Выявление личных имен в новостных текстах на материале коллекций Persons-1000/1111-F

© И. В. Трофимов
Институт программных систем им. А.К. Айламазяна РАН,
Переславль-Залесский
itrofimov@gmail.com

Аннотация

Методы извлечения информации из текстов позволяют автоматически структурировать содержащуюся в документах информацию и играют важную роль в ряде приложений, связанных с автоматической обработкой больших документальных массивов и материалов электронных библиотек. В работе рассматриваются возможности простых словарно-эвристических алгоритмов выявления упоминаний лиц в текстах на материале двух новых русскоязычных новостных коллекций.

Работа выполнена при поддержке РФФИ, грант № 13-06-00483а.

1 Введение

Недавно в интернете были опубликованы две коллекции (Persons-1000[1] и Persons-1111-F[2]), созданные для оценки методов автоматического извлечения из текстов упоминаний лиц в форме личного имени. Задача извлечения предполагает нормализацию целевой информации в расчете на помещение ее во внешние по отношению к тексту структуры. Для личных имен в русском языке это означает приведение к именительному падежу. Это довольно сложная задача, предполагающая нормализацию фамилий [1]. В то же время, для ряда прикладных задач достаточно решения задачи выявления (обнаружения) информации, то есть определения мест в тексте, где упоминается целевая информация. В данной работе мы попытались оценить «нижний порог» качества выявления по коллекциям Persons-1000/1111-F, а именно, какой F-меры выявления можно достичь, используя достаточно простые инструменты. Так как коллекции новые и исследования на их основе еще не проводились, полученные в рамках данной работы результаты послужат отправной точкой для дальнейших научных изысканий.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

2 Современный технический уровень

К настоящему моменту проведено немало исследований по распознаванию именованных сущностей (NER), в том числе и упоминаний лиц. В обзорной части мы рассмотрим лишь небольшое число работ, для которых известны количественные результаты по выявлению лиц, полученные на размеченной коллекции.

Уже в ранних работах по выявлению лиц в текстах были получены результаты с F-мерой, превышающей 90. На MUC-7 победившей в треке по распознаванию именованных сущностей системе удалось достичь значения 96 по классу сущностей "person" и 93.39 по всей совокупности именованных сущностей трека [2]. Это была гибридная система, в основе которой лежали контекстные правила распознавания и алгоритмы частичного сопоставления, опирающиеся на предварительно обученный классификатор на базе метода максимальной энтропии. Следует отметить, что NER-трек на MUC-7 состоял всего из 100 новостных сообщений (язык английский). Каждая же из исследуемых нами коллекций (Persons-1000/1111-F) в 10 раз больше.

Позднее в рамках ориентированных на машинное обучение треков CoNLL предлагалось решить задачу распознавания именованных сущностей алгоритмами, не зависящими от языка. В основе подхода, победившего на CoNLL-2003, лежало комбинирование результатов четырех классификаторов [3]; исследовались несколько способов комбинирования. При выявлении лиц на тестовых множествах системе удалось получить следующие результаты: 93.85 для английского языка и 82.8 для немецкого. Снова отметим, что тестовые коллекции CoNLL в разы меньше [4] коллекций Persons-1000/1111-F.

Высокие результаты получены также для арабского языка. На корпусе ANERCorp (150000 слов) была получена F-мера=94.5 на задаче выявления лиц при помощи гибридного подхода, использующего правила и машинное обучение[5].

Для русского языка не так много работ содержат количественные оценки, полученные на каких-либо коллекциях. В работе Подобряева [6] описан подход

к выявлению лиц в текстах на базе CRF-модели, использующей преимущественно графематические признаки и группы лексических признаков. На коллекции из 600 новостных сообщений данный подход позволил получить величину F-меры = 88.32. Близкая задача выявления именных групп, содержащих собственные имена решалась Крейдлиным [7]. Для групп, содержащих личные имена, на коллекции, содержащей 155 таких групп, была получена F-мера = 88.8.

3 Возможности словаря имен

Не имея статистических данных, мы предположили, что в текстах новостного жанра на русском языке для личных имен чаще используется шаблон, который условно можно назвать Имя+Фамилия. Если это так, то при помощи словаря имен, учитывающего словоизменение, можно обнаружить значительную долю личных имен в тексте.

