Метод условных случайных полей в задачах обработки русскоязычных текстов

А.Ю. Антонова
НИУ ВШЭ, Москва, Россия
a-antonova@list.ru

А.Н. Соловьев
СПбГУ, Санкт-Петербург, Россия
lechat1@mail.ru

Аннотация

Работа посвящена исследованию применимости 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.

1. Введение

Статистические методы для обработки текста традиционно применяются в тех случаях, когда большой объем получаемой информации требует быстрой и качественной обработки. Традиционная ситуация: обработка "на лету"непрерывно поступающих из Интернета данных и представление заказчику результата анализа новостных лент, блогов, сообщений в Твиттере, соцсетях.

Применимость к русскому языку метода классификации, использующего языконезависимые параметры для построения статистических языковых моделей, вообще говоря, не вызывает сомнений. Авторы хотели показать, что при использовании простейших эвристик метод условных случайных полей (CRF) для поставленной задачи дает уровень качества сопоставимый/превышающий уровень традиционных методов. Для иллюстрации этого утверждения авторы рассмотрели три популярные задачи: частеречное тэгирование, выделение именованных сущностей и анализ тональности коротких сообщений относительно объекта сообщения. В каждом случае материалом являлся плоский текст, так как привлечение лингвистических ресурсов неизбежно сказывается на скорости обработки данных.

2. Метод CRF. Формальное описание

Метод CRF, впервые предложенный в [13], имеет двух непосредственных предшественников, от каждого из которых он унаследовал часть свойств. Прежде всего, это метод скрытых марковских моделей (HMM) [4, 8, 16], которые успешно используются для моделирования последовательностей, а также метод моделей максимальной энтропии (MaxEnt) [7]. CRF рассматривает условное распределение $(y \mid x)$ последовательности меток $y \in Y$, где вектор $x \in X$ состоит из наблюдаемых элементов, и наряду с MaxEnt, принадлежит к категории дискриминативных методов. Из наблюдаемых и выходных элементов конструируется набор бинарных функций-признаков (feature functions, potential functions, factors), которые могут задаваться произвольно, включать в себя любое количество элементов. Например,

$$f_i(x,y) = \begin{cases} 1, & \text{если } y =< GEO >,\ y \text{ начинается} \\ & \text{с большой буквы, } x\text{="улица";} \\ 0, & \text{иначе.} \end{cases}$$

На графе каждая такая признак-функция соответствует клике.

Условным случайным полем называется распределение вида:

$$p(y \mid x) = \frac{1}{Z(X)} \prod_k exp(\sum_k \lambda_k f_k(y_t, y_{t-1}, x_t)), \qquad (1)$$

где f – признак-функция, λ - множитель Лагранжа, Z – коэффициент нормализации, сумма по всем $y \in Y$. Вычисление модели $p^*(y \mid x)$ решается как оптимизационная задача с заданными ограничениями [12] (разность между наблюдением и его оценкой должна быть нулевой и выполняться условие $\sum_{y \in Y} p(y \mid x) = 1$ по всем $x \in X$). На каждой итерации пересчитываются множители Лагранжа (аналогично с MaxEnt), вычисление проводится с использованием традиционных алгоритмов “forward-backward” и Витерби (точнее, их модификаций).

Среди преимуществ CRF называют отсутствие презумпции условной независимости наблюдаемых переменных, а также частичное устранение т.н. label bias problem – ситуации, когда преимущество получают состояния с меньшим количеством переходов, поскольку строится единое распределение вероятностей и нормализация (коэффициент $Z(X)$ из формулы) производится в целом, а не в рамках отдельного состояния.

