Региональный классификатор текстов для поиска упоминаний лиц в новостных текстах

© А.В. Подобряев
Институт программных систем имени А.К. Айламазяна РАН,
Переславль-Залесский
alex@alex.botik.ru

Аннотация

Рассматривается задача выявления в новостных текстах на русском языке упоминаний людей в форме личных имен в условиях отсутствия словаря имен и фамилий. Приводятся два подхода к использованию национально-регионального классификатора текстов.

1 Введение

Существует ряд естественно возникающих задач автоматической обработки естественных текстов, требующих выявления именованных сущностей. Например, в задачах извлечения информации о событиях необходимо находить в тексте упоминания географических объектов, названий организаций, имен лиц. В настоящей работе мы ограничиваемся рассмотрением задачи поиска имен лиц в новостных текстах на русском языке.

Обычно для решения этой задачи используют те или иные словари имен и фамилий. Однако, такой подход не работает при анализе новостных текстов, в которых упоминаются экзотические имена и фамилии. Например, китайские, корейские, японские, арабские и другие.

Кроме того, обычно используемые признаки в системах, основанных на статистическом обучении, а также правила, используемые в системах, основанных на правилах, могут не работать. Например, арабское имя может быть сколь угодно длинным, состоять не только из слов, начинающихся с прописной буквы, части имени могут иметь префиксы и т.п.

Например,

«*3 февраля суд Кувейта приговорил местного жителя Мухаммеда Эйд аль-Аджми (Mohammad Eid al-Ajmi) к пяти годам тюрьмы за оскорбление эмира Сабаха аль-Ахмада аль-Сабаха в Twitter*».

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

Поэтому возникает естественное решение – ввести нужные признаки в существующую систему признаков. Как показывают эксперименты, такой подход приводит к некоторому увеличению полноты выявления имен лиц, но при этом страдает точность выявления на всей коллекции.

Мы предлагаем создать предварительный обучаемый классификатор текстов на несколько классов, соответствующих национально-географическому принципу. Этот классификатор работает перед основным распознавателем лиц в тексте.

Данная работа выполнена в рамках существующей и развивающейся системы автоматического извлечения информации из текстов ИСИДА-Т [1] и использует имеющийся там ресурс знаний.

2 Предварительный классификатор

Предварительный классификатор состоит из следующих классов: Европа, Китай, Индия, Япония, Средняя Азия, Арабский мир.

В качестве признаков используются основы слов, соответствующие концептам геополитических единиц из используемого нами ресурса знаний (существующего в рамках системы ИСИДА-Т [1]), их частоты, теги новостного текста (при их наличии).

В качестве предварительно классификатора мы используем SVM [2] с квадратичным ядром. Применялась собственная реализация SVM на основе алгоритма SMO [3].

Выходные данные SVM, имеющие геометрический смысл расстояния до разделяющей поверхности, пересчитываются в величины, имеющие вероятностную интерпретацию, согласно алгоритму Платта [4].

Для получения многоклассового классификатора использовался «турнир» двуклассовых классификаторов, отделяющих каждый класс от каждого.

Таким образом, в качестве выхода предварительного классификатора мы получаем вектор, размерность которого равна числу классов, в каждой компоненте которого содержится

вероятность принадлежности данного текста данному классу.

3 Варианты использования предварительного классификатора

Здесь приведены два возможных варианта использования предварительного классификатора.

1. Использование вектора вероятностей принадлежности данного текста имеющимся классам в качестве дополнительных признаков для основного алгоритма обнаружения лиц в тексте.

При этом основной алгоритм использует, наряду с универсальными признаками, также признаки, отвечающие за специфику каждого конкретного класса-региона.

2. Второй вариант заключается в том, что для различных классов-регионов используются различные системы признаков и обучаются отдельные распознаватели лиц. Результат предварительного классификатора используется для выбора нужного распознавателя из имеющегося набора.

4 Основной алгоритм выявления лиц

В качестве алгоритма выявления лиц мы используем CRF [5].

Система универсальных признаков (не зависящих от региона, о котором идет речь в данном новостном тексте) описана в работе [6]. Для построения признаков используется существующий в системе ИСИДА-Т ресурс знаний.

По сравнению с описанными ранее признаками мы добавили признаки, отвечающие за то, что данное слово соединено отношениями типа UP и AKO с концептами «геополитическая единица» и «национальность». Учитывается наличие слов написанных в скобках латинскими буквами, так как часто в текстах иностранное имя дублируется в скобках латиницей.

Кроме того, имеется ряд признаков специфичных для каждого конкретного региона.

Использовался следующий набор словарей.

