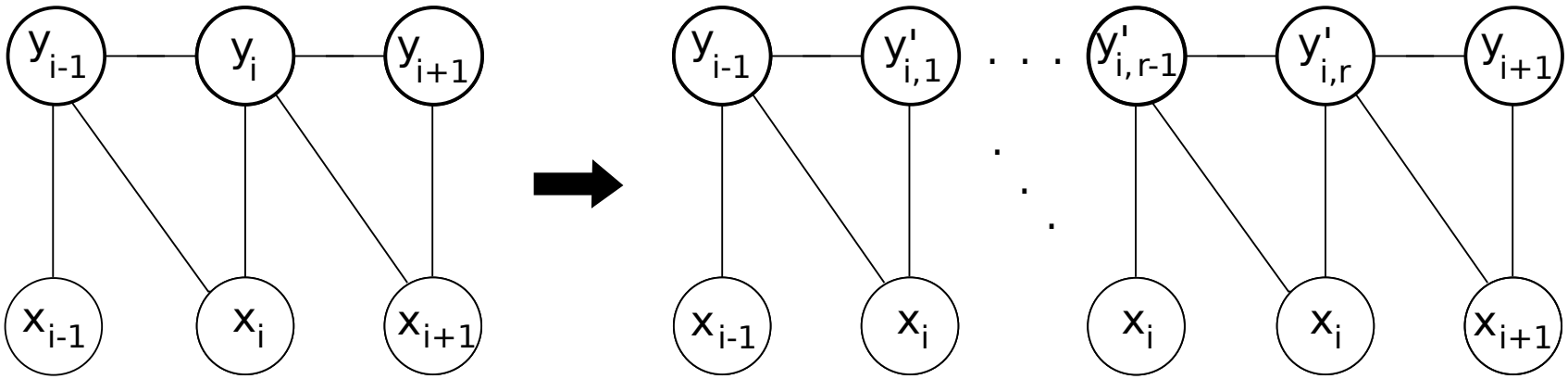


Модификация линейной модели

Для задач, где скрытую переменную $y_i \in \mathcal{Y}$ нужно классифицировать на несколько классов, т. е.

$$\mathcal{Y} \equiv \mathcal{Y}'_1 \times \mathcal{Y}'_2 \times \cdots \times \mathcal{Y}'_r,$$

предлагается использовать модификацию линейной модели:



Свойства модифицированной модели

Применение модифицированной модели

- позволяет моделировать зависимости между классами путём выбора последовательности, в которой разворачиваются размерности $1, \ldots, r$ декартова произведения
- уменьшает количество допустимых меток на $\prod_{j=1}^{r} |\mathcal{Y}'_j| - \sum_{j=1}^{r} |\mathcal{Y}'_j|$

За счет этого

- уменьшается количество параметров
- снижается эффект переобучения
- ускоряется подсчет $Z(\mathbf{x})$, а значит и процесс обучения

Понятие временного выражения

Временное выражение

Временны́м выражением (*temporal expression, timex*) называется выражение естественного языка, несущее временную окраску и обозначающее точку во времени, промежуток времени или периодичность некоторого события.

Примеры временных выражений:

- Что будут показывать *сегодня ночью* по пятому каналу?
- Встреча с руководством состоится *через 2 недели*.
- *Ежедневно в 7 часов вечера* в больнице делают обход.

Задача выделения временного выражения

Задачу можно свести к задаче разметки последовательности:
$\mathbf{x}$ — последовательность слов и их свойств (последовательность наблюдаемых переменных);
$\mathbf{y}$ — последовательность меток, $y_i \in \{B, I, O\}$;

$$p(\mathbf{y}|\mathbf{x}) \to \max_{\mathbf{y}}$$

Здесь:

- B — метка начала выражения;
- I — метка любого не первого слова в выражении;
- O — метка для слов, не входящих в выражение.