Для проверки этой гипотезы мы использовали морфологический словарь имен (7893 словарных входа). Словарь преимущественно состоял из традиционных славянских, романских, германских и других европейских имен, содержал небольшие подборки имен, распространенных на Кавказе и в мусульманских странах, и был дополнен именами известных современных общественно-политических деятелей из различных регионов мира.

Затем мы написали регулярное выражение (в терминах языка PSL [8]), которое содержательно можно выразить следующим образом. Последовательность будем считать личным именем, если она состоит из следующих компонентов:

1) слово с заглавной буквы, являющееся словарным именем;

2) за которым идет одно или более кириллических слов с заглавной буквы;

3) за которыми в круглых скобках может идти последовательность, в составе которой допускаются только некириллические слова и некоторые знаки препинания (точка, дефис, апостроф).

Первые два компонента позволяют обнаруживать как типичный новостной шаблон (например, *Иван Иванов*), так и ряд более редких (таких как, *Иван Иванович Иванов*, *Сергей Витальевич*, *Хосе Антонио Рейес Кальдерон* и т.п.). Третий компонент предназначен для работы с упоминаниями вида *Джеймс Ставридис (James G. Stavridis)*, которые согласно принятой в коллекциях разметке должны извлекаться полностью.

Применение указанного регулярного выражения позволило получить следующие результаты (таблица 1).

Как видно из результатов, относящихся к Persons-1000, шаблон Имя+Фамилия действительно имеет большое распространение — более 53% личных имен записаны в такой форме. Мы говорим «более», имея в виду неполноту словаря имен и недостатки использованного регулярного выражения (например, во фразе *с сыном генпрокурора РФ Юрия Чайки Артемом Чайкой* оба личных имени не будут обнаружены; вместо двух имен будет выявлена ошибочная последовательность из четырех слов, начиная с *Юрия*).

Таблица 1. Результаты выявления для регулярного выражения, основанного на словаре имен

Коллекция	F-мера	Точность	Полнота
Persons-1000	68,63	96,64	53,21
Persons-1111-F	44,39	87,14	29,78

Наиболее распространенными видами ошибок, допущенных в Persons-1000, стали: соединение двух подряд идущих личных имен, омонимия словарных имен с другими собственными именами (например, во фразе *премьер-министр Израиля Биньямин Нетаньяху* наличие словарного имени *Израиль* приводит к выявлению более длинной цепочки слов, чем необходимо) и наличие в коллекции шаблонов Фамилия+Имя+Отчество (регулярным выражением ошибочно выделяется лишь Имя+Отчество).

Результаты по коллекции Persons-1111-F существенно хуже, главным образом за счет полноты. Это связано с тем, что в использованном словаре слабо представлены имена из региона Юго-Восточной и Средней Азии. Кроме того, значительная доля текстов освещает события в странах арабского мира, где личные имена часто включают в себя префиксы (*аль-*, *эль-*, *бин-*, *абд-* и др.), записываемые со строчной буквы (иногда отдельными словами). Используя словарь таких префиксов и модифицировав соответствующим образом регулярное выражение, можно улучшить результат (таблица 2).

Таблица 2. Эффект арабских префиксов

Коллекция	F-мера	Точность	Полнота
Persons-1000	69,28	97,25	53,80
Persons-1111-F	50,19	91,49	34,58

4 Фамилия с инициалами

Другой распространенный шаблон для личного имени в новостных сообщениях — Инициалы+Фамилия. Чтобы оценить долю таких случаев среди всех личных имен, мы модифицировали предыдущее регулярное выражение, заменив его следующим:

1) инициал (одна заглавная буква или *Дж*, за которыми следует точка);

2) точка;

3) за которой может следовать еще один инициал с точкой, причем допускается дефис между инициалами (например, *Ж.-К. Трише*);

4) затем следует одно кириллическое слово с заглавной буквы;

5) после которого может присутствовать некириллическая запись в круглых скобках (аналогично предыдущему регулярному выражению).

Полученные с помощью этого регулярного выражения результаты представлены в таблице 3.