3. Метод CRF для обработки текста

Метод CRF для обработки текста применим всюду, где задача сводится к задаче классификации: приписывание словам (и их сочетаниям варьируемой длины) тэгов из ограниченного набора: частей речи, имен сущностей, тональности объекта, семантических ролей и т.д. Ниже будут подробнее рассмотрены три первые задачи. С 2001-го года было опубликовано много работ, исследующих применимость метода для каждой из перечисленных (а также и других) задач для европейских и азиатских языков. Насколько нам известно, для русского языка наиболее часто применяемыми по-прежнему остаются методы SVM, HMM, MEMM. Это побудило авторов опубликовать опыт применения метода CRF в разработках компании "Ай-Теко"[1]. В качестве основы был использован открытый исходный код проекта CRF-Suite (http://www.chokkan.org/software/crfsuite/), с помощью которого были реализованы три инструмента. Ниже описываются некоторые детали реализации, приводится сравнение результата с результатами разработанных ранее алгоритмов (более подробно см. [1]). Реализация метода позволяет использовать несколько алгоритмов оптимизации параметров, мы упомянем лишь те два, которые показали наилучший результат для каждой задачи: это Averaged Perceptron (AP) [9] и Passive-Aggressive (PA) [10]. Оба эти алгоритма родственны алгоритму обучения перцептрона, и могут самостоятельно работать для задачи классификации или регрессии. Алгоритм PA строит разделяющую гиперплоскость, в качестве ограничения выступает условие на поворот гиперплоскости на каждой следующей итерации, который предполагается минимальным. Эвристика для AP еще проще: после того, как с помощью алгоритма Витерби [20] найдена оптимальная последовательность меток, на каждой итерации параметры пересчитываются, так что к прежнему значению добавляется разница между значением фактор-функции на эталонной последовательности и на той, которая получена с помощью алгоритма Витерби. Вопрос о том, почему эти два алгоритма оказались самыми успешными, мог бы стать темой для отдельного исследования.

3.1. Частеречное тэгирование

Определение частей речи (ЧР) – следующий этап обработки текста после его очистки и сегментирования. Качество статистических тэггеров для английского языка [8] преодолело порог в 97% правильно определяемых частей речи [14]. Для русского языка подобный результат сумели достичь парсеры, основанные на правилах [3], статистический парсер подобного качества (97.3% правильно определенных тэгов) описывается в [18]).

Для обучения и тестирования CRF-моделей был собран корпус новостных лент (объемом 2.2 млн словоформ – обучающий, 0.67 млн – тестовый) и размечен с помощью морфологического модуля системы "Аналитический курьер"[2]. (Особенности этой системы определили набор частеречных тэгов, а также характер наиболее частотных ошибок). Так в разные ЧР выделялись глаголы, причастия и деепричастия, поскольку на уровне синтаксического анализа им соответствуют разные типы зависимостей, отдельно выделялись личные местоимения, был введен тип “остальное”, предназначенный для экзотических случаев. В качестве параметров для обучения использовались триграммы словоформ, частеречные триграммы, характеристики написания слова (большая буква, дефис, цифры и т.п.), последние три буквы слова (признак, который мы использовали вместо окончания). Таблица 1 иллюстрирует статистику выделения каждой ЧР по отдельности.

Качество CRF-тэггера сопоставлялось с качеством Стенфордского тэггера и TreeTagger’a. Основу Стенфордского тэггера (http://nlp.stanford.edu/software/tagger.shtml) составляет метод максимальной энтропии [19]. В свою очередь, инструмент TreeTagger (представленный еще в 1994-м году [17]) использует марковские модели и деревья решений для оценки вероят-

Таблица 1. Качество выделения каждой части речи с помощью CRF-теггера

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

Таблица 2. Сравнение тэггеров на одном корпусе

	Accuracy, %
Stanford	79.39
TreeTagger	93.33
CRF	**96.75**

ности перехода между состояниями. Обученная модель для русского языка была получена Сергеем Шаровым и находится в открытом доступе (http://corpus.leeds.ac.uk/mocky/russian.par.gz).

В Табл. 2 приводится сравнение качества методов. Под Accuracy в данном случае понимается процент правильно классифицированных слов от объема тестового корпуса. Следует отметить, что трудность сопоставления результатов состояла в том, что для TreeTagger'а мы использовали заранее обученные модели с заданными частеречными классами, которые не полностью соответствуют классам нашей разметки. Таким образом, в сравнительную таблицу попали только результаты по совпавшим классам. В случае Стенфордского тэггера проверялось пересечение по всем классам (функционал инструмента позволяет задавать список классов вручную), он обучался на том же корпусе, что и CRF-тэггер.

Оговоримся, что 96.75% – процент правильно определенных слов. Для CRF-тэггера F-мера, усредненная по каждой части речи (см. Табл. 1), составляет 94.14% (от 82.48% для причастий и до 100% для знаков препинания, которые в системах обработки традиционно выделяются в отдельную "часть речи").

