Использование методов извлечения информации при географической привязке текстов на русском языке

Прокофьев Петр Александрович

Компания «ЛАН-ПРОЕКТ»
p_prok@mail.ru

Аннотация

В работе предлагается метод выделения фрагментов текста, позволяющих осуществлять географическую привязку текстов на русском языке. Настраиваются и оцениваются обучаемые методы MaxEnt, MEMM, CRF выделения объектов для разрешения неоднозначностей. Оценивается комбинированный метод выделения объектов. Методы сравниваются по показателям полноты и точности.

Введение

Текст на естественном языке может содержать информацию о географических объектах, с которыми связан текст. Процедуру установления такой связи будем называть географической привязкой текста.

В более узком смысле под географической привязкой текста будем понимать установку связей между фрагментами текста и объектами географического справочника, содержащего информацию о названиях, типах и связях географических объектов. Таким справочником может быть электронная карта с именованными объектами, классификаторы по странам, регионам или адресам.

За рубежом задаче географической привязке текстов уделяется большое внимание. Периодически проводятся семинары: Geographic Information Retrieval (в рамках конференции Special Interest Group of Information Retrieval), GeoCLEF (в рамках конференции Cross-Language Evaluation Forum), Analysis of Geographic References (в рамках конференции North American Chapter of the Association for Computational Linguistics - Human Language Technologies). На этих семинарах представлен широкий спектр статей, описывающих методы и системы выделения географических объектов и разрешения неоднозначностей. Однако оценка этих методов и систем осуществлялась авторами преимущественно для англоязычных текстов.

В работе [10] описан обучаемый метод выделения именованных объектов, лежащий в основе метода разрешения неоднозначностей географической привязки, описанного в работе [4]. Разработчики системы MetaCarta в статье [8] описывают метод разрешения неоднозначностей на основе весов, назначаемых при анализе контекста.

Активно развиваются системы географического поиска и анализа географически привязанных текстов: MetaCarta [5], Google Maps [2]. Среди российских разработок можно выделить Yandex Карты [12]. Географическая привязка в этих системах используется на стадии индексации текстов для поиска по пространственным запросам.

Перед началом исследования на качественном уровне было оценено несколько систем выделения объектов в русскоязычных текстах [9], [11] и [14]. Оценка показала, что эти системы с географическими объектами работают на основе регулярных выражений и словарей названий. Эти методы часто не позволяют разрешать неоднозначности в текстах на естественном языке.

Ниже приведены неоднозначности, влияющие на географическую привязку текстов на русском языке, выявленные и классифицированные при анализе текстов новостных и энциклопедических статей.

1. Омонимия географических названий и имен нарицательных или их форм: «поселок Строитель», «город Чехов».
2. Географические названия включают имена или фамилии известных людей: город Энгельс Саратовской области.
3. Географические названия входят в состав названий организаций. Например: Администрация Воронежской области.
4. Синонимия различных названий:
«Россия» = «Российская федерация»,
«Париж» = «столица Франции»,
«Япония» = «родина японцев».
5. Исторические изменения географических названий.
6. Омонимия географических названий: Алтайский край и Алтайский район, город Железногорск Красноярского края и Железногорск Курской области.

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

7. Контекстная зависимость относительной географической привязки: «50 километров от Москвы по Дмитровскому шоссе».

В работе описывается метод выделения и разрешения первых трех видов неоднозначностей. Для разрешения неоднозначностей используются обучаемые методы извлечения объектов и признаков в текстах.

Цель этого исследования – разработать метод выделения фрагментов русскоязычных текстов позволяющих осуществлять географическую привязки текстов.

Задачи исследования:

- разработать метод выделения фрагментов, не позволяющих однозначно осуществить географическую привязку;
- выбрать характеристики слов, используемые в обучаемых методах;
- сравнить качество разрешения неоднозначностей при использовании MaxEnt [10] [7], MEMM [6], CRF [3] и комбинированного методов с различными наборами характеристик;
- сделать вывод, какой метод лучше использовать для разрешения неоднозначностей при определении типа географического объекта, название которого употребляется в тексте.

Метод выделения неоднозначностей

Географическая привязка текстов осуществляется путем поиска в тексте географических названий из справочника. Этапы выделения неоднозначностей привязки описаны ниже.

1. Перед поиском текст разбивается на лексемы, каждая из которых описывается морфологической и графематической информацией. Разбивка осуществляется с использованием морфологического модуля RML [13].