Таблица 3. Результаты для регулярного выражения, работающего с инициалами

Коллекция	F-мера	Точность	Полнота
Persons-1000	27,51	98,43	15,99
Persons-1111-F	4,26	93,94	2,18

Как видно из результатов, отечественные журналисты избегают использования инициалов при упоминании лиц азиатского региона (за исключением России) и в то же время активно пользуются таким шаблоном в новостях, затрагивающих российский и западный мир.

5 Отдельные имена и фамилии

Использованные регулярные выражения (Имя+Фамилия и Инициалы+Фамилия) не должны иметь существенного пересечения по покрытию текста, поэтому эффект их совместного применения приближенно может быть рассчитан путем суммирования полноты. Тем не менее, мы выполнили эмпирическую оценку (таблица 4).

Таблица 4. Совместный эффект регулярных выражений для шаблонов Имя+Фамилия и Инициалы+Фамилия

Коллекция	F-мера	Точность	Полнота
Persons-1000	81,36	97,52	69,79
Persons-1111-F	52,47	91,63	36,76

Теперь остается открытым вопрос, что представляют собой оставшиеся 30% (в коллекции Persons-1000). Довольно легко проверить, какую долю составляют упоминания лиц в форме отдельно употребленной фамилии или имени. Достаточно разметить все слова с заглавной буквы[3] как отдельные упоминания лиц и взглянуть на полноту (учтем также арабские префиксы); результат в таблице 5.

Таблица 5. Доля личных имен в форме отдельных употреблений имени или фамилии

Коллекция	Полнота
Persons-1000	26,64
Persons-1111-F	35,74

Теперь, если сложить полноту по коллекции Persons-1000, то становится очевидным, что доля каких-либо еще шаблонов (кроме Имя+Фамилия, Инициалы+Фамилия, отдельные имена и фамилии) не превышает 3,6%. Мы говорим «не превышает», поскольку первые два регулярных выражения не гарантируют обнаружения всех шаблонов Имя+Фамилия и Инициалы+Фамилия.

По коллекции Persons-1111-F аналогичные выводы пока сделать нельзя, так как слишком велика неопределенность, обусловленная неполнотой словаря имен.

В общем случае выявление отдельных имен и фамилий нельзя отнести к задачам, эффективно решаемым простыми средствами. Для их обнаружения требуются довольно сложные шаблоны, способные учитывать разнородный (порой достаточно широкий) контекст. На наш взгляд, для решения этой подзадачи целесообразно использовать методы, основанные на машинном обучении.

Тем не менее, в текстах новостного жанра эта задача может решаться достаточно просто. Это связано с тем, что журналисты почти для каждого упоминаемого в тексте лица хотя бы раз используют форму Имя+Фамилия или Инициалы+Фамилия (обычно в момент интродукции лица в тексте). Учитывая эту особенность жанра, мы разработали простой программный модуль, который составлял словарь фамилий, ранее обнаруженных в тексте регулярным выражением для Имя+Фамилия (фамилией считалось последнее слово, обнаруженное регулярным выражением). Для каждого анализируемого текста составлялся свой индивидуальный словарь из обнаруженных в этом тексте фамилий. Затем мы использовали частичное сопоставление по этому словарю, проверяя в тексте каждое слово, начинающееся с заглавной буквы[4]. У алгоритма частичного сопоставления были выделены два параметра:

– усечение — задает, сколько символов с конца слова можно не рассматривать при сопоставлении;

– минимальная длина слова, при котором включается усечение. Если длина проверяемого слова меньше, то проверяется точное совпадение текстовой и словарной формы.

Мы эмпирически подобрали значения этих параметров по коллекции Persons-1000 (таблица 6).

Из таблицы становится понятно, что из 26% полноты (в коллекции Person-1000), приходящихся на отдельно стоящие *имена и фамилии*, более 24% составляют фамилии — максимальная достигнутая полнота (94,15) минус полнота без алгоритма частичного сопоставления (69,79).

Попытки дополнительно обнаружить отдельные фамилии словарными методами не привели к существенным изменениям результатов. Мы пытались использовать морфологический словарь частотных русских фамилий [9], а также экспериментировали с аналогичным словарем, составленным для лиц, часто фигурирующих в СМИ (были использованы списки «Персон года» за последние 3 года публикуемые на сайте «Медиалогия» — 146 различных фамилий). Рост полноты составлял доли процента.

