Поиск упоминаний лиц в новостных текстах с использованием модели условных случайных полей

© А.В. Подобряев
ИПС имени А.К.Айламазяна РАН,
г.Переславль-Залесский
alex@alex.botik.ru

Аннотация

В статье описывается подход к выявлению упоминаний людей (в форме личных имен) в новостных текстах на русском языке при помощи модели условного случайного поля с использованием признаков, основанных на имеющемся ресурсе знаний о данной предметной области, без использования лексических признаков.

1 Введение

Извлечение именованных сущностей из текста на естественном языке важная подзадача во многих задачах автоматического извлечения информации из текстов, например, в задаче извлечения событий (это могут быть назначения и отставки лиц, слияния и поглощения компаний и т.п.). Под именованной сущностью мы понимаем объект, имеющий имя или название и относящийся к определенному типу. Возможные типы: имена лиц, названия организаций, названия геополитических единиц, названия химических веществ и т.п. В общем виде ставится задача обнаружения и классификации именованных сущностей. При этом использование только лишь словаря малоэффективно, т.к. важен контекст. Например, слово «Питер» в разных контекстах может означать имя человека или названия города. В настоящей работе мы ограничиваемся поиском имен лиц в новостных текстах на русском языке.

Для решения данной задачи мы используем метод машинного обучения, основанный на модели условного случайного поля (conditional random field, далее CRF). Это модель, в частности, предназначена для классификации элементов последовательностей, и учитывает структуру последовательности, в отличие от классических методов машинного обучения, которые предназначены для классификации каждого элемента последовательности в отдельности.

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

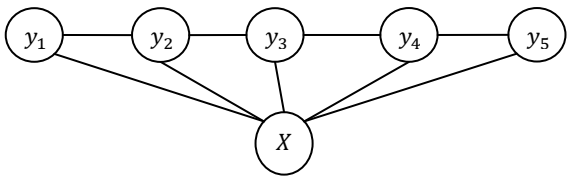
В нашем случае последовательностью является предложение, а ее элементами – токены, т.е. отдельные слова, знаки препинания и т.п. Действительно, для задачи обнаружения имен нельзя считать слова предложения независимыми, а каждое слово необходимо рассматривать в контексте. В тоже время достаточно далекие слова можно считать независимыми. Этому наблюдению отвечает свойство марковости CRF.

CRF успешно применяется для извлечения именованных сущностей из текстов на английском языке, а также для определения фраз, частей речи, см., например, [1], [2]. В этих работах используются лексические признаки, что требует большой обучающей выборки. Если же встреченное при тестировании слово отсутствует в обучающей выборке, то применяются специальные приближенные или эвристические методы, например, основанные на рассмотрении различных окончаний данного слова [3].

Здесь приводится описание подхода, в котором лексические признаки отсутствуют, но используются признаки, основанные на имеющемся ресурсе знаний о данной предметной области. Приводятся экспериментальные результаты для коллекции новостных текстов на русском языке.

2 Условные случайные поля

CRF – это графическая модель, предназначенная для оценки условных вероятностей событий, соответствующих вершинам некоторого графа Γ, при условии наблюдаемых данных. Пусть $X = \{x_1, \dots, x_N\}$ – последовательность наблюдаемых данных. В нашем случае это токены одного предложения. Пусть $Y = \{y_1, \dots, y_N\}$ – последовательность случайных величин, связанных с вершинами графа Γ.



В нашем случае графическая модель выглядит приведенном на рисунке образом, а случайные

величины y_i суть метки токенов, которые мы хотим научиться предсказывать.

Такой набор данных называется CRF, если для каждой вершины v графа Γ выполнено марковское условие

$$P(y_v|Y_{V\setminus\{v\}}, X) = P(y_v|Y_{O(v)}, X),$$

где V – множество вершин графа Γ, а $O(v)$ – множество соседних с v вершин в графе Γ. Т.е. метка зависит только от близкого контекста.

Тогда ([5], [6])

$$P(Y|X) = \frac{1}{Z(X)}\, exp\left(\sum_{c\in C} \lambda_c f_c(y_c, X)\right),$$

где $Z(X) -$ нормализующий множитель, C – множество всех клик графа Г, f_c – признаки, а λ_i – коэффициенты.

Эти коэффициенты подбираются в процессе обучения данной модели так, чтобы максимизировать логарифм функции правдоподобия на обучающем наборе X, Y:

$$\max_{\lambda_c}\left(\sum_{i=1}^{N} P(y_i \mid x_i) - \sum_{c\in C}\frac{\lambda_c^2}{2\,\sigma^2}\right),$$

где вычитаемое – регуляризационный член с настраиваемым параметром σ, нужный для того, чтобы избежать переобучения.

3 Система признаков

В пунктах 3.1 и 3.2 описаны признаки, вычисляемые для каждого конкретного токена и конкретной позиции в последовательности. Один признак формируется для каждого наблюдаемого признака (вычисленного по токену) и каждой метки. В пункте 3.3 описаны признаки, учитывающие контекст.

3.1 Признаки токена

Нами были использованы следующие признаки написания токена:

-использованный для написания алфавит,

-является ли первая буква прописной,

-есть ли прописные буквы внутри токена,

-является ли данный токен знаком препинания и каким,

-количество символов в токене.

Также использовался глобальный признак количество вхождений данного токена в текст, посчитанное по совпадению хотя бы одной из возможных канонических форм.

Кроме того, морфологический признак – часть речи, которая определялась с помощью HMM-алгоритма, описанного в работе [3]. Его обучение производилось на коллекции национального корпуса русского языка [4], содержащей порядка миллиона словоупотреблений.

Для слов, не входящих в тексты обучающего множества, используется алгоритм, основанный на анализе окончаний (в данном случае имеются в виду наборы нескольких последних букв слова, от 1 до 5). Для вычисления априорной вероятности части речи при условии данного слова суммируются с весами доли данной части речи в обучающем множестве для каждого из возможных окончаний. Эти веса являются несмещенными оценками дисперсии долей частей речи с данным окончанием. Поэтому, например, вклад окончания, для которого доли частей речи в обучающем множестве примерно одинаковы, будет меньше вклада окончания с большим разбросом долей частей речи в обучающем множестве.

3.2 Основанные на знаниях признаки

Можно выделить следующую группу признаков, основанных на знаниях.

1. Является ли данный токен аббревиатурой из словаря аббревиатур. (Постоянно пополняемый список аббревиатур и их расшифровок.)

2. Является ли данный токен возможным характерным префиксом фамилии (например, «д'», «де», «О'», «Тер-» и т.п.).

Имеет ли данный токен окончание, характерное для фамилии или отчества (например, «-ович», «-овна», «-енко» и т.п.).

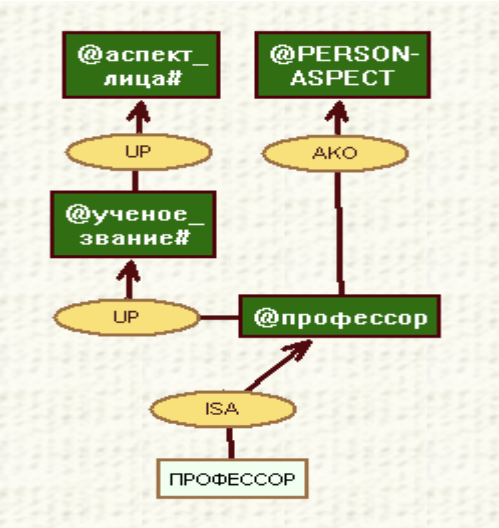
3. Является ли данный токен характерным словом, обычно находящимся перед фамилией («господин», «месье», «товарищ» и т.п.).

Средство вычисления последних двух признаков реализовано в виде конечного автомата, определяемого словарем (окончаний или характерных слов) с учетом морфологии.

4. Для описания слов, которые обозначают должности, профессии, роли и т.п., был использован признак, получаемый из ресурса знаний, структура которого описана в работе [7]. Ресурс знаний обеспечивает доступ к априорным знаниям о предметной области. Это модель знаний, которая содержит онтологическую информацию (классы сущностей, отношения между ними, атрибуты сущностей и отношений, таксономию сущностей и отношений) и фактографическую информацию (конкретные экземпляры сущностей и отношений между ними, конкретные их атрибуты). Модель знаний состоит объектов четырех типов: концепты, типы отношений, экземпляры концептов, экземпляры отношений. Первые два типа объектов отвечают за описание онтологической информации, а второе два типа за описание фактографической информации.

Токену соответствует экземпляр концепта, который экземпляром отношения ISA (is a) связан с некоторым концептом.

Например, токену «профессор» соответствует экземпляр концепта «профессор», связанный экземплярами отношений следующим образом.



Здесь экземпляр концепта профессор связан отношением «ISA» (is a, является экземпляром) с концептом «@профессор». Этот концепт связан отношением «AKO» (a kind of) с концептом «@PERSON_ASPECT» и путем из отношений «UP» (выше в какой-либо онтологической иерархии) с концептами «@ученое_звание#» и «@аспект_лица#».

В качестве признака берется наличие направленного пути от данного экземпляра концепта по отношениям типа «ISA», «AKO» и «UP» до концептов «@PERSON_ASPECT» и/или «@аспект_лица#».

3.3 Признаки контекста

Для описания контекста для каждого токена вычисляются вышеописанные признаки для его окрестности радиуса 2, т.е. для еще четырех токенов, из которых два стоят перед данным токеном, и два после данного токена.

Один признак формируется для каждого наблюдаемого признака (для каждой позиции) и каждой метки.

4 Полученные результаты

Описанный здесь подход был протестирован на корпусе, состоящим из 376 новостных текстов из коллекции “Persons-600” [11], взятых из лент информационных агентств. Всего имеется 3820 вхождений имен лиц в этих текстах, 93603 токена. Эти тексты были размечены аннотационной разметкой в полуавтоматическом режиме (т.е. разметка, сгенерированная основанным на правилах алгоритмом, была откорректирована человеком). Аннотационная разметка для использования CRF была переведена в разметку типа BIO (begin, inside, outside – маркеры токенов начала, продолжения имени и не имени, см. например, [2], где таким образом размечаются именные группы).

Множество текстов было разбито на два – обучающее (251 документ, 2501 вхождений имен лиц, 62251 токен) и тестовое (125 документов, 1319 вхождений имен лиц, 31352 токена). Кроме того, производилось скользящее обучение на всем множестве текстов. Т.е. из всего корпуса исключался один документ, на остальных производилось обучение, и полученная модель тестировалась на исключенном документе, затем документ возвращался в корпус, и такая процедура проводилась для всех документов. Известно (например, [8]), что средняя ошибка скользящего обучения дает несмещенную оценку вероятности ошибочной классификации.

При подсчете результатов имя лица считалось правильно извлеченным, если границы аннотации, полученной нашим алгоритмом, в точности совпадают с границами аннотации эталонной разметки.

Использовалась реализация условных случайных полей в библиотеке hCRF, версии 2.0 [9], [10].

Ниже приведены результаты (в процентах) для использованной CRF-модели.

	Тестовое множество	Обучающее множество	Скользящее обучение
Точность	86.26	90.63	90
Полнота	83.93	88.16	86.7
F1-мера	86.28	89.38	88.32

Работа выполнена при поддержке РФФИ, грант № 13-07-00307а.

Литература

[1] A. McCallum, W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the Seventh Conference on natural Language Learning (CoNLL-2003), Edmonton, Canada, 2003.

[2] F. Sha, F. Pereira. Shallow parsing with conditional random fields. In Proceedings of HLT-NAACL 2003, p. 134-141, Edmonton, May-June 2003.

[3] T. Brants. TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000, Seattle, 2000.

[4] Морфология. Национальный корпус русского языка, URL: http://www.ruscorpora.ru/corpora-morph.

[5] J. Lafferty, A. McCallum, F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proeedings of 18th International Conference on Machine Learning, p. 282-289, Morgan Kaufmann, San Francisco, 2001.

[6] C. Sutton, A. McCallum. An Introduction to Conditional Random Fields for Relational Learning, 2010. arXiv:1011.4088v1.

[7] Кормалев Д., Куршев Е., Сулейманова Е., Трофимов И. Технология извлечения информации из текстов, основанная на знаниях. Программные продукты и системы, 2009, № 2, с.62–66.

[8] M. Stone. Asymptotics for and against cross-validation. Biometrika 64(1), p. 29-35, 1977.

[9] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, T. Darrell. Hidden conditional random fields. In IEEE Transactions on Pettern Analysis and Machine intelligence, volume 29, issue 10, p. 1848-1852, October 2007.

[10] HCRF, URL: http://sourceforge.net/projects/hcrf

[11] Колекция “Persons-600”, Исследовательский центр искусственного интеллекта ИПС РАН, 2013, URL: http://ai-center.botik.ru/Airec/index.php/ru/collections/27-persons-600

Persons recognition using CRF model

Alexey V. Podobryaev

We study CRF-based approach to persons recognition in the Russian news texts. It contains some knowledge based features instead of lexical features.