3.2. Выделение именованных сущностей

Метод CRF был применен для задачи распознавания в тексте именованных сущностей. Мы ограничились пятью традиционно выделяемыми типами: физические лица, юридические лица, названия географических объектов, названия продуктов, события, бренды. Для обучения модели вручную был размечен корпус новостных сообщений на разные темы, всего порядка 1.5 млн словоформ (около 71 тыс. предложений). Корпус формировался так, чтобы каждое предложение содержало по крайней мере одну именованную сущность. В итоге 90% всех выделенных сущностей относится к первым трем типам и только 10% осталось на долю событий и продуктов (чем, по-видимому, и объясняется невысокий результат для этих двух типов). В качестве признаков для обучения использовались n-граммы и упоминавшиеся графематические признаки написания. (Мы пробовали также нормализовывать словоформы, но это не дало значимого вклада.)

Лучшие результаты показывали методы оптимизации AP и PA на триграммной модели (см. Табл. 3).

Таблица 3. Качество распознавания пяти типов сущностей (оптимизация с помощью AP).

NAME	GEO	ORG	PROD	EVENT	AVG
точность, %					
92.58	91.22	85.23	76.09	79.43	84.91
полнота, %					
94.23	95.01	88.02	87.14	83.84	89.65
F1-мера					
93.39	93.08	86.60	81.24	81.57	87.18

Здесь точность и полнота рассчитывались по следующим формулам:

$$Precision = \frac{A}{A+C+D} * 100\%; \quad (2)$$

$$Recall = \frac{A}{A+B+C} * 100\%, \quad (3)$$

где каждая буква соответствует количеству
A – верных срабатываний системы;
B – пропусков;
C – случаев типизации нетипизированной сущности;
D – неверных определений типа сущности.

Этот результат сравнили с результатами, полученными словарным методом и методом MEMM (Табл. 4). Словарь был составлен на основе размеченного корпуса (его части, предназначенной для обучения) и, очевидно, показал высокий результат точности при низкой полноте (за счет того, что в тестовом корпусе содержались сущности, не вошедшие в словарь). То, что и точность не стопроцентная, объясняется тем, что одна и та же сущность может быть отнесена сразу к двум типам. Например, в предложении *"Кремль дал понять Киеву, что..." Кремль* и *Киев* выполняют функцию юридического лица, но не географического названия, как это могло бы быть в другом контексте. Метод MEMM был выбран, поскольку он больше всего похож на CRF и, подобно ему, относится к дискриминативным.

3.3. Сентимент-анализ

Метод CRF для сентимент-анализа применялся в рамках задачи определения тональности (под)предложения относительно объекта тональности. Объектом тональности назовем слово (в об-

Таблица 4. Сравнение качества распознавания пяти типов сущностей с помощью разных методов.

	Dict	MEMM	CRF
Точность, %	90.35	89.08	84.91
Полнота, %	46.64	72.63	89.65
F1-мера, %	59.39	79.89	**87.18**

щем случае, именную группу), относительно которого определяется эмоциональная окраска предложения. Объект тональности, в зависимости от решаемой задачи, или выделяется автоматически, или задается вручную. В нашем случае список объектов тональности формировался априори.

Материалом для обучения послужил корпус из порядка 20 тыс. русскоязычных твит-сообщений, в которых вручную было выделено около 21 тыс. заранее заданных объектов (для простоты однословных). Корпус не содержал бессмысленных или бессодержательных текстов, благодаря чему, вообще говоря, результат оказался довольно высоким. Заданным объектам приписывалась тональность: нейтральная (ок. 40% всех объектов в корпусе), позитивная или негативная (поровну). Для обучения модели использовался признак вхождения слова/коллокации из предложения в словари тональной лексики. Тональные словари были составлены при разработке системы сентимент-анализа, основанного на правилах [5]. Они представляют собой разделенные по частям речи словари позитивных и негативных слов и коллокаций, в том числе глагольных, усиливающих и нейтральных (подробнее в [5]). Имевшийся набор словарей был пополнен словарями инверторов и шифтеров, а также списками словоизменений (для работы с плоским текстом). Результат тестирования модели показал хорошую точность, в Таблице 5 приведена точность для оптимизационного метода AP и 4-граммного окна.