2. Каждой лексеме присваивается набор словарей, в которых найдена эта лексема. Используются словари имен, фамилий, отчеств, составленные по телефонному справочнику МГТС, родовые словари для географических объектов, «общий» словарь, составленный по набору художественных и научных произведений, а также словари географических названий, сгруппированные по типам объектов. Географические словари построены с использованием общероссийского классификатора стран мира (ОКСМ) и общероссийского классификатора объектов административно-территориального деления (ОКАТО). Используется 8 типов географических объектов:

- страны,
- федеральные округа,
- края,
- республики,
- автономные округа
- области,
- районы,
- города.

3. Для каждой лексемы проверяется выполнение предикатов, зависящих от лексемы и ее контекста. Далее такие предикаты будем называть запросами. Запросы формулируются на специальном языке, и их значения зависят от ранее вычисленных характеристик лексемы и лексем, входящих в ее контекст с учетом порядка. Запросы сгруппированы по типам географических объектов. Также выделены группы запросов для фамилий, имен и отчеств. Каждый запрос составлен для максимизации точности так, что если для лексемы выполнен только один запрос, то лексему можно однозначно классифицировать по типу географических объектов. В результате вычисления запросов, каждой лексеме присваивается набор групп, к которым относятся выполняющиеся для лексемы запросы. Примеры запросов приведены в таблице 1.

Город	`город @%ГОРОД`
	`г. @%ГОРОД`
	`г @%ГОРОД`
	`@%ГОРОД #2 столица %СТРАНА`
	`столица %СТРАНА #2 @%ГОРОД`
	`житель @%ГОРОД`

Страна	`житель @%СТРАНА`
	`гражданин @%СТРАНА`
	`гражданка @%СТРАНА`
	`экономика @%СТРАНА`
	`политика @%СТРАНА`
	`бюджет @%СТРАНА`
	`политика против @%СТРАНА`
	`страна @%СТРАНА`
	`граница @%СТРАНА`

Фамилия	`@%ФАМИЛИЯ %ИМЯ`
	`%ИМЯ @%ФАМИЛИЯ`
	`%ИМЯ %ОТЧЕСТВО @%ФАМИЛИЯ`
	`*ИНИЦИАЛ . @%ФАМИЛИЯ`
	`@%ФАМИЛИЯ *ИНИЦИАЛ .`
	`господин @%ФАМИЛИЯ`
	`госпожа @%ФАМИЛИЯ`
	`%ЗВАНИЕ @%ФАМИЛИЯ`
	`%ПРОФЕССИЯ @%ФАМИЛИЯ`

Таблица 1. Примеры запросов, взятые из разных групп. Синтаксис языка: @ – текущая лексема, % – словарь, * – регулярное выражение.

4. Лексемы, для которых выполняются запросы из разных групп, или находящиеся в нескольких словарях, считаются неоднозначными и для их классификации используются обучаемые методы извлечения информации и метод максимального веса.

Таким образом, в тексте выделяются лексемы, которые согласно предварительной классификации по словарям и запросам относятся к нескольким классам. Эти лексемы будем называть «неоднозначностями». Их классификация осуществляется несколькими методами, описание и сравнение которых приведено ниже.

Метод максимального веса

Запросам и словарям присваиваются веса. При классификации среди выполняемых запросов и содержащих слово словарей выбирается тот, у которого максимальный вес. Выбранный словарь или запрос определяют класс лексемы, исходя из группы, к которой принадлежит словарь или запрос.

Обучаемые методы для разрешения неоднозначностей

Обучаемые методы, описанные ниже, решают одну и ту же задачу классификации лексем текста, поэтому можно ввести общую терминологию и обозначения.

Каждую лексему необходимо отнести к некоторому классу $c \in C, s = |C| < \infty$. Множество классов включает в себя набор типов объектов и служебные классы для обозначения «других» лексем (без типа), границ фрагментов и контекстных лексем (суффиксных и префиксных).

Перед классификацией для каждой лексемы b вычисляется двоичный вектор длины k, координаты которого вычисляются, как значения характеристических функций $f_i(b), i \in \overline{1,k}$ и показывают, какие характеристики лексемы активны. Выбор характеристик влияет на результаты классификации, что будет показано ниже.

Обучение осуществляется по выборке $T = \left(\left(b^{(1)}, c^{(1)}\right), \left(b^{(2)}, c^{(2)}\right), \ldots, \left(b^{(h)}, c^{(h)}\right)\right) \in \left(B \times C\right)^h$, где h - длина выборки, B - множество лексем. Из обучающего набора вычисляется эмпирическое распределение $\tilde{p}(b,c) = \frac{n(b,c)}{h}$, где $n(b,c)$ – мощность $\left\{m \mid (b,c) = \left(b^{(m)}, c^{(m)}\right), m \in \overline{1,h}\right\}$.