1. Для арабского мира словарь префиксов (например, «*аль-*», «*эль-*», «*ибн*» и т.п.).
2. Для Японии словарь характерных окончаний имен и фамилий (например, «*-хито*», «*-яма*» и т.п.).
3. Для Средней Азии словарь характерных окончаний русифицированных фамилий (например, «*-ов*», «*-инов*», «*-беков*», «*-ев*», «*-аев*» и т.п.).

Для китайских имен учитывается типичная длина имени и количество слогов в отдельном слове.

Для арабских имен было использовано увеличенное окно CRF-метода длины восемь, в то время как для остальных регионов длина окна была равна пяти.

5 Размеченная коллекция

В Исследовательском центре искусственного интеллекта ИПС им. А.К. Айламазяна РАН созданы две размеченные коллекции новостных текстов на русском языке Persons-1000 [7] и Persons-1111-F [8].

Первая коллекция содержит 1000 документов, 156 тысяч словоупотреблений и 6455 вхождений европейских личных имен.

Во второй коллекции собраны новости из юго-восточной Азии, Китая, Японии, Средней Азии и т.д. Эта коллекция содержит 1111 документ, 160 тысяч словоупотреблений и 6993 вхождения восточных имен. Ниже в таблице приведено количество текстов в этой коллекции, соответствующих определенному региону.

Регион	Количество текстов
Китай	113
Япония	190
Индия	257
Средняя Азия	220
Арабский мир	331

5 Результаты

При использовании системы, обученной только на коллекции Persons-1000, содержащей в основном европейские имена и фамилии, получаются следующие результаты при тестировании на части коллекции Persons-1000, не участвовавшей в обучении, и коллекции Persons-1111-F.

	Европейские имена, %	Восточные имена, %
Точность	88.1	65.61
Полнота	83.98	56.39
F-мера	85.99	60.68

Результат выявления восточных имен не удовлетворительный, поэтому необходимо включать содержащие их тексты в обучающее множество, а также учитывать возможную структуру таких имен для построения дополнительных признаков.

Для обучения SVM брались 2/3 коллекции Persons-1111-F с равномерным распределением по классам.

Эти же 2/3 добавлялись к 2/3 коллекции Persons-1000 для обучения CRF.

Тестирование производилось на оставшихся третях каждой коллекции.

При использовании первого из вышеописанных подходов, т.е. вероятностный выход SVM (вектор размерности равной числу классов) добавляется в систему признаков для последующего обучения CRF, получаются следующие результаты на тестовом множестве.

	Европейские имена, %	Восточные имена, %
Точность	87.66	82.03
Полнота	79.28	72.09
F-мера	83.26	76.74

Здесь видно, что на основной коллекции европейских имен произошло падение результатов. Это говорит о том, что стоит попробовать использовать разные системы признаков и по-разному обученные системы для каждого региона. О каком из регионов идет речь в данном тексте при этом нужно определять с помощью отдельного распознавателя.

При проведении эксперимента, соответствующего этому подходу, были получены следующие результаты.

	Европейские имена, %	Восточные имена, %
Точность	87.93	85.83
Полнота	80.71	76.52
F-мера	84.16	80.91

Видно, что использование отдельных распознавателей для каждого региона превосходит по результатам использование одного распознавателя имен лиц с единой системой признаков.

При этом более трудоемким является обучение системы, а ее применение имеет одинаковую вычислительную сложность для обоих подходов.

Настоящая работа выполнена при поддержке РФФИ, проект № 13-07-00307А.

Литература

[1] 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.

[2] V. Vapnik. Statistical learning theory. John Wiley and Sons, Inc., New York, 1998.

[3] J.C. Platt. Sequential Minimal Optimization: A fast approach algorithm for training support vector machines. Microsoft Research Technical Report MSR-TR-98-14, April, 1998.

[4] J.C. Platt. Probabilistic output for support vector machines and comparisons to regularized likelihood methods. In Advances of large margin classifiers, A. J. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans eds., MIT Press, 1999.

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

[6] А.В. Подобряев. 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, 2013. С. 255—258.

[7] Коллекция “Persons-1000”, Исследовательский центр искусственного интеллекта ИПС им. А.К. Айламазяна РАН, 2013. URL: http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

[8] Коллекция “Persons-1111-F”, Исследовательский центр искусственного интеллекта ИПС им. А.К. Айламазяна РАН, 2013. URL: http://ai-center.botik.ru/Airec/index.php/ru/collections/29-persons-1111-f

Regional Classification of Russian News Texts for Person Recognition

Alexey V. Podobryaev

We study the problem of persons recognition in Russian news texts. If the text is about non-European countries, we can’t use the dictionaries of names and surnames. We investigate two approaches of using the results of preliminary classification of texts to several classes depending on nation or geographical region. One of them consists in using the results of such classifier as additional features for names detection, and the second one consists in making different feature systems for different classes.