Таблица 6. Величина F-меры в зависимости от параметров алгоритма частичного сопоставления (в скобках указаны точность и полнота соответственно); выделены максимальные значения

		Усечение (в символах)			
		1	2	3	4
Минимальная длина слова для усечения (в символах)	3	95,11 (97,30 / 93,02)	95,55 (97,19 / 93,96)	—	—
	4	95,10 (97,31 / 92,98)	**95,57** (97,26 / 93,92)	95,48 (96,80 / 94,19)	—
	5	95,05 (97,31 / 92,90)	95,52 (97,26 / 93,84)	95,53 (97,00 / 94,10)	95,01 (95,88 / **94,15**)
	6	94,75 (**97,32** / 92,32)	95,21 (97,29 / 93,22)	95,30 (97,19 / 93,48)	95,14 (96,83 / 93,52)

Кроме того, мы пытались применить метод частичного сопоставления для выявления отдельных имен (словарь имен составлялся первым регулярным выражением). Такой метод также не позволил улучшить значение F-меры, а рост полноты составлял доли процента.

6 Нижний порог

Совместный эффект вышеупомянутых регулярных выражений и алгоритма частичного сопоставления для фамилий сведен в таблицу 7.

Таблица 7. Совместный эффект регулярных выражений для шаблонов Имя+Фамилия, Инициалы+Фамилия и алгоритма частичного сопоставления на базе динамически формируемого словаря фамилий

Коллекция	F-мера	Точность	Полнота
Persons-1000	95,57	97,26	93,92
Persons-1111-F	64,43	88,44	50,56

Таким образом, для коллекции Persons-1000 величина F-меры, равная 95, достигается довольно простыми словарными и эвристическими методами, которые тем не менее зависят от полноты словаря имен. Касательно низкого результата по коллекции Persons-1111-F имеется две гипотезы: 1) неполнота словаря имен, 2) весомая доля других шаблонов для личного имени, распространенных в азиатском регионе.

7 Мелкие детали

Продолжив работу с Persons-1000, мы написали еще несколько регулярных выражений, нацеленных на обработку ряда частных аспектов, учитывающих региональные особенности.

Так, например, в европейских личных именах встречаются компоненты, которые исторически указывали на местность (*ван*, *фон*, *де*), а впоследствии трансформировались в часть фамилии. Графематически они записываются со строчной буквы. Мы составили словарь такого рода компонентов и модифицировали регулярное выражения для Имя+Фамилия, чтобы оно могло их учитывать (таблица 8).

Таблица 8. Результат с европейскими частицами, входящими в состав личного имени

Коллекция	F-мера	Точность	Полнота
Persons-1000	95,73	97,39	94,12

Еще одна национальная особенность, которую мы учли в работе, — строение корейских личных имен. Корейское личное имя, как правило, состоит из трех односложных компонентов, первым из которых является фамилия. Корейские фамилии немногочисленны. Мы составили словарь из 179 таких фамилий и написали регулярное выражение для трех компонентов корейского личного имени. Первый должен быть словарной корейской фамилией, а два других — словами с заглавной буквы и длиной не более 5 символов (таблица 9).

Таблица 9. Результат с регулярным выражением для корейских личных имен

Коллекция	F-мера	Точность	Полнота
Persons-1000	95,82	97,43	94,26

Следующим шагом была работа с шаблоном Фамилия+Имя+Отчество для ограниченного списка характерных окончаний фамилий и отчеств. Регулярное выражение требует наличия последовательности из следующих компонент:

1) слово с заглавной буквы, заканчивающееся на *-ов/-ова*, *-ев/-ева*, *-ин/-ина*, *-ский/-ская*, *-цкий/цкая*, *-нко*, *-швили*, *-дзе* (с учетом словоизменения при склонении);

2) слово, являющееся словарным именем;

3) слово с заглавной буквы, заканчивающееся на *-вич/-вна* (также с учетом словоизменения при склонении).

Вероятно, списки характерных окончаний можно расширить. Результат сведен в таблицу 10.

Таблица 10. Результат с ограниченным шаблоном Фамилия+Имя+Отчество

Коллекция	F-мера	Точность	Полнота
Persons-1000	95,95	97,59	94,36

Кроме этого, мы предприняли шаги для повышения точности выявления за счет исключения из словаря «проблемных» имен, омонимичных другим объектам (*Израиль*, *Банка*, *Ливия*, *Рада* и т.д.), и составления регулярного выражения, запрещающего выявлять конструкции вида *имени А.С.Пушкина*, которые не должны выявляться согласно инструкции к Persons-1000 (таблица 11).

Таблица 11. Эффект удаления омонимичных имен и конструкций с *имени*

Коллекция	F-мера	Точность	Полнота
Persons-1000	96,62	98,75	94,58

8 Заключение

В работе исследованы две новые коллекции для оценки качества алгоритмов извлечения и выявления личных имен в текстах — Persons-1000/1111-F. Показано, что задача выявления личных имен в новостных сообщениях может эффективно решаться простыми словарными и эвристическими методами при наличии достаточно полного словаря имен. На коллекции Persons-1000 удалось преодолеть планку значения 95 для F-меры. Отражен вклад каждого отдельного метода выявления в общий результат.

Раскрыта структура коллекции Persons-1000: более 53% приходится на шаблон Имя+Фамилия, около 16% — Инициалы+Фамилия, более 24% — отдельные фамилии.

Литература

[1] Сулейманова Е.А. и Константинов К.А. Об эвристическом методе разрешения неоднозначности при морфологическом анализе незнакомых фамилий // Машинное обучение и анализ данных. — 2013. — Т. 1, № 5. — С. 519—525.

[2] A. Mikheev, C. Grover, M. Moens. Description of the LTG System Used for MUC-7 // Proceedings of the Seventh Message Understanding Conference. Fairfax, Virginia, 29 April – 1 May, 1998.

[3] Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named Entity Recognition through Classifier Combination // Proceedings of CoNLL-2003. Edmonton, Canada, 2003, pp. 168–171.

[4] Tjong Kim Sang, E.F. and De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition // Proceedings of CoNLL-2003. Edmonton, Canada, 2003. — pp. 142–147.

[5] Oudah, Mai and Khaled Shaalan. Person name recognition using the hybrid approach. // Natural Language Processing and Information Systems, volume 7934 of Lecture Notes in Computer Science. Springer, Berlin Heidelberg, 2013, pages 237–248.

[6] Подобряев А.В. Поиск упоминаний лиц в новостных текстах с использованием модели условных случайных полей // Труды XV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14–17 октября 2013 г. — Ярославль: ЯрГУ, 2013. — С. 255—258.

[7] Крейдлин Л.Г. Программа выделения русских индивидуализированных именных групп TAGLITE // Компьютерная лингвистика и интеллектуальные технологии (Диалог'2005): Труды международной конференции. Звенигород, 1–6 мая 2005 г. — М. : Наука, 2005. С. 292—297.

[8] Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Извлечение информации из текста в системе ИСИДА-Т // Труды XI Всероссийской научной конференции RCDL'2009. — Петрозаводск : КарНЦ РАН, 2009. — С. 247—253.

[9] Журавлев А.Ф. К статистике русских фамилий. I. Вопросы ономастики. №2. — Екатеринбург : Изд-во Уральского ун-та, 2005. — С. 126—146.

Person Name Recognition in News Articles Based on the Persons-1000/1111-F Collections

Igor V. Trofimov

Information extraction methods allow for structuring information contained in documents. They play an important role in a number of applications involved in the processing of large document collections and digital library content. The paper evaluates simple heuristic dictionary-based algorithms for person mentioning information extraction in two Russian-language collections of news articles.

[1] http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

[2] http://ai-center.botik.ru/Airec/index.php/ru/collections/29-persons-1111-f

[3] Будем полагать, что употребления фамилий вида *де Гроот*, *фон Беттихер* и т.д. довольно редки.

[4] Мы не стали брать фамилии из шаблона Инициалы+Фамилия из-за неполноты словаря имен. Например, если у нас в тексте встречаются личные имена *Наото Кан* и *Н.Кан*, а в словаре имен нет *Наото*, то, обнаружив фамилию *Кан* в конструкции с инициалами, мы выявим ее и после *Наото*, совершив ошибку, т.к. имя останется невыявленным.