Другие инструменты на этом же корпусе не тестировались (это не входило в планы авторов). Уровень результата можно *приблизительно* оценить, сравнив с имеющимися недавними публикациями, посвященными сентимент-анализу твиттер-сообщений. Полученный нами результат превосходит классификаторы для Twitter-сообщений, основанных на:

- SVM [6] – точность достигает 80% на вычищенных текстах, а в [11] – 68.3% на предобработанных текстах,
- MaxEnt [15] – 64%, бинарная классификация, зашумленные тексты,
- "наивном Байесе"[21] – 59% точности, четыре категории, зашумленные тексты.

Таблица 5. Точность определения класса объекта тональности методом CRF.

Класс	**Точность, %**
негативный	84.44
позитивный	84.89
нейтральный	90.93
среднее	**86.75**

Приведем для сравнения и данные о качестве системы сентимент-анализа, основанной на правилах [5], тональными словарями которой мы воспользовались для CRF-классификатора. Эта система тестировалась на “реальных” текстах блогов и показала 74.2% F-меры.

4. Выводы

В статье иллюстрируется применимость и конкурентоспособность CRF-метода в обработке текста на русском языке на примере задач выделения именованных сущностей, сентимент-анализа коротких высказываний, частеречной классификации. Результаты тестирования метода условных случайных полей на русскоязычном материале позволяют сделать вывод о том, что

- данный метод применим для различных видов обработки текста;
- выбором функций-признаков можно достичь высокого качества результата;
- метод является достаточно гибким в выборе функций-признаков и не требует условной независимости переменных.

Список литературы

[1] А. Антонова, А. Соловьев. *Использование метода условных случайных полей для обработки текста на русском языке.* Труды международной конференции "Диалог 2013 в печати.

[2] С. Киселев. *Системы "Аналитический курьер"и X-Files — основа технологии извлечения знаний текстов из произвольных источников.* Бизнес и безопасность в России. 2007. - № 48. с. 102–106

[3] О. Ляшевская и др. *Оценка методов автоматического анализа текста: морфологические парсеры русского языка* В кн.: Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог"(2010). Москва: РГГУ, 2010. С. 318–326

[4] С. Николенко. *Скрытые марковские модели.* ИТМО, 2006.

[5] А. Пазельская, А. Соловьев. *Метод определения эмоций в текстах на русском языке*: Компьютерная лингвистика и интеллектуальные технологии: "Диалог-2011". Вып. 10 (17). – М.: Изд-во РГГУ, 2011. С. 510-522.

[6] L. Barbosa, J. Feng. 2010. *Robust sentiment detection on twitter from biased and noisy data.* Proceedings of Coling.

[7] A.L. Berger, S.A.D. Pietra, V.J.D. Pietra (1996). *A maximum entropy approach to natural language processing.* Computational Linguistics, 22, 39–71.

[8] Th. Brants. 2000. *TnT - A Statistical Part-of-Speech Tagger.* "6th Applied Natural Language Processing Conference".

[9] M. Collins. *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.

[10] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, Y. Singer. *Online Passive-Aggressive Algorithms.* Journal of Machine Learning Research. 7. Mar. 551–585. 2006.

[11] L. Jiang et al, *Target-dependent Twitter Sentiment Classification*, ACL 2011.

[12] R. Klinger, K. Tomanek. *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.

[13] J. Lafferty, A. McCallum, F. Pereira. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* Proceedings of the 18th International Conference on Machine Learning. 282-289. 2001.

[14] C.D. Manning. 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] R. Parikh, M. Movassate. *Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques*, CS224N Final Report, 2009.

[16] L.R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* In: Proceedings of the IEEE 77 (1989). No. 2. pp. 257–286.

[17] H. Schmid. *Probabilistic Part-of-Speech Tagging Using Decision Trees.* Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.

[18] S. Sharoff, J. Nivre, (2011) *The proper place of men and machines in language technology*: Processing Russian without any linguistic knowledge. Dialog 2011.

[19] K. Toutanova, 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.

[20] A. Viterbi (April 1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". IEEE Transactions on Information Theory 13 (2): 260–269.

[21] H. Wang, D. Can, A. Kazemzadeh, F. Bar and Sh. Narayanan. *A System for Real-Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle.* Proceedings of ACL 2012.