Порождение признаков

Все признаки слова можно разделить на группы:

- Грамматические признаки слова
  («сущ», «род. падеж» и т. д.)
- Положение слова в предложении
  («первое слово в предложении» и т. д.)
- Свойства написания слов
  («есть заглавные буквы», «есть пунктуация»и т. д.)
- Является ли слово специфическим
  («название месяца», «число» и т. д.)
- Признаки соседних слов, перечисленных выше групп

Отбор признаков

- В модель включались наиболее информативные признаки
- Информативность признака — разность числа ошибок алгоритма Random Forest на обучающих данных с включенным в модель признаком и исключенным
- Сам алгоритм Random Forest использовался для классификации
- Наиболее важные группы признаков:
  - является ли слово специфическим (день, год, и т. д.),
  - часть речи,
  - падеж

Условия эксперимента

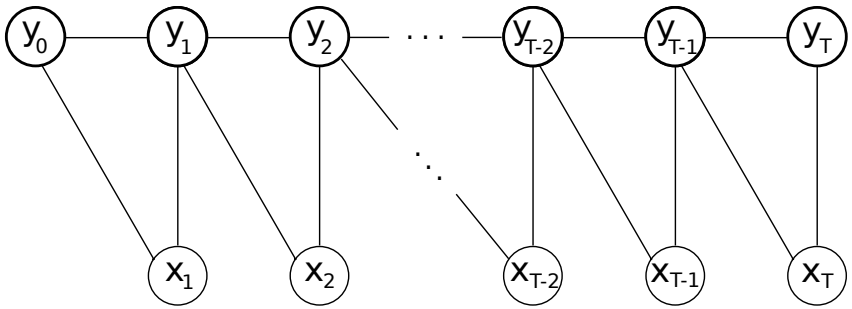
Данные

- Обучение: ≈ 380000 предложений, ≈ 5000 выражений.
- Контроль: ≈ 2000 предложений, ≈ 500 выражений.

Меры качества

- полнота $R = \frac{tp}{tp+fn}$,
- точность $P = \frac{tp}{tp+fp}$,
- F_1 – мера $F_1 = \frac{2PR}{P+R}$.

Применяемая модель CRF



Результаты

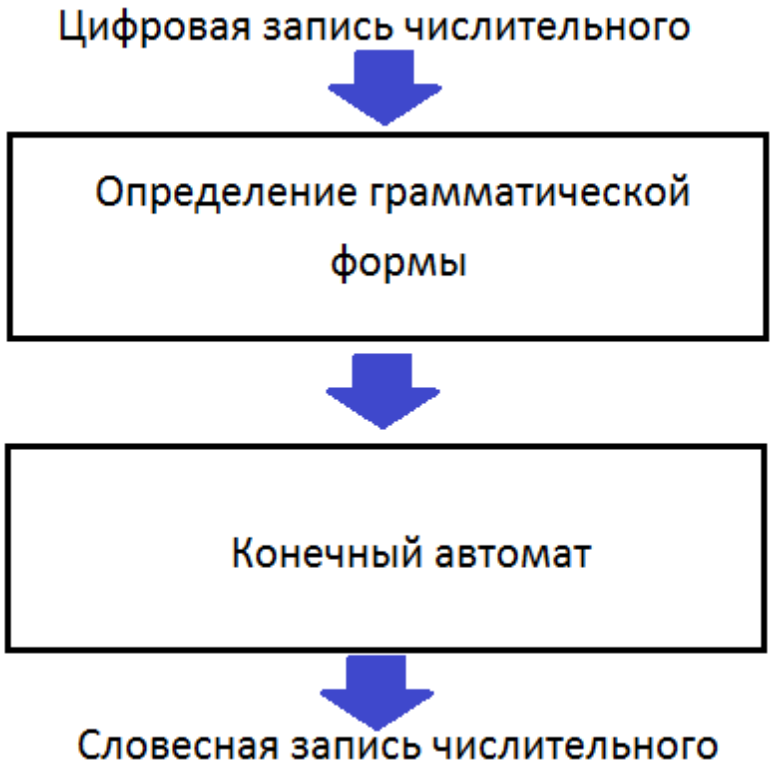
Результаты работы алгоритмов при наилучшей конфигурации признаков

Алгоритм	P	R	F_1
Base-line	95,7	85,7	90,4
RF	96,4	87,1	91,5
CRF	96,3	89,9	93,05

Base-line — шаблонный алгоритм
RF — алгоритм Random Forest
CRF — классическая линейная модель CRF

Задача нормализации числительных

Задача нормализации цифровой записи числительных возникает при построении Text-to-Speach систем, когда машина должна правильно произнести числительное.



Пример

Виктор Ан на олимпиаде 2014 года занял 1 место в забеге на 500 метров

$\Downarrow$

Виктор Ан на олимпиаде две тысячи четырнадцатого года занял первое место в забеге на пятьсот метров

Признаковое описание

Множество $\mathcal{X}$ всех признаков наблюдаемой переменной:

$$\mathcal{X} = \mathrm{GRAM} \cup \mathrm{SPEL} \cup \mathrm{SPEC} \cup \mathrm{NEAR}.$$

- GRAM — грамматические метки слов («существительное», «предлог» и т. д.);
- SPEL — метки особенностей написания числительного (длина в символах, последняя цифра числительного);
- SPEC — является ли слово «характерным» для употребления с количественными или порядковыми числительными;
- NEAR — признаки соседних слов.

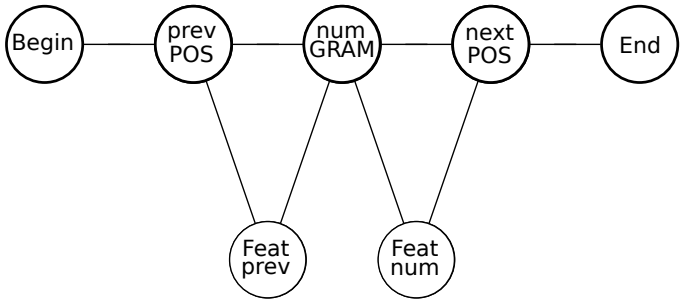
Грамматические метки числительного

Грамматическое описание числительного состоит из пяти меток

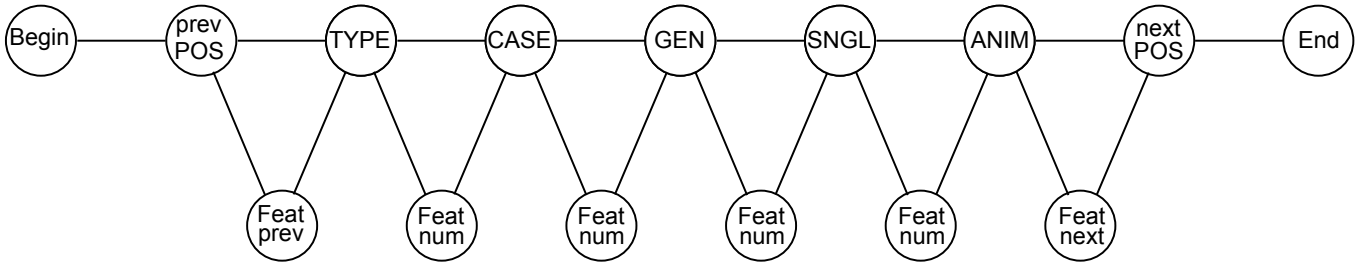
- TYPE — тип числительного
  (количественное или порядковое)
- CASE — падеж числительного
  (именительный, родительный, ...)
- GEND — род числительного
  (мужской, женский, средний, неизвестно)
- SNGL — число числительного
  (единственное, множественное, неизвестно)
- ANIM — одушевленность числительног
  (одушевленное, неодушевленное, неизвестно)

Используемые модели CRF

Классическая модель:



Модифицированная модель:



Условия эксперимента

Данные

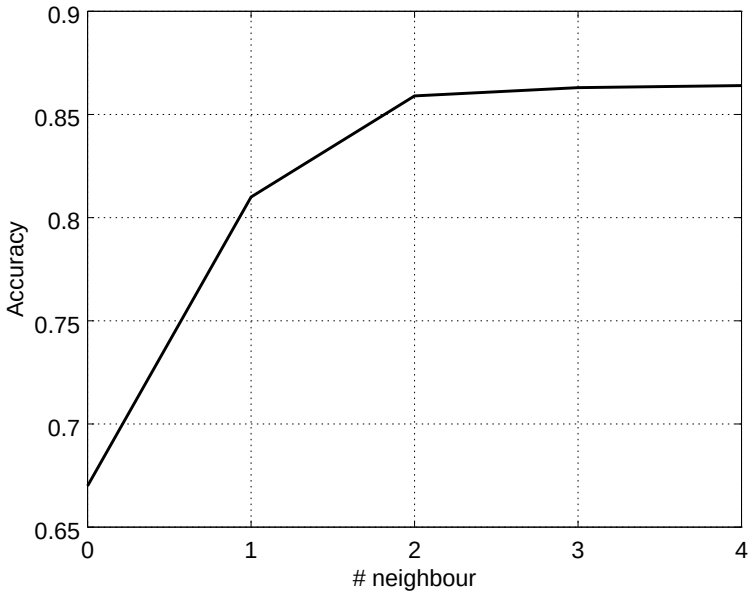
- Национальный корпус русского языка
- Обучение: 8251 предложение
- Контроль: 2017 предложений

Измерение качества

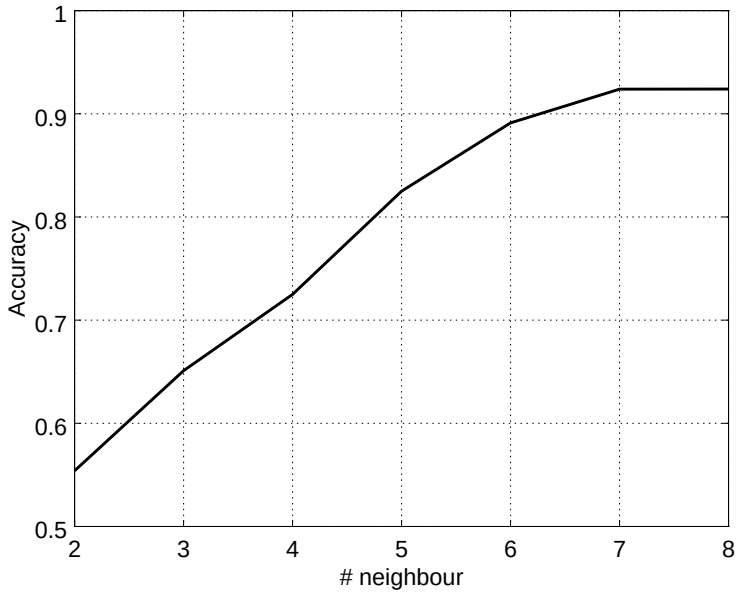
- Измерялось только качество определения грамматических меток числительных
- Использовались полнота R, точность P и F_1 – мера для измерения качества по меткам в отдельности
- Использовалась аккуратность $Acc = \frac{\text{число верных ответов}}{\text{число ответов}}$ для измерения качества работы в целом

Результаты

Зависимость качества Acc от числа слов, включаемых в признаковое описание



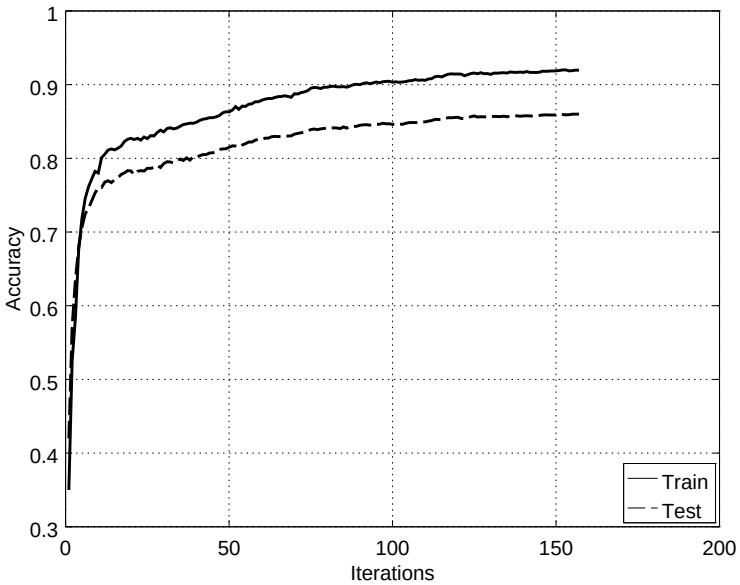
Классическая модель



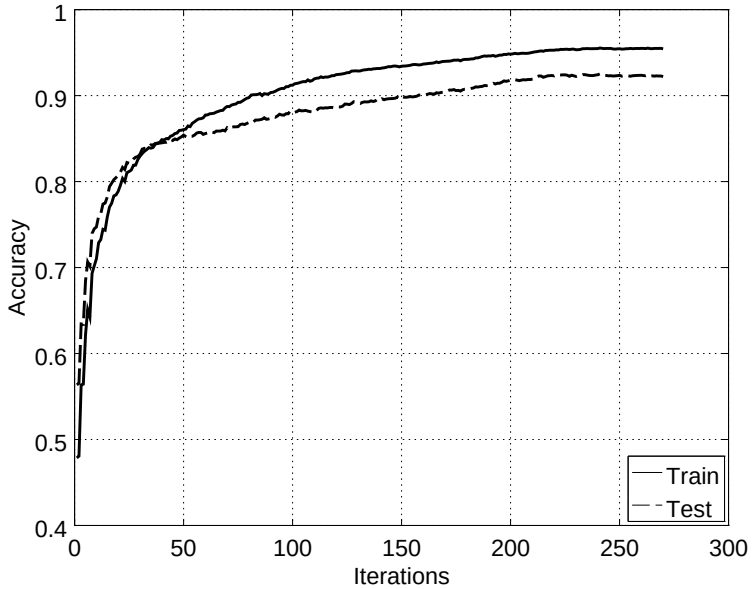
Модифицированная модель

Результаты

Качество Acc моделей, получаемых в процессе обучения



Классическая модель



Модифицированная модель

Результаты

Доля верных ответов на тестовом множестве $A = 92{,}39\%$.

Качество определения грамматических характеристик, усредненное по группам меток:

Мера качества	TYPE	CASE	GEN	SNGL	ANIM
P	97,21	91,33	89,77	82,39	87,66
R	97,21	92,93	90,74	85,97	95,05
F_1	97,21	92,10	90,24	84,05	91,11

Результат 5-fold CV: $A_{CV} = 92{,}21\%$.

Заключение

Результаты, выносимые на защиту

1. Предложена модификация линейной модели CRF
2. Предложено решение двух задач NLP с помощью CRF
3. В экспериментах показано, что точность предложенного метода выше, чем у классического

Публикации

- Muzychka S., Romanenko A., Piantkovskaya I. *CRF for morphological disambiguation in Russian*, Computational Linguistics and Intelligent Technologies, 2014
- Kudinov S., Romanenko A., Piantkovskaya I. *CRF in segmentation and noun phrase inclination tasks for Russian*, Computational Linguistics and Intelligent Technologies, 2014