При классификации выбирается класс, для которого условная вероятность $p(c \mid b)$ – наибольшая. Задача методов найти (оценить) распределение $p(c \mid b)$ по обучающей выборке.

Maximum Entropy

Метод Maximum Entropy (MaxEnt) основан на принципе максимальной энтропии, и широко используется для извлечения информации и выделения именованных сущностей [10]. Метод строит вероятностную модель с явным заданием условной вероятности:

$$p(c \mid b) = \frac{1}{Z(b)} \prod_{i \in \overline{1,k}, j \in \overline{1,s}} u_{i,j}^{f_{i,j}(b,c)}, \qquad \textbf{(1)}$$

где $Z(b)$ – нормирующий множитель,

Множители $u_{i,j}$ вычисляются при решении оптимизационной задачи максимизации энтропии условного распределения модели:

$H(p) = -\sum_{c,b} \tilde{p}(b)\, p(c \mid b) \log\left(p(c \mid b)\right)$, то есть

$$p^* = \arg\max_{p \in P}\left(H(p)\right), \text{ где}$$

$P = \left\{p \mid M_p f_{i,j} = M_{\tilde{p}} f_{i,j}, i \in \overline{1,k}, j \in \overline{1,s}\right\}$ - ограничения на распределения, заданные через математические ожидания характеристик вычисленные по обучающей выборке.

Поиск решения можно осуществлять с помощью алгоритма Generalized Iterative Scaling (GIS-алгоритм) [1].

Maximum Entropy Markov Model

Метод Maximum Entropy Markov Model (MEMM) [6] задает в явном виде распределение $p(c \mid b, c_{-1})$, где c_{-1} – класс предыдущей лексемы. Распределение задается через совокупность распределений $\left\{p_{c_{-1}}(c \mid b), c_{-1} \in C\right\}$, каждое из которых описывается формулой (1) из метода MaxEnt. То есть модель можно также обучать с использованием GIS-алгоритма, но для каждого класса обучение проводится отдельно.

При классификации лексем последовательно определяют класс лексемы, в зависимости от результатов классификации предыдущей лексемы. В целом метод аналогичен MaxEnt, однако дает более точные результаты в задачах извлечения информации в текстах.

Conditional Random Fields

Метод Conditional Random Fields (CRF) [3] аналогично MaxEnt определяет в явном виде $p(c \mid b)$, но кроме характеристик лексем используются также характеристики переходов классов.

$$p(c \mid b, c_{-1}) = \frac{1}{Z(b)} \prod_{i,j} u_{i,j}^{f_{i,j}(b,c)} \prod_{t=1}^{l} \eta_t^{g_t(b,c,c_{-1})}, \text{ где}$$

$g_t(b,c,c_{-1})$ – характеристика перехода, зависящая от текущей лексемы, а также классов предыдущей и текущей лексемы.

При тестировании метода использовались следующие характеристики переходом:

$$g_{i,j,r}(b,c,c') = \begin{cases} 1, \textit{если } f_i(b) = 1, \widetilde{n} = \widetilde{n}_j, c' = c_r \\ 0, \textit{иначе} \end{cases}$$

Аналогично MaxEnt для вычисления множителей использовался GIS-алгоритм.

Комбинированный метод

В процессе обучения и классификации тестовой выборки осуществляется оценка результатов классификации. Оценка вычисляется с использованием показателей точности, полноты, F-меры, для каждого типа:

$$P_c = \frac{A}{A+B}; R_c = \frac{A}{A+C}; F_c = \frac{2P_cR_c}{P_c+R_c},$$

где A – число лексем отнесенных к типу c, как в оцениваемом результате классификации, так и в тестовой выборке, B – число лексем отнесенных к типу в оцениваемом результате классификации, но не отнесенных в тестовой выборке, C – число

лексем отнесенных к типу в тестовой выборке, но не отнесенных в оцениваемом результате классификации. При подсчете показателей используются все лексемы (не только неоднозначные), отнесенные к какому-либо типу географических объектов либо в тестовой выборке, либо в оцениваемой выборке.

Комбинированный метод использует результаты классификации несколькими методами (MaxEnt, MEMM, CRF, максимального веса), а также оценки этих результатов. Если лексема отнесена к нескольким классам $\left\{c_{i_1}, c_{i_2}, ..., c_{i_d}\right\} =: C' \subset C$ по результатам разных классификаций, то ей присваивается класс $c_0 = \arg\max_{c \in C'} F_c$, выбранный методом, у которого показатель F-меры результатов классификации максимален для данного класса.

Сравнение методов

Обучение и классификация осуществлялись на размеченном в полуавтоматическом режиме корпусе новостных сообщений по регионам России и странам мира. Каждый метод проверялся за 5 итераций, каждая из которых состояла из 2-х этапов: обучение на одной части корпуса и классификация оставшейся части корпуса. Для каждой итерации использовались различные деления корпуса на обучающий и тестовый наборы.

На каждой итерации производится вычисление интегральных показателей точности и полноты:

$$P^{(i)} = \frac{1}{|C|}\sum_{c \in C} P_c^{(i)}; R^{(i)} = \frac{1}{|C|}\sum_{c \in C} R_c^{(i)},$$

где i – номер итерации. Для каждого метода выводились средние показатели, объединяющие результаты всех итерации тестирования метода.

Обучающие выборки

Размеченный корпус состоит из 846 текстов, содержащих 372367 лексем (в то числе знаков пунктуации), из которых 5431 отмечены одним из 8 типов географических объектов.

При обучении на каждой итерации используется часть корпуса из 400 текстов, остальные 446 используются как тестовый набор.

Характеристики слов и переходов

Тестирование осуществлялось с разными наборами характеристик. В качестве характеристик использовались функции, приведенные в таблице 2.

1) словарные	- принадлежность к словарю текущего, следующего или предыдущего слов; - выполнение запросов в контекстах текущего, предыдущего или следующего слов; - равенство предыдущего или следующего слова одному из часто употребляемых слов в контекстах помеченных в обучающей выборке слов.
2) морфологические и графематические (вычисляются для текущего, предыдущего и следующего слова)	- графематические дескрипторы, например, «первая заглавная», «знак пунктуации»; - морфологические дескрипторы, такие как падеж, число, род, часть речи.

Таблица 2. Характеристики, используемые при тестировании обучаемых методов.

Первый набор характеристик содержит только словарные характеристики, второй – словарные, морфологические и графематические.

Результаты сравнения

Отправной точкой для сравнения методов будем считать показатели качества метода максимального веса. Метод дал результаты: $P = 74,8\%; R = 83,4\%$. Показатели обучаемых методов приведены в таблице 3.

	Словарные (233)		**Сл. + морф. + граф. (285)**	
	P	***R***	***P***	***R***
MaxEnt	87,9%	87,7%	86,6%	84,5%
MEMM	86,8%	88,2%	86,8%	87,5%
CRF	87%	89,6%	86,9%	89,4%
Комбин.	*86,7%*	*88,2%*	*86,9%*	*87,5%*

Таблица 3. Показатели качества обучаемых методов при определении типов географических объектов. Рассмотрены 2 группы характеристик: только словарные и словарные с морфологическими и графематическими дескрипторами.

Результаты сравнения показали, что комбинированный метод не оправдывает себя, поскольку дает показатели хуже, чем метод CRF.

Все методы дают показатели выше, чем метод максимального веса. Метод CRF показывает наилучшие результаты.

Кроме этого, использование морфологических и графематических дескрипторов ухудшает качество классификации. Возможно, это происходит по причине недостаточного объема размеченного корпуса.

Сравним методы, при использовании только словарных характеристик. При фильтрации характеристик по частоте встречаемости их в корпусе были получены наборы с различным числом характеристик. Результаты сравнения методов в зависимости от числа характеристик изображены на графике (см. рис. 1).

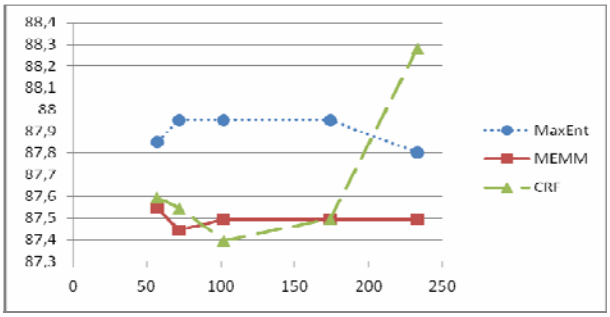

Рисунок 1. Сравнение показателя F-меры качества классификации обучаемыми методами при использовании