ПРИМЕНЕНИЕ N-ГРАММ И ДРУГИХ СТАТИСТИК УРОВНЯ СИМВОЛОВ И СЛОВ ДЛЯ СЕМАНТИЧЕСКОЙ КЛАССИФИКАЦИИ НЕЗНАКОМЫХ СОБСТВЕННЫХ ИМЕН

Нехай И. В. (nekhayiv@gmail.com)

Кафедра распознавания изображений и обработки текста, ФИВТ МФТИ, Москва, Россия

Автоматическая семантическая классификация неизвестных собственных имен в текстах является значимой проблемой при разработке средств автоматического анализа и обработки текста. В данной работе мы исследуем возможности классификации часто используемых в русском языке собственных имен при помощи машинного обучения и побуквенных n-грамм, а также других признаков уровня букв и слов, таких как капитализация слов, общее число слов, наличие в имени аббревиатур, чисел и т.п. Мы используем только внутренние признаки, однако, поскольку использование внешних признаков ожидается в будущих работах, мы тестируем также сценарии, которые могут оказаться полезными при использовании внешних признаков. В работе показано, что при помощи простых признаков и средств классификации можно достичь достоверности 80–99 % для биномиальной классификации, 89–99 % при отделении одной категории от остальных и 85 % достоверности для задачи мультиномиальной классификации. При этом используется обучающая выборка, содержащая порядка 2000 примеров в каждой категории. Аналогичные результаты для английского языка были продемонстрированы в работе (Patel & Smarr, 2001), где применялся аналогичный подход.
Работа выполнена в рамках исследований, проводящихся под эгидой проекта Министерства Образования и Науки № 13.G25.31.0088.

Ключевые слова: n-граммы, семантическая классификация, имена собственные, именованные сущности

APPLICATION OF N-GRAMS AND OTHER LETTER- AND WORD-LEVEL STATISTICS TO SEMANTIC CLASSIFICATION OF UNKNOWN PROPER NOUNS

Nekhay I. V. (nekhayiv@gmail.com)

Department of image recognition and text processing,
DIHT MIPT, Moscow, Russian Federation

Automatic semantic classification of unknown proper nouns is a significant problem in the field of automatic text analysis. We investigate the ability to classify proper noun phrases frequent in Russian language using machine learning approaches and n-gram statistics, as well as other letter- and word-level feature statistics such as word capitalization, number of words, presence of abbreviations and numbers, etc. We use only internal information, but, as we expect to use external information in future, we test scenarios that can arise when the former is utilized for classification. We show that such simple features allow to achieve accuracies up to 80–99,9 % in 1–1 trials, 89–99 % in 1-other trials and 85 % accuracy in 5-way trial in 5 pre-selected categories, using training set of about 2000 examples in each category. These results come very close to results of (Patel & Smarr, 2001), achieved by a similar system designed to classify English proper nouns.

Key words: n-grams, semantic classification, proper nouns, named entities

Введение

При разработке систем автоматического анализа и перевода текстов, использующих словари, необходимо решать проблемы, связанные с обработкой собственных имён. Не представляется возможности исчерпывающе перечислить в словарях собственные имена, такие как антропонимы, топонимы, названия компаний, фильмов, лекарств. Это вызвано следующими двумя причинами:

1. на текущий момент существует огромное число подобных имён;
2. непрерывно порождаются новые имена.

Таким образом, неизвестные системе разбора текста слова необходимо обрабатывать автоматически. Существует ряд известных процедур, выполняемых при такой обработке:

1. выделение собственных имён в тексте, в частности:
   1.1. детектирование присутствия имени и
   1.2. определение границ имён;
2. приведение к нормальной форме;
3. семантическая классификация имён.

Пример реализации процедуры 1 можно найти в (Baluja, 2000), процедуры 2 — в (Mikheev, 1997).

В данной работе мы рассматриваем вариант реализации процедуры семантической классификации имен, т. е. отнесения их к какой-либо из заранее заданных нами категорий. При этом предполагается, что предшествующие процедуры уже выполнены, и входными данными для системы классификации являются выделенные и приведённые к нормальной форме имена.

Основной целью данной работы является получение программы-классификатора, использующего внутренние признаки имён собственных для классификации их по следующим категориям: антропонимы, названия городов, фильмов, компаний, лекарств. В статье (Patel & Smarr, 2001) описана сходная реализация классификатора англоязычных имён, использующего 4 категории, поэтому дополнительной целью является сравнение характеристик полученных классификаторов. Отметим особо, что система классификации исследовалась на примерах, широко распространённых в современном русском языке.

1. Основные принципы

Классификатор разработан по принципу «обучения с учителем»: программе предоставляются «обучающий» набор имён, на основе которого вычисляются статистические параметры модели, которые в дальнейшем используются для классификации «неизвестных» имён.

В процессе разработки были решены следующие вопросы:

- составление набора «обучающих» данных;
- определение набора признаков, на основании которых осуществляются обучение и классификация;
- выбор алгоритма классификации;
- определение и вычисление метрик качества классификации;

Эти вопросы рассматриваются в последующих разделах.

2. Внешние и внутренние признаки

Как отмечается в (McDonald, 1993), для классификации собственных имен можно использовать два вида признаков.

1. Внешние признаки — морфосинтаксическая информация о словоизменении имени и его роли в тексте или предложении.
2. Внутренние признаки — лексическая информация, содержащаяся в самом имени. Примерами их для английского языка, согласно (Baluja, 2000), являются:
   a. капитализация слов;

б. написание слов заглавными буквами;

в. наличие чисел;

г. количество цифробуквенных последовательностей;

д. наличие слов, состоящих из одной буквы.

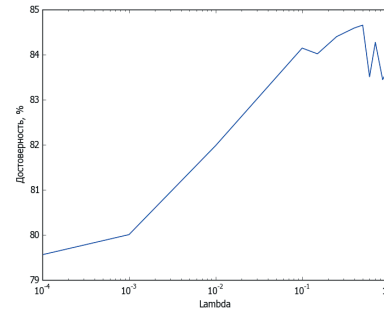

Рис. 1. Зависимость достоверности классификатора от значения параметра сглаживания λ, задача мультиномиальной классификации

В данной работе внешние признаки не принимаются во внимание, так как рассматриваются уже выделенные и лемматизированные имена, поэтому используются только внутренние признаки.

Основным внутренним признаком для классификации, как и в (Patel & Smarr, 2001), является статистика символьных n-грамм. На этапе обучения по принципу максимального правдоподобия эмпирически оцениваются вероятности $P(l_i|c_i,l_{i-k}\ldots l_{i-2}\ l_{i-1})$ того факта, что символ l_i в именах категории c встречается после последовательности символов $l_{i-k}\ldots l_{i-2}\ l_{i-1}$ длины k. На этапе классификации вероятность того, что имя w принадлежит категории c оценивается как

$$P(w|c) = \prod_{i=1\ldots len(w)} P(l_i|c_i,l_{i-k}\ldots l_{i-2}\ l_{i-1}) \qquad (1)$$

При этом значение k, определяющее глубину учитываемого контекста, является параметром и выбирается, исходя из максимизации достоверности классификации. Для корректного учёта статистики по начальным и конечным символам имени в начало имени приписывается k условных символов “$”, а в конец — один символ “#”.

Значения вероятности сглаживаются при помощи правила Лидстоуна:

$$\hat{P}(l_i|c, l_{i-k}\ldots l_{i-2}l_{i-1}) = \frac{N_{l_i}+\lambda}{N_{l_{i-k}\ldots l_{i-2}l_{i-1}}+N_l\cdot\lambda}, \qquad (2)$$

где N_{l_i} — число появлений символа l_i после последовательности $l_{i-k}\ldots l_{i-2}\ l_{i-1}$, $N_{l_{i-k}\ldots l_{i-2}\ l_{i-1}}$- число появлений этой последовательности, N_l — число уникальных символов во всех примерах, λ — параметр сглаживания.

Значение параметра λ выбрано равным 0.15, так как значительное число экспериментов показали, что это значение обеспечивает максимальную достоверность классификатора. Экспериментальная зависимость достоверности классификатора от значения λ приведена на рис. 1. Это значение также может выбираться при помощи кросс-валидации.

Остальные признаки были выбраны путём ручного анализа основных классов ошибок классификатора. В результате были получены следующие

признаки (в скобках **полужирным шрифтом** отмечены используемые в графиках сокращения):

- *бинарные, значения которых есть «истина» или «ложь»*:
  - капитализация всех слов имени (**ВсеКап**);
  - наличие чисел в имени (**СодЧисл**);
  - наличие аббревиатур (последовательностей длины 3 и более заглавных букв) (**СодАббр**);
  - наличие слов, содержащих заглавные буквы после строчных (пример: *НоваТЭК*) (**СлитКап**);
  - имя состоит из слов, разделённых дефисами (пример: *Ростов-на-Дону*) (**Дефисоид**);
  - имя полностью написано заглавными буквами (**ВсеЗагл**);
- *числовые, значения которых есть натуральные числа:*
  - число последовательностей цифр или букв (**ЧислоПослед**);
  - общее число символов (**ЧислоСимв**);
- *числовые, значения которых есть произвольные вещественные числа:*
  - средняя длина последовательности цифр или букв (**СредДлПослед**);
  - средняя длина слов (последовательностей букв) (**СредДлСлов**).

3. Алгоритм классификатора

В основе классификатора лежит машина опорных векторов (SVM), использующая в качестве ядра радиальную базисную функцию. Входными значениями признаков для неё являются:

- значения $\log_2 P(w|c)$ вероятностей принадлежности имени w классу c, вычисляемые по формуле (1);
- значения остальных признаков.

Заметим, что наш классификатор не использует при обучении априорные вероятности классифицируемых категорий. Это ослабляет ограничения на использование классификатора, т. е. полученные оценки справедливы для случая, когда об априорных вероятностях ничего не известно и они считаются одинаковыми, т. е. $P(c) = \frac{1}{N_c}$ для всех категорий c, где N_c — общее число всех категорий. В таком случае значения $P(w|c)$ можно использовать в качестве значений $P(w|c)$, поскольку $P(c)$ не зависит от c, и

$$argmax_c P(w|c)P(c) = argmax_c P(c|w) \tag{3}$$

Однако, если априорные вероятности категорий можно некоторым образом оценить, то они могут быть использованы путём следующего изменения формулы (1):

$$argmax_c P(c|w) = argmax_c P(c) \prod_{i=1\ldots len(w)} P(l_i|c_i l_{i-k} \ldots l_{i-2}\, l_{i-1}) \tag{4}$$

Процесс обучения построен следующим образом. 80% примеров из обучающей выборки используются для оценки вероятностей n-грамм. После этого по оставшимся 20% примерам обучается машина опорных векторов с вышеперечисленными входными значениями. При этом все входные значения нормируются к отрезку [–1,1] и исключаются линейно зависимые значения. Далее статистика n-грамм перестраивается по всем примерам обучающей выборки.

Тем самым, машина опорных векторов в некоторой степени компенсирует ошибки, возникающие при классификации на основе n-грамм, т. к. примеры, на которых она обучается, не использовались для сбора статистики и отражают характер входных данных, к которым будет применяться классификатор.

4. Способы применения

В данной работе не рассматривается использование внешних признаков для семантической классификации имён. Однако мы планируем использовать полученный классификатор в рамках более сложной системы обработки текста, в которой внешние признаки также будут доступны для классификации. Поэтому мы будем осуществлять оценку результатов, исходя из различных способов применения классификатора при наличии внешних признаков.

Используемый метод опорных векторов осуществляет биномиальную классификацию, но может быть преобразован к мультиномиальной (многозначной) путём попарной классификации между всеми имеющимися категориями и выбора категории с самой высокой оценкой.

В более сложной системе анализа текста при помощи внешних признаков можно ограничить множество рассматриваемых категорий для каждого анализируемого имени, тем самым повышая достоверность классификации на основе внутренних признаков.

Таким образом, мы будем оценивать классификатор при решении следующих задач:

- биномиальная классификация имени в рамках произвольной пары категорий — частный случай задачи с уточнением в рамках заранее выбранной пары категорий;
- отделение одной категории от остальных — при решении задач выделения собственных имён определённой категории из текстов;
- мультиномиальная классификация — определение категории имени из полного списка возможных категорий.

Отметим, что возможны также задачи классификации в рамках тройки, четвёрки и т. п. категорий, но изложение экспериментальных результатов по ним было бы весьма объёмным, а никакой новой информации эти результаты не открывают, т. к. все такие задачи сводимы к перечисленным.

5. Методы оценки результатов

Для оценки результатов работы применяется оценка достоверности классификатора (в англоязычной литературе — *accuracy*), определяемая как:

$$accuracy = \frac{\text{число верно классифицированных примеров}}{\text{общее число примеров}} \quad (5).$$

Все полученные результаты приводятся в сравнении с работой (Patel & Smarr, 2001) в тех случаях, когда доступны результаты для сравнения.

Отметим, что в работе (Patel & Smarr, 2001) контрольная выборка содержала различные доли примеров из каждой категории, в связи с чем достоверность константного или произвольного классификаторов в ней может составлять до 85 % для отдельных категорий и режимов работы. В настоящей работе во всех контрольных выборках доля примеров каждой категории составляет $^{1}/_{\text{число категорий}}$, поэтому достоверность константного классификатора составляет:

- 50 % — в задачах биномиальной классификации;
- 80 % — в задачах выделения одной категории (классификатор, определяющий все примеры как не принадлежащие заданной категории);
- 20 % — в задаче мультиномиальной классификации.

6. Источники и характеристики экспериментальных данных

В качестве экспериментальных данных использовались примеры, полученные из различных интернет-источников.

В категории «Имена людей» использовались списки, содержащие имена и фамилии сотрудников кафедры распознавания изображений и обработки текста ФИВТ МФТИ, а также имена, фамилии и отчества депутатов Государственной Думы РФ V и VI созывов. К спискам случайным образом применялись 10 распространённых правил сокращения и перестановки имён, фамилий и отчеств, что позволяет получить практически неограниченное число правдоподобных имён. Общее число примеров данной категории, использованное в экспериментах — 2500.

В категории «Названия компаний» использовались списки рейтинга Эксперт-400 (http://www.expert.ru/) и других рейтингов российских компаний, доступных в Интернете. Отметим, что в полученных списках отсутствуют обозначения организационно-правовой формы компании и другие признаки (кавычки и т. п.), которые существенно упрощают классификацию имён как названий компаний. Объём категории — 2509 названий.

В категории «Фильмы» использовались материалы сайта http://www.kinokadr.ru/allfilms/year/2010/ — названия фильмов, вышедших в прокат в России в период 2000–2010 гг., очищенные от пометок «DVD», «сериал» и т. п. Объём — 2713 названий.

В категории «Лекарства» использованы материалы сайта http://leim.ru/Practica/Drugs/druglist.htm, включающие как торговые марки лекарств, так и названия действующих компонентов. Объём — 1984 примера.

В категории «Названия городов» использованы материалы сайта http://autotravel.ru/towns.php, включающие 1621 название.

Примеры имён указанных категорий приведены в таблице 1.

Таблица 1. Примеры экспериментальных наборов данных

Категория	**Примеры**
Имена людей	*Бочкарев Андрей Валерьевич; Г. Г. Семёнов; Муцоев, Александр Генрихович; Антонов Е. А.*
Названия компаний	*Аптечная сеть 36,6; Волжская ТГК (ТГК-7); ТАИФ-НК; Авиакомпания «Домодедовские авиалинии»*
Фильмы	*Ледниковый период 4: Континентальный дрейф; Эволюция; 11.11.11; Сумерки 4: Рассвет. Часть 1*
Лекарства	*Адапален; АКДС/вакцина полиомиелитная инактивированная; Этидронат натрия; Бензокаин; Берлинская лазурь*
Названия городов	*Бережная Дуброва; Интерпосёлок; Юрьев-Польский; ЫБ; Умет; Урочище Калбак-Таш*

7. Экспериментальные результаты

7.1. Биномиальная классификация

Таблица 2. Характеристики достоверности для режима биномиальной классификации

Категории	Достоверность	Эквивалентные категории (Patel & Smarr, 2001)	Достоверность (Patel & Smarr, 2001)
Люди — лекарства	99,87%		
Люди — фильмы	99,70%		
Люди — компании	99,35%		
Люди — города	99,36%		
Лекарства — города	96,34%	Drug — Place	94,406%
Фильмы — лекарства	96,46%	Drug — Movie	94,875%
Лекарства — компании	94,32%	Drug — NYSE	98,881%
Фильмы — города	87,26%	Movie — Place	88,955%
Фильмы — компании	80,39%	NYSE — Movie	98,887%
Города — компании	82,80%	NYSE — Place	98,721%
Средние значения	**93,585%**		**95,787%**
Средние значения без категории «Люди»	**89,595%**		**95,787%**

7.2. Задача выделения одной категории

Таблица 3. Характеристики достоверности для режима выделения одной категории

Категория	Достоверность	Эквивалентная категория (Patel & Smarr, 2001)	Достоверность (Patel & Smarr, 2001)
Люди	99,36%		
Фильмы	91,15%	Movie — Others	91,987%
Лекарства	96,82%	Drug — Others	95,722%
Города	87,39%	Place — Others	90,325%
Компании	88,98%	NYSE — Others	99,952%
Средние значения	**92,74%**		**94,321%**
Средние значения без категории «Люди»	**91,085%**		**94,321%**

7.3. Задача мультиномиальной классификации

Общая достоверность классификатора — 84–85 %, при этом контрольная выборка содержит одинаковое число примеров каждой из 5 категорий.

По данным (Patel & Smarr, 2001) — 88,971 %, при этом контрольная выборка содержит различное число примеров из 4 категорий.

7.4. Детализация экспериментальных результатов

В экспериментах использовалась статистика символьных n-грамм длины 3, поскольку было проверено, что именно эта длина обеспечивает максимальную достоверность классификатора (см. рис. 2).

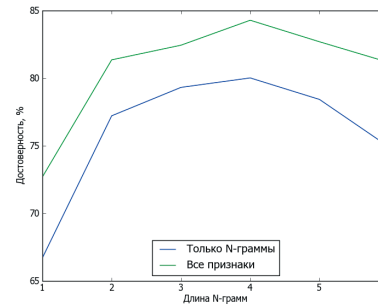

Рис. 2. Зависимость достоверности классификатора от длины используемых n-грамм, мультиномиальная классификация

Были проверены классифицирующие способности каждого признака в отдельности (рис. 3). Поскольку, согласно рис. 2, статистика n-грамм без дополнительных признаков обеспечивает достоверность порядка 80 %, проверены также достоверности классификатора на основе сочетания n-грамм и одного дополнительного признака (рис. 4). На (рис. 4) также приведены для сравнения результаты классификаторов, использующих все признаки кроме N-грамм, и только N-граммы без дополнительных признаков.

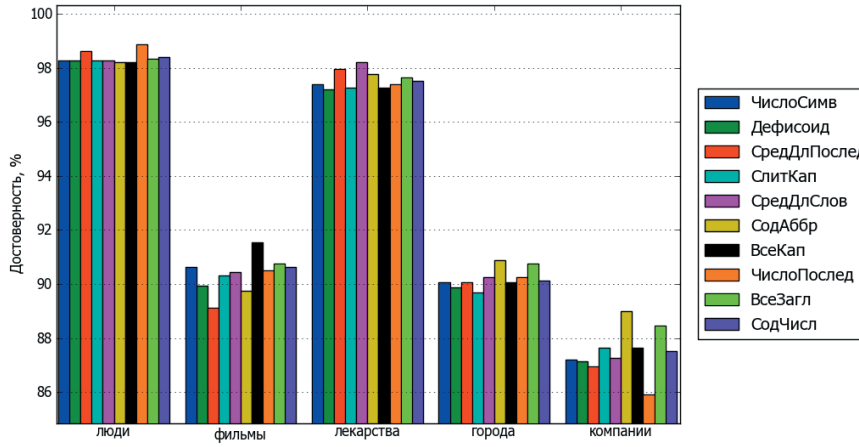

Рис. 3. Классифицирующие способности отдельных признаков, показатель достоверности, задача выделения одной категории

Подчеркнём, что достоверность 80% на рис. 3 соответствует достоверности константного классификатора (не относящего к выбираемой категории ни одного примера), что свидетельствует о полной бесполезности признака для той или иной категории. Достоверности меньшего, чем 80%, уровня говорят о том, что признак позволяет отделить от требуемой категории только некоторую часть примеров, принадлежащих другим категориям.

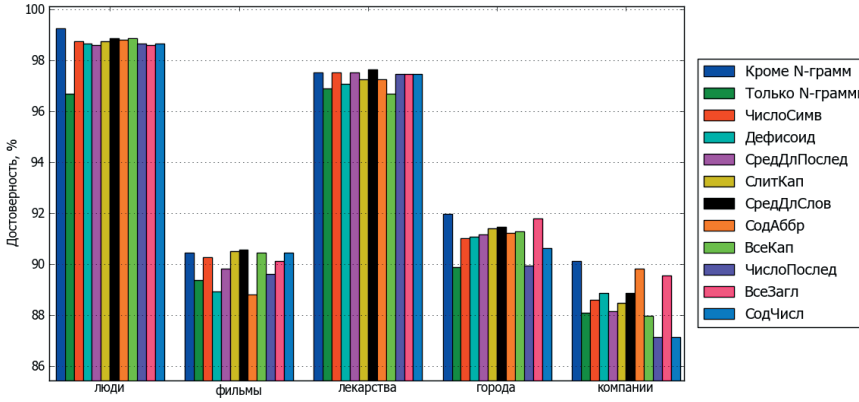

Рис. 4. Классифицирующие комбинации n-грамм и одного дополнительного признака, показатель достоверности, задача выделения одной категории

Для оценки источников ошибок при классификации можно использовать ручной анализ типичных ошибок. Некоторые ошибки приведены в таблице 4.

Таблица 4. Примеры ошибок классификации

Истинная категория	Результат классификатора	Примеры
Компания	Фильм	Бархатные ручки; Вкус лета; Высота; Гармония; Независимость; Пятерочка; Нокиан Шина; Дикси групп
Компания	Город	Дары Покрова; Волжанин; Корнет; Кронверк; Магнит; Молот; Русал; Север; Останкинское; Кукуй
Город	Компания	Владивосток; Москва; Санкт-Петербург; Шишкин Лес; Янтарный; Йошкар-Ола; Сыктывкар
Город	Фильм	Анна; Свобода; Опочка; Воскресенск; Казахский аул; Прислониха
Лекарство	Компания	АДС; Муромонаб-CD3; Антитромбин III
Фильм	Город	Ад; Бог; Вздох; Дзифт; Дитя; Интервью; Окись; Сын; Халк; Кислород
Фильм	Компания	Generation П; Вдох-выдох; Королева; Консервы; Здравозахоронение; Отважная; Рестлер; Макс Пэйн; Корпорация монстров; Рожденные в СССР

7.5. Выводы из экспериментальных результатов

Анализ ошибок в экспериментах показывает следующие основные причины ошибок.

Наиболее достоверно классифицируемыми категориями являются имена людей и названия лекарств. Это объясняется тем фактом, что данные категории обладают наиболее жесткой структурой, характерными префиксами или суффиксами.

Будем обозначать как «**X→Y**» ошибки отнесения имени собственного из категории **X** к категории **Y**.

Ошибки вида «компания→фильм» вызваны использованием в названиях компаний наиболее общей лексики, также употребляемой в названиях фильмов.

Ошибки вида «компания→город», как правило, происходят при классификации коротких названий компаний, состоящих из одного слова и содержащих редкие n-граммы.

Ошибки вида «компания→город» и «город→компания» также вызваны использованием названий городов в названиях компаний, в особенности наиболее значимых городов (*Москва, Санкт-Петербург*).

Ошибки вида «город→фильм» и «фильм→город» вызваны наличием в названиях городов коротких общеупотребительных слов или их фрагментов, также присутствующих в названиях фильмов.

Ошибки вида «фильм→компания» вызваны структурным сходством названия фильма с названиями компаний: наличием дефисов, аббревиатур, капитализации всех слов.

Путём исследования ошибок классификации можно сделать выводы о наиболее статистически значимых характеристиках имён тех или иных категорий, которые приведены в таблице 5.

Таблица 5. Значимые характеристики, полученные в результате ручного анализа ошибок классификации

Категория	**Значимые характеристики**
Имена людей	Структура, префиксы и суффиксы, число слов
Лекарства	Префиксы, суффиксы, n-граммы, число слов
Компании	Специфические свойства структуры: аббревиатуры, капитализация, дефисы, числа; использование географических названий, написание заглавными буквами.
Город	Часто — короткие, мало других признаков.
Фильм	Использование общей лексики, число слов, наличие некапитализованных слов

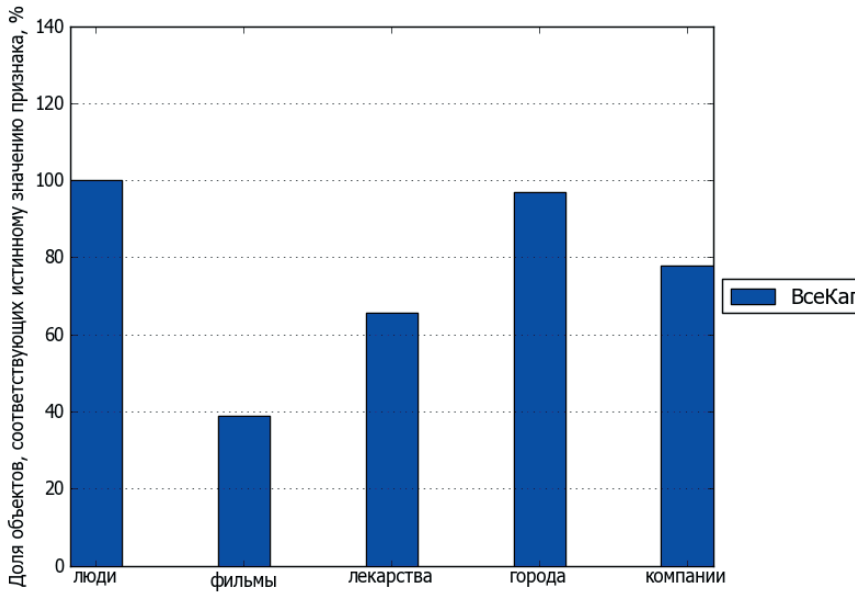

Рис. 5. Доля примеров, в которых все слова капитализованы

Приведём распределения основных признаков по словам из использованного набора. Эти распределения позволяют оценить особенности примеров. По соображениям объёма, они приводятся без дополнительного расширенного описания.

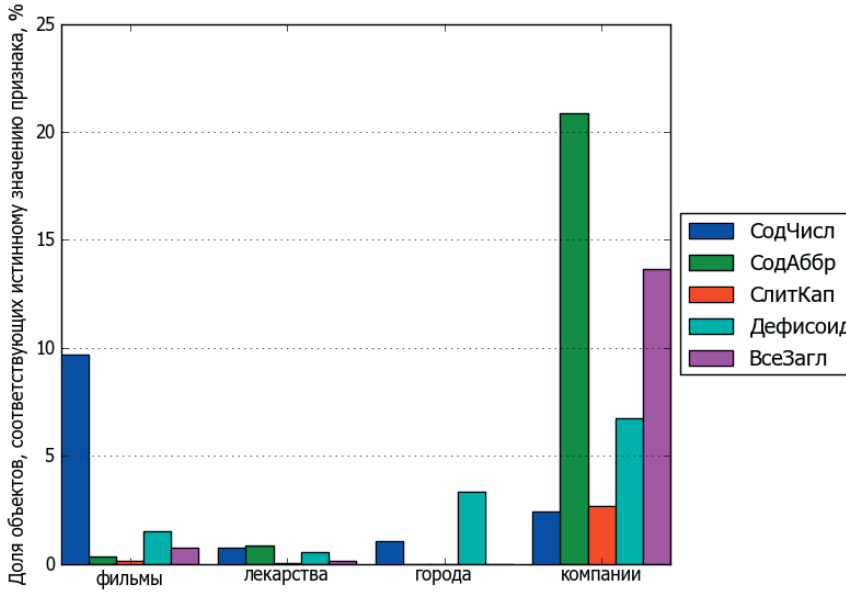

Рис. 6. Распределение истинных значений прочих бинарных признаков по категориям (в категории имён людей все эти признаки имеют значение «ложь»)

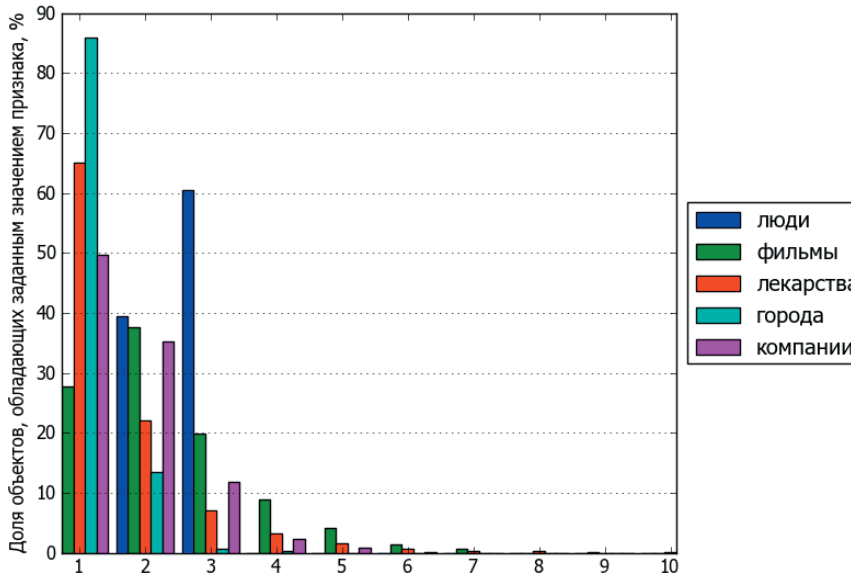

Рис. 7. Распределение числа цифробуквенных последовательностей по категориям

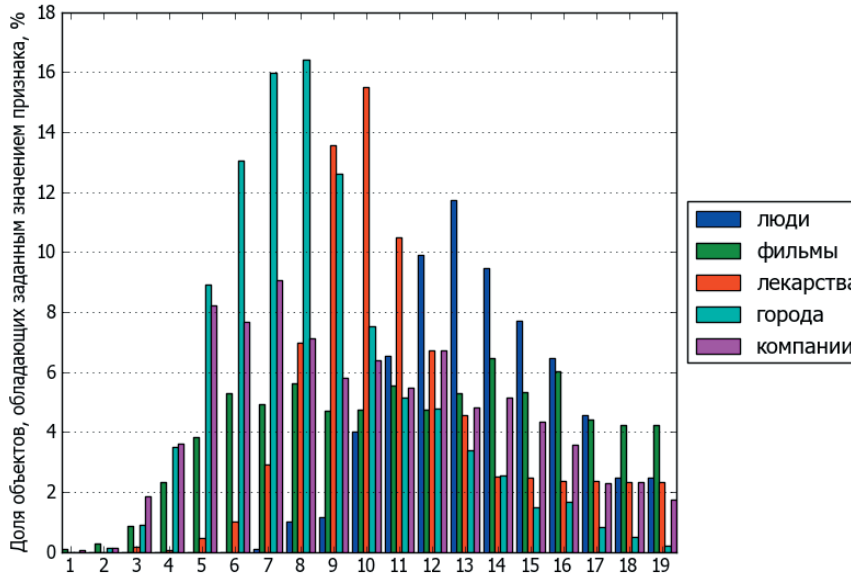

Рис. 8. Распределение общей длины примера в символах по категориям

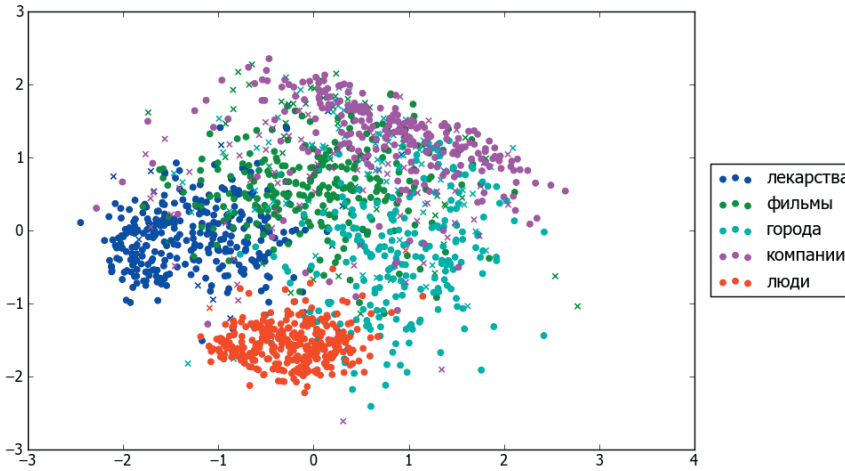

Рис. 9. Применение метода анализа главных компонент (PCA) к набору признаков для визуального отображения взаимного расположения исследуемых категорий. Точки — верно классифицированные примеры, крестики — ошибки, цвет соответствует известной категории имени.

8. Сравнение результатов (Patel & Smarr, 2001)

Таблицы 1, 2, 3 показывают, что наиболее значительные отличия в достоверности между нашим классификатором и (Patel & Smarr, 2001) наблюдаются в категориях «Названия городов» и «Названия компаний», где достоверность нашего классификатора значительно ниже. В остальных категориях достигнуты сходные результаты.

Эти отличия объясняются тем, что в нашем наборе названий компаний практически отсутствуют обозначения организационно-правовых форм (*ООО, ОАО, Группа компаний*) и видов деятельности (*Банк*). В то же время в примерах (Patel & Smarr, 2001) эти атрибуты обнаруживаются столь часто (*Inc., INC, Corporation, Fund*), что в цитируемой работе был введён дополнительный внутренний признак — последнее слово имени, обозначаемый в оригинале как «suffix». В нашем случае подобный признак не вносит ощутимого увеличения достоверности классификатора.

В то же время, указанные категории названий городов и компаний обнаруживают существенное сходство в статистике n-грамм, вследствие чего значительная доля ошибок классификации проявляется в неспособности различать именно эти категории.

Заключение

Нами была продемонстрирована возможность создания достаточно простого классификатора, достигающего достоверности классификации 80–99,9 % между двумя категориями, 89–99 % при отделении одной категории от остальных и общей достоверности 85 % при определении категории имени из 5 заранее заданных категорий.

При этом использовались исключительно внутренние признаки, извлекаемые непосредственно из классифицируемых имен и набора обучающих примеров, имеющего объём порядка 2000 примеров для каждой категории.

Полученные результаты аналогичны приводимым в сходной работе (Patel & Smarr, 2001), авторы которой применяли аналогичный подход, но несколько иной набор признаков, для классификации англоязычных имен.

Наиболее перспективным подходом к дальнейшему совершенствованию классификатора, ограниченного использованием сугубо внутренних признаков, представляется использование частотного общелексического словаря для повышения достоверности классификации названий компаний и фильмов.

В наши дальнейшие планы входят испытания классификатора на большем числе категорий, увеличение набора обучающих примеров, поиск дополнительных внутренних признаков. Помимо этого, значительный интерес представляет расширение классификатора путём интеграции его с системой автоматического разбора текста и анализа синтактико-семантической структуры контекста собственного имени — фактически, источника внешней информации.

Литература

1. *Baluja, S.* (2000). Applying Machine Learning for High Performance Named Entity Extraction.
2. *McDonald, D. D.* (1993). “Internal and External Evidence in the Identification and Semantic Categorization of Proper Names.” available at: Acquisition of Lexical Knowledge: http://aclweb.org/anthology-new/W/W93/W93-0104.pdf
3. *Mikheev, A.* (1997). Automatic Rule Induction for Unknown Word Guessing. Computational Linguistics , 23(3), 405–423.
4. *Patel, S., & Smarr, J.* (2001). “Automatic Classification of Previously Unseen Proper Noun Phrases into Semantic Categories using an N-Gram Letter Model.” available at: http://nlp.stanford.edu/courses/cs224n/2001/jsmarr/NGramWordClassifier.pdf