ИСПОЛЬЗОВАНИЕ МЕТОДА УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ ДЛЯ ОБРАБОТКИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Антонова А. Ю. (a-antonova@list.ru)
НИУ Высшая школа экономики, Москва, Россия

Соловьев А. Н. (a.solovyev@i-teco.ru)
ЗАО «Ай-Теко», Москва, Россия

Работа посвящена исследованию метода условных случайных полей (Conditional Random Fields — CRF) на русскоязычных текстах. В частности, продемонстрированы результаты использования CRF в задачах распознавания именованных сущностей, определения частей речи и сентимент-анализа сообщений относительно объекта тональности. Результаты CRF сравниваются с результатами, полученными другими методами.

Ключевые слова: Марковские поля, метод условных случайных полей (CRF), машинное обучение, распознавание именованных сущностей, анализ тональности сообщений, определение частей речи

CONDITIONAL RANDOM FIELD MODELS FOR THE PROCESSING OF RUSSIAN

Antonova A. Y. (a-antonova@list.ru)
Higher School of Economics, Moscow, Russia

Soloviev A. N. (a.solovyev@i-teco.ru)
CJSC I-Teco, Moscow, Russia

The paper aims to illustrate the applicability of conditional random field (CRF) models to Russian texts. Introduced in 2001, CRF method has been successfully exploited and proved its efficiency for a variety of NLP tasks. Its main advantage over HMM is the possibility to model the dependencies and interdependencies in sequential data. Yet this approach has not been widely used for Russian.

Since CRF operates with language-independent features, its initial adaptation for Russian can be minimalistic. We show how CRF models produce state-of-the-art quality for several basic NLP tasks, including named entity recognition, part-of-speech tagging and object-oriented sentiment analysis. We exploited CRF-Suite tool to train and evaluate our models. We used a corpus of news texts for NER and POS-tagging tasks and a subcorpus from Russian Twitter for SA. The results of the evaluation were compared to other existing methods for each of the three tasks.

Key words: Markov fields, conditional random fields (CRF), machine learning, named entity recognition, object oriented sentiment analysis of short messages, part-of-speech tagging

1. Введение

Обработка больших массивов данных поставила перед разработчиками инструментальных систем задачу обеспечения высокой скорости получения результата без существенной потери качества. Так, например, только в русскоязычном сегменте Twitter ежедневно публикуются около 8-10 млн. твитов. Увеличение объема потока текстовых данных привело к тому, что статистические методы стали неотъемлемой частью области text mining.

К числу методов, успешно применяемых в обработке текста, относится метод случайных Марковских полей и его модификация — метод условных случайных полей (CRF — Conditional Random Fields), который нашел широкое применение в лингвистических приложениях, требующих разметки больших объемов текста на основе некоторых параметров. Чаще всего этот метод применяют в задачах распознавания специальных терминов [Finkel at al. 2004, Dingare at al. 2004], именных групп [McCallum 2003, Ratinov 2009], поверхностного синтаксиса (pos-taggers и shallow parsing) [Sha 2003, Sutton 2004] и т. п. Также данный метод находит свое применение в задачах разрешения лексической омонимии [Sutton 2004], анафорических ссылок [McCallum 2005], сентимент-анализе [Choi 2005, Sadamitsu 2008, Mao 2006], машинном переводе [Lavergne 2011]. (Этот ряд можно продолжить набором задач из других предметных областей: биоинформатики, компьютерной графики и пр.)

Метод CRF хорошо исследован для английского, немецкого, арабского, китайского и некоторых других широко распространенных языков. К сожалению, для русского языка этот метод еще не нашел столь широкого применения.

Целью нашей статьи является апробировать возможности CRF применительно к русскому языку на примере определения частей речи, выделения именованных сущностей и сентимент-анализа текста относительно объекта тональности. Важной особенностью нашего исследования являлось отсутствие лингвистической предобработки текста, т. е. анализ проводился на плоском тексте.

2. Описание CRF

Метод CRF (Conditional Random Fields) [Lafferty 2001, Klinger 2007] относится к статистическим лингвистическим методам. Данный метод является одной из возможных реализаций Марковских случайных полей.

Марковским случайным полем или Марковской сетью (Markov random field, Markov network) называют графовую модель, которая используется для представления совместных распределений набора нескольких случайных переменных. Формально Марковское случайное поле состоит из следующих компонентов:

- неориентированный граф или фактор-граф $G = (V, E)$,где каждая вершина $v \in V$ является случайной переменной X и каждое ребро $\{u,v\} \in E$ представляет собой зависимость между случайными величинами u и v.
- набор потенциальных функций (potential function) или факторов $\{\varphi_k\}$, одна для каждой клики в графе G (полный подграф). Функция φ_k ставит

каждому возможному состоянию элементов клики в соответствие некоторое неотрицательное вещественнозначное число.

Вершины, не являющиеся смежными, должны соответствовать условно независимым случайным величинам. Группа смежных вершин образует клику, набор состояний вершин является аргументом соответствующей потенциальной функции.

Совместное распределение набора случайных величин $X=\{x_k\}$ в Марковском случайном поле вычисляется по формуле:

$$P(x) = \frac{1}{z}\prod_k \varphi_k(x_{\{D\}}) \quad (1)$$

где $\varphi_k(x_{\{k\}})$ — потенциальная функция, описывающая состояние случайных величин в k-ой клике; Z — коэффициент нормализации вычисляется по формуле:

$$Z = \sum_{x \in X} \prod_k \varphi_k(x_{\{k\}}) \quad (2)$$

Множество входных лексем $X=\{x_t\}$ и множество соответствующих им типов $Y=\{y_t\}$ в совокупности образуют множество случайных переменных $V=X \cup Y$. Для решения задачи извлечения информации из текста достаточно определить условную вероятность $P(Y \mid X)$. Потенциальная функция имеет вид:

$$\varphi_k(x_{\{k\}}) = \exp\left(\sum_k \lambda_k f_k(y_y, y_{t-1}, x_t)\right) \quad (3)$$

где $\Sigma\{\lambda_k\}$ вещественнозначный параметрический вектор, и $\Sigma\{f_k(y_t, y_{t-1}, x_t)\}$ — набор признаковых функций. Тогда линейным условным случайным полем называется распределение вероятности вида:

$$p(y \mid x) = \frac{1}{z(x)}\prod_k \exp\left(\sum_k \lambda_k f_k(y_y, y_{t-1}, x_t)\right) \quad (4)$$

Коэффициент нормализации тогда Z(x) вычисляется по формуле:

$$Z(x) = \sum_y \prod_k \exp\left(\sum_k \lambda_k f_k(y_y, y_{t-1}, x_t)\right) \quad (5)$$

Метод CRF, как и метод MEMM (Maximum Entropy Markov Models), относится к дискриминативным вероятностным методам, в отличие от генеративных методов, таких как HMM (Hidden Markov Models) [Rabiner 1989, Николенко 2006] или метод «Наивного Баеса» (Naïve Bayes [McCallum 1998]).

По аналогии с MEMM [Bishop 2006, McCallum 2000], выбор факторов-признаков для задания вероятности перехода между состояниями при наличии наблюдаемого значения x_t зависит от специфики конкретных данных, но в отличии от того же MEMM, CRF может учитывать любые особенности и взаимозависимости в исходных данных. Вектор признаков $L=\{\lambda_k\}$ рассчитывается

на основе обучающей выборки и определяет вес каждой потенциальной функции. Для обучения и применения модели используются алгоритмы, аналогичные алгоритмам HMM: Витерби и его разновидность — алгоритм «вперед-назад» (forward-backward) [Sutton 2008].

Как показано в [Sutton 2006], скрытую Марковскую модель можно рассматривать как частный случай линейного условного случайного поля (CRF). В свою очередь, условное случайное поле можно рассматривать как разновидность Марковского случайного поля (см. рис. 1).

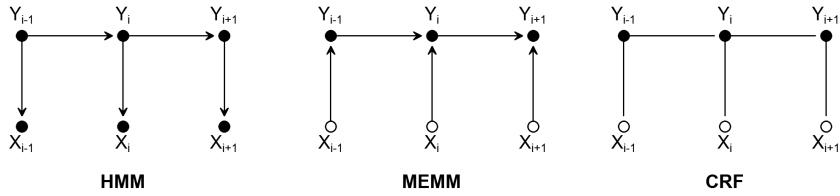

Рис. 1. *Изображение в виде графов для методов HMM, MEMM и CRF. Незакрашенные окружности обозначают, что распределение случайной величины не учитывается в модели. Стрелки указывают на зависимые узлы*

В условных случайных полях отсутствует т. н. label bias problem — ситуация, когда преимущество получают состояния с меньшим количеством переходов [Lafferty at al. 2001], так как строится единое распределение вероятностей и нормализация (коэффициент Z(x) из формулы (5)) производится в целом, а не в рамках отдельного состояния. Это, безусловно, является преимуществом метода: алгоритм не требует предположения независимости наблюдаемых переменных. Кроме того, использование произвольных факторов позволяет описать различные признаки определяемых объектов, что снижает требования к полноте обучающей выборки. Недостатком подхода CRF является вычислительная сложность анализа обучающей выборки, что затрудняет постоянное обновление модели при поступлении новых обучающих данных.

На сегодняшний день именно метод CRF является наиболее популярным и точным способом извлечения объектов из текста [Sarawagi 2008]. Например, он был реализован в проекте Стэндфордского университета Stanford Named Entity Recognizer [Stanford NER]. С той же целью этот метод с этого года успешно применяется в системах «X-Files» и «Аналитический Курьер» [Киселев 2007].

Как уже было отмечено, качество выделения сущностей методом, основанном на статистическом подходе, определяется полнотой обучающей выборки. Для этого необходимы размеченные корпуса, содержащие не только лексику, характеризующую данную предметную область, но и статистически значимые N-граммы, наиболее полно покрывающие произвольный текст. В отличии от других статистических методов, метод CRF, основанный на фактор-графах, требует гораздо меньшей обучающей выборки, поскольку статистически значимые сочетания могут быть определены как набор клик для исследуемого

объекта. В зависимости от решаемой задачи, на практике достаточно объема от нескольких сотен тысяч до миллионов термов. При этом точность будет определяться не только объемом выборки, но и выбранными факторами.

3. Эксперимент

Для экспериментов был выбран проект с открытым кодом http://www.chokkan.org/software/crfsuite/. Его преимущества относительно других открытых проектов (CRF++; HCRF; CRF package; набор утилит, основанных на алгоритме CRF Стенфордского Университета и др.) заключаются в его простоте, универсальности и скорости работы. Этот алгоритм универсален для любых видов классификационных задач, при этом включает в себя набор разных оптимизационных обучающих методов:

- Limited-memory Quasi-Newton method (LBFGS) [Andrew 2007] — квазиньютоновский метод с ограничением памяти BFGS [Nocedal 1980].
- Stochastic Gradient Descent (L2SGD) [Shalev-Shwartz 2007] — метод градиентного спуска.
- Averaged Perceptron (AP) [Collins 2002] — метод усредненного персептрона[1].
- Passive Aggressive (PA) [Crammer 2006] — алгоритм, основанный на бинарной классификации.
- Adaptive Regularization Of Weight Vector (AROW) [Mejer 2010] — метод адаптивной регуляризации весов вектора.

Все эксперименты проводились с различными методами оптимизации, из которых выбирался наилучший. Используемый алгоритм относится к линейным методам CRF.

3.1. Определение именованных сущностей (NER)

Метод CRF был использован в задаче распознавания именованных сущностей — NER (Name Entity Recognition). Для этого вручную был размечен небольшой корпус из новостных лент СМИ объемом более 71 тыс. предложений на самую разную тематику (более 1,5 млн. словоформ). Каждое предложение содержало хотя бы одну именную сущность. Для проведения тестов было выбрано пять типов сущностей (Табл. 1): физические лица (имена, фамилии, отчества), юридические лица (названия организаций, компаний и пр.), географические объекты (названия городов, улиц, рек и пр.), продукты (названия продуктов, включая марку и бренд) и события (названия форумов, съездов, мероприятий и пр.).

---

[1] Метод усредненного персептрона, как и алгоритм Passive Aggressive, часто используются в качестве самостоятельных методов в решении ряда лингвистических задач (например, классификации). В методе CRF эти алгоритмы используются только для оптимизации потенциальных функций (см. уравнение (3)).

Табл. 1. Частотное распределение именованных сущностей в размеченном корпусе. NAME — физ.лица, ORG — юр.лица, GEO — географические названия, PROD — продукты, EVENT — события

Тип сущности	Абс. частота в корпусе	Отн. частота в корпусе, %
NAME	37 729	24,99
ORG	43 646	28,91
GEO	52 889	35,04
PROD	6 425	4,26
EVENT	10 263	6,80

Для обучения и тестирования с кросс-валидацией полученный корпус был разделен на 4 части: три части использовались для обучения и одна для тестов. На одном и том же корпусе обучались модели с использованием каждого из перечисленных оптимизационных методов с различными экспертно задаваемыми параметрами.

В качестве факторов были выбраны только n-граммы (длины от двух до пяти) и графематические особенности написания именованных сущностей. Например, все заглавные буквы без точек (*МТС*), одна заглавная буква с точкой (*W.*), первые строчные с точкой, затем заглавная (*ул.Мира* — при слитном написании) и т. д. Всего было определено 14 параметров.

3.2. Сентимент-анализ

Современный сентимент-анализ текста включает в себя по крайней мере три вида задач [Liu 2010]:

1. Классификация тональных сообщений (позитив/негатив или более тонкая градация);
2. Определение сентимента относительно заданного объекта тональности (ОТ) (часто с последующей визуальной разметкой дерева зависимостей предложения, например, «***Правительство*** *одобрило новый указ президента, нарушив конституцию*», тут ОТ — «*Правительство*», у которого два противоположных сентимента);
3. Определение сентимента объекта тональности относительно его имплицитных и эксплицитных атрибутов (feature-based). Например, «У этого телефона большой аккумулятор, правда и вес немаленький», тут ОТ — *телефон*, а его атрибуты — *вес* и *аккумулятор* — имеют разную полярность.

В данном исследовании мы ограничимся второй областью задач сентимент-анализа, а именно: будем определять сентимент заданного объекта тональности в произвольном сообщении.

Для этого по заданным объектам (числом более 20, например, *МГУ, РЖД, радио, пальто, кино* и пр.) нами был собран корпус коротких твит-сообщений и экспертно размечен. Всего в 20 тысячах твитах были отмечены около 21 тысячи ОТ.

Для разметки слов использовались тональные словари, которые были собраны при разработке метода сентимент-анализа, основанного на правилах (описание и список словарей см. в [Pazelskaya 2011, Solovyev 2012]). К данным словарям были добавлены словари инверторов и шифтеров. Число всех словарей составило 34. Тональные словари были дополнены полными наборами словоформ (т. к. мы используем плоский текст без предобработки). Таким образом, общее число словарных форм получилось более 20 тыс., а полное число всех слов — более 400 тыс.

Вхождение слова или словосочетания в тональные словари определенного типа рассматривалось в качестве факторов CRF, сами слова не учитывались. Слово, не получавшее значения, исключалось из анализа, при этом знаки пунктуации сохранялись с нулевым весом. Ширина окна анализа составила 2–5 грамм. Классификация производилась на три класса: позитивный, негативный и нейтральный (Табл. 2).

Табл. 2. Частотное распределение в корпусе объектов тональности (ОТ) по классам тональности

Класс	Кол-во ОТ	%
Позитивный	6435	31,08
Негативный	6034	29,14
Нейтральный	8236	39,78

3.3. Определение частей речи (POS-tagger)

Цель данного раздела — показать, что метод CRF, примененный в задаче определения частей речи (частеречного тэгирования), дает результаты не ниже уровня, достигаемого в данный момент статистическими системами на материале русского языка. В этом случае мы можем говорить, что CRF метод не хуже или даже лучше других, ранее разработанных методов. Полученный результат мы будем сравнивать с результатами статистических систем. (Словарные системы представляют собой другой подход.)

Задачу определения части речи, традиционно рассматриваемую в прикладных системах как задачу классификации, решают с помощью 1) метода опорных векторов (SVM [Giménez 2004]), адаптированный для многозначной классификации,[2]) HMM [Brants 2000] и ряда других (AP, ME и т. п.) методов.

Для европейских языков (и прежде всего английского) статистические тэггеры уже давно [Brants 2000] приблизились и преодолели барьер в 97 % [Manning 2011]. В случае русского языка ситуация иная. Почти все участники соревнования морфологических парсеров, прошедшего в 2010-м году

[2] http://www.lsi.upc.edu/~nlp/SVMTool/

[Ляшевская и др. 2010], представляли инструменты, в которых работали алгоритмы, основанные на правилах. По дорожке PoS-tagging был достигнут результат, близкий к абсолютному: 99,4 % правильно определенных частей речи. Что касается, уровня качества статистических систем, в работе [Sharoff, Nivre 2011] описывается тэггер, обученный на корпусе SinTagRus, показавший результат 97 %.

В настоящей работе мы проводим сравнение между полученным CRF-классификатором, тэггером разработки Стенфордского университета и TreeTagger'ом[3], которые являются свободно распространяемыми как в виде приложений, так и в виде исходного кода.

Размер обучающего корпуса 2.2 млн. словоформ, тестового 670 тыс. словоформ. Как и в случае задачи выделения сущностей, основную часть обоих корпусов составили сообщения новостных лент на разную тематику. Разметка корпуса производилась с помощью морфологического модуля системы «Аналитический курьер» [Киселев 2007], наиболее частотные случаи омонимии размечались вручную. Список выделяемых частей речи был ограничен возможностями системы (так, например, совсем не выделялись частицы и междометия — и те, и другие попали в категорию «прочее» и были обозначены в разметке одинаковым тэгом).

4. Результаты

4.1. NER

Результаты тестирования NER представлены в Табл. 3–5. Лучшие результаты показывали методы оптимизации Averaged Perceptron и Passive Aggressive на триграммной модели. Измерения проводились как по пяти, так и по трем типам сущностей (Табл. 3 и 4).

Точность и полнота рассчитывались по следующим формулам:

(6) $$Precision = \frac{A}{A + C + D} * 100\%$$

(7) $$Recall = \frac{A}{A + B + C} * 100\%$$

Здесь A — количество верных срабатываний системы;
B — количество пропусков;
C — количество случаев типизации нетипизированной сущности;
D — тип сущности определен неверно.

Этот же принцип оценивания использовался и для задач, разбираемых ниже.

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Табл. 3. Оценка качества определения пяти типов сущностей с разными методами оптимизации

	NAME	GEO	ORG	PROD	EVENT	Среднее
Точность						
AROW	91,36	89,57	83,49	75,08	77,61	83,43
L2SGD	91,19	89,52	83,88	70,14	75,69	82,09
LBFGS	91,39	89,53	84,25	69,95	76,05	82,23
PA	92,79	91,26	85,61	75,01	79,22	84,78
AP	92,58	91,22	85,23	76,09	79,43	**84,91**
Полнота						
AROW	93,85	94,05	86,48	84,02	80,65	87,81
L2SGD	93,61	94,22	87,23	84,28	83,27	88,61
LBFGS	93,57	94,34	87,10	83,39	83,32	88,34
PA	94,17	95,08	87,92	86,5	83,23	89,38
AP	94,23	95,01	88,02	87,14	83,84	**89,65**
F1						
AROW	93,60	91,75	84,96	79,30	79,10	85,54
L2SGD	92,38	91,81	85,53	76,71	79,34	85,15
LBFGS	92,47	91,87	85,65	76,08	79,52	85,12
PA	93,48	93,13	86,75	80,34	81,17	86,98
AP	93,39	93,08	86,60	81,24	81,57	**87,18**

Табл. 4. Оценка качества определения трех типов сущностей (результаты показаны только для двух методов оптимизации)

	NAME	GEO	ORG	Среднее
Точность, %				
AP	92,52	91,39	85,02	89,64
PA	93,05	91,47	85,5	90,01
Полнота, %				
AP	93,48	93,99	82,38	89,95
PA	93,25	93,96	82,23	89,81
F1, %				
AP	93,00	92,67	83,68	89,78
PA	93,15	92,7	83,83	89,89

Несколько меньшие значения полноты и точности для продуктов и событий объясняются недостаточной сбалансированностью исходного корпуса: продуктов и событий в текстах встречается гораздо меньше, чем геообъектов, физических и юридических лиц (см. Табл. 1).

Несмотря на то, что важным пунктом нашего исследования было какое-либо отсутствие предобработки текста, мы провели тест с нормализованным текстом, в котором все слова приведены к словарной форме. Нормализация в среднем не меняет результаты (см. Табл. 5).

Табл. 5. Средние значения полноты и точности при определении трех и пяти типов сущностей при нормализации слов

	Точность, %	**Полнота, %**	**F1, %**
Пять типов	84,76	89,67	87,06
Три типа	90,52	89,97	90,22

Результаты, полученные с помощью метода CRF, сравнивались с другими методами на том же корпусе (см. Табл. 6 и 7): MEMM и обычным словарным методом (обозначен Dict). MEMM был выбран, поскольку, как и CRF, принадлежит к дискриминативным методам, в отличие от генеративного HMM. Словарь для словарного метода был получен из обучающей выборки. Расчеты проводились как для трех, так и для пяти типов сущностей.

Табл. 6. Сравнительные результаты полноты и точности при определении пяти типов сущностей, полученные различными методами

	Dict	**MEMM**	**CRF**
Точность, %	90,35	89,08	84,91
Полнота, %	46,64	72,63	89,65
F1, %	59,39	79,89	**87,18**

Табл. 7. Сравнительные результаты полноты и точности при определении трех типов сущностей, полученные различными методами

	Dict	**MEMM**	**CRF**
Точность, %	94,04	93,26	90,01
Полнота, %	55,37	75,80	89,81
F1, %	69,60	83,59	**89,89**

Как и следовало ожидать, простая словарная разметка дает высокую точность (около 90%) при низкой полноте (47%). Следует заметить, что точность словарного метода не достигает 100%, т. к. одна и та же сущность может относиться к разным типам. Например, в предложении «*Кремль дал понять Киеву, что...*» *Кремль* и *Киев* выполняют функцию юридического лица, но не географического названия.

Наилучшие результаты для метода MEMM показал метод оптимизации Averaged Perceptron. Как видно из таблиц, результаты сравнительных методов уступают по F1-мере методу CRF.

4.2. Сентимент-анализ

Результаты тестирования показали, что наилучшую точность показывают, опять же, оптимизационные методы Averaged Perceptron и Passive Aggressive, причем для трех и четырех-граммных построений результаты получились примерно одинаковыми с небольшим перевесом у четырех-граммного окна. В Табл. 8 приведена точность полученной четырех-граммной модели для метода AP (Полнота всюду составила 100%, поскольку объекты тональности были заданы заранее.)

Табл. 8. Точность определения класса объекта тональности методом CRF

Класс	Точность, %
Негативный	84,44
Позитивный	84,89
Нейтральный	90,93
Среднее	**86,75**

Несмотря на то, что самым популярным статистическим методом сентимент-анализа текста является метод SVM [Liu 2010], мы для сравнения результатов тестирования использовали метод, основанный на правилах. На это у нас было две причины:

1. SVM классифицирует документы (предложения) относительно их полярности без какой-либо привязки к объекту тональности. В нашем алгоритме предполагается, что полярность жестко связана именно с объектом тональности (объект тональности был одним из факторов CRF[4]).
2. в методе, основанном на правилах, не только задается объект, относительно которого определяется тональность, но и используется тот же набор тональных словарей и шифтеров, *придающих лингвистический смысл результату.*

Сравнение с результатами ручного тестирования метода сентимент-анализа, основанного на правилах (Табл. 9) [Pazelskaya 2011], показывают более высокую степень точности CRF.

---

[4] В принципе, метод SVM позволяет сделать привязку к объекту мониторинга, но это скорее искусственный прием, связанный с особой компоновкой векторов, а не лексико-семантическими характеристиками тональности.

Табл. 9. Качество сентимент-анализа, основанного на правилах

	СМИ	Блогосфера
Точность, %	86,66	80,42
Полнота, %	82,37	68,88
F1, %	84,46	**74,20**

Это может быть вызвано разными причинами, одна из которых та, что обучающий корпус составлялся экспертами вручную, поэтому лишен разного рода «кривых» текстов с грубыми нарушениями грамматики и синтаксиса, в то время как модель сентимент-анализа на правилах тестируется на реальных текстах. Другой причиной повышения качества мог послужить удобный формат твитов: короткие сообщения не более 140 символов не имеющие сложной иерархической системы, как, например, блоги или форумы.

4.3. Part-of-Speech tagger

Для POS-tagger'а использовались те же оптимизационные алгоритмы, которые перечислены в разделе 3. Поскольку мы не ставили себе целью исследовать качество каждого из этих алгоритмов, то приведем только наилучший результат (Табл. 10). Он был получен с помощью алгоритма Passive Aggressive (вторым по качеству был метод Averaged Perceptron) на триграммах, с использованием «хвоста» слова длины три и графематических характеристик, упоминаемых в п. 3.1. Отметим, что упомянутый в [Manning 2011] барьер в 97% правильно классифицированных частей речи в нашем случае практически достигается: 96,7% (см. Табл. 12).

Табл. 10. Оценка результата частеречной классификации

	Точность, %	Полнота, %	F1, %
Биграммы, иных параметров нет	90,70	86,66	88,22
Триграммы, иных параметров нет	88,14	85,99	86,91
Триграммы + «хвосты» длины 3	93,79	92,47	93,10
«Хвосты» длины 3, без n-грамм	76,32	73,29	74,28
Триграммы + «хвосты» длины 3 + графематические характеристики	94,95	93,43	**94,14**

Для наилучшего случая приводится таблица (Табл. 11) с показателями качества классификации для каждого выделяемого частеречного класса.

Табл. 11. Результаты классификации для каждой выделяемой части речи

Часть речи	Относительная встречаемость частеречного класса, %	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	0,64	90,27	89,22	89,74
Числительное, записанное цифрами	1,56	92,80	94,78	93,78
Местоимение-существительное (личное)	1,20	99,31	99,84	99,57
Остальные 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

Чтобы сопоставить качество CRF-тэггера, были использованы два другие инструмента. Основу Стенфордского тэггера[5] составляет метод максимальной энтропии [Toutanova 2000, Manning 2011]. В свою очередь, инструмент TreeTagger (представленный еще в [Schmid 1994]) использует марковские модели и деревья решений для оценки вероятности перехода между состояниями. Обученная модель для русского языка была получена Сергеем Шаровым и находится в открытом доступе[6].

В таблице 12 приводится сравнения качества методов[7]. Под Accuracy в данном случае понимается процент правильно классифицированных слов от объема тестового корпуса.

---

5 http://nlp.stanford.edu/software/tagger.shtml

6 http://corpus.leeds.ac.uk/mocky/russian.par.gz

7 Трудность сопоставления результатов состояла в том, что для TreeTagger'а мы использовали заранее обученные модели с заданными частеречными классами, которые не полностью совпали с классами нашей разметки. Таким образом, в сравнительную таблицу попали только результаты по совпавшим классам. В случае Стенфордского тэггера проверялось пересечение по всем классам, поскольку Стенфордский тэггер обучался на том же корпусе, что и CRF-тэггер.

Табл. 12. Значение F1-меры для трех систем

	Accuracy, %
Stanford	79,39
TreeTagger	93,33
CRF	**96,75**

5. Выводы

В статье показана применимость CRF-метода в обработке текста на русском языке на примере задач выделения именнованных сущностей, сентимент-анализа коротких высказываний, частеречной классификации. Полученные результаты сравниваются с данными, полученными с помощью других подходов к рассматриваемым задачам. Как показывают результаты тестов, метод условных случайных полей (CRF) может составить существенную конкуренцию другим статистическим методам, используемым при лингвистической обработке текста.

Литература

1. *Bishop. Ch.* Pattern Recognition and Machine Learning. Springer. 2006.
2. *Brants, Th.* 2000. TnT — A Statistical Part-of-Speech Tagger. «6th Applied Natural Language Processing Conference»
3. *Choi Y., Cardie Cl., Riloff E., Patwardhan S.* Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. HLT ‹05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Pages 355–362. 2005.
4. *Collins M.* "Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). 1–8. 2002.
5. *Dingare Sh., Finkel J., Nissim M., Manning Ch., Grover C.* A System For Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations. Comparative and Functional Genomics, Volume 6 (2005). Issue 1–2. Pages 77–85. 2004.
6. *Giménez J., Márquez L.* 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC›04). Lisbon, Portugal.
7. *Finkel J., Dingare Sh., Manning Ch.D., Nissim M., Alex B., Grover C.* Exploring deep knowledge resources in biomedical name recognition. JNLPBA 04 Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Pages 96–99. 2004.
8. *Galen A. and Jianfeng G.* "Scalable training of L1-regularized log-linear models". Proceedings of the 24th International Conference on Machine Learning (ICML 2007). 33–40. 2007.
9. *Klinger R., Tomanek K.* Classical Probabilistic Models and Conditional Random Fields. Algorithm Engineering Report TR07–2-013. Department of Computer Science. Dortmund University of Technology. December 2007. ISSN 1864–4503.
10. *Koby C., Ofer D., Joseph K., Sh. Shalev-Shwartz. and Singer.* "Yoram Online Passive-Aggressive Algorithms". Journal of Machine Learning Research. 7. Mar. Pages 551–585. 2006.
11. *Lafferty J., McCallum A., Pereira F.* "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". Proceedings of the 18th International Conference on Machine Learning. Pages 282–289. 2001.
12. *Lavergne T., Allauzen A., Yvon F.* "From n-gram based to CRF-based translation models" EMNLP'11, 6th workshop on statistical machine translation (WMT'11), Edinburgh, UK, July 2011
13. *Liu Bing.* "Sentiment Analysis and Subjectivity". Handbook of Natural Language Processing. Second Edition. (editors: N. Indurkhya and F. J. Damerau). 2010.
14. *Manning C. D.* 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander Gelbukh (ed.), Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I. Lecture Notes in Computer Science 6608, pp. 171–189. Springer.

15. *Mao Yi., Lebanon G.* Isotonic Conditional Random Fields and Local Sentiment Flow. In proceeding of: Advances in Neural Information Processing Systems 19. Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems. Vancouver. British Columbia. Canada. December 4–7. 2006
16. *McCallum A., Li W.* Early results for named entity recognition with conditional random fields. feature induction and web-enhanced lexicons. In Seventh Conference on Natural Language Learning (CoNLL). 2003.
17. *McCallum A. and Nigam K.* "A Comparison of Event Models for Naive Bayes Text Classification". In AAAI/ICML-98 Workshop on Learning for Text Categorization. pp. 41–48. Technical Report WS-98-05. AAAI Press. 1998.
18. *McCallum A., Wellner B.* Conditional models of identity uncertainty with application to noun coreference. In Lawrence K. Saul. Y. Weiss. and L. Bottou. editors. Advances in Neural Information Processing Systems 17. Pages 905–912. MIT Press. Cambridge. MA. 2005.
19. *McCallum A., Freitag D. and Pereira F.* "Maximum entropy markov models for information extraction and segmentation." ICML-2000. pp. 591–599.
20. *Mejer A. and Crammer K.* "Confidence in Structured-Prediction using Confidence-Weighted Models". Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010). Pages 971–981. 2010.
21. *Nocedal.* "Updating Quasi-Newton Matrices with Limited Storage". Mathematics of Computation Jorge. 35. 151. Pages 773–782. 1980.
22. *Pazelskaya A., Solovyev A.* A Method of Sentiment analysis of Russian Text." Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011". Bekasovo. 2011. pp. 510–523.
23. *Rabiner, L. R.* A Tutorial on Hidden Markov Models andSelected Applications in Speech Recognition. In: Proceedings of the IEEE 77 (1989). No. 2. pp. 257–286.
24. *Ratinov L., Roth D.* Design Challenges and Misconceptions in Named Entity Recognition. CoNLL'09 Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Pages 147–155. 2009.
25. *Sadamitsu K., Sekine S., Yamamoto M.* Sentiment analysis based on probabilistic models using inter-sentence information. International Conference on Language Resources and Evaluation. 2008.
26. *Sarawagi S.* Information extraction. Foundations and Trends in Databases. 1(3). 2008.
27. *Schmid H.* Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.
28. *Sha F., Pereira F.* Shallow Parsing with Conditional Random Fields. NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology — Volume 1, pp. 134–141. 2003.
29. *Shalev-Shwartz Sh., Singer Y., and Srebro N.* "Pegasos: Primal Estimated sub-GrAdient SOlver for SVM". Proceedings of the 24th International Conference on Machine Learning (ICML 2007). Pages 807–814. 2007.

30. *Serge Sharoff, Joakim Nivre,* (2011) The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Dialog 2011.
31. *Singla P., Domingos P.* Discriminative training of Markov logic networks. In Proceedings of the Twentieth National Conference on Artificial Intelligence, pp. 868–873. Pittsburgh. PA. 2005. AAAI Press.
32. *Solovyev A. N., Antonova A. Ju., Pazelskaya A. G.* Using Sentiment-Analysis for Text Information Extraction. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012". Bekasovo. 2012. pp. 616–627.
33. *Stanford Named Entity Recognizer* // http://www-nlp.stanford.edu/software/CRF-NER.shtml
34. *Sutton C.* Conditional probabilistic context-free grammars. Master's thesis. University of Massachusetts. 2004.
35. *Sutton C., McCallum A.* Introduction to Conditional Random Fields for Relational Learning. MIT Press. 2006.
36. *K. Toutanova and C. D. Manning. 2000.* Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63–70
37. *Kiselev S.* Sistemy «Analiticheskij kur'er» i X-Files — osnova tehnologii izvlechenija znanij tekstov iz proizvol'nyh istochnikov ["Analytical Courier" and X-Files Systems — mining data from various sources] // Biznes i bezopasnost' v Rossii [Business and Security in Russia]. 2007. — № 48. c. 102–106.
38. *Ljashevskaja O., Astaf'eva I., Bonch-Osmolovskaja A., Garejshina A., Grishina Ju., D'jachkov V., Ionov M., Koroleva A., Kudrinskij M., Litjagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., Koval' S.* NLP Evaluation: Russian Morphological Parsers [Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskie parsery russkogo jazyka] // Komp›juternaja lingvistika i intellektual'nye tehnologii [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"] (Bekasovo, May 26–30, 2010). Moscow, RSUH, 2010
39. *Nikolenko S.* Skrytye markovskie modeli [Hidden Markov Models]. ITMO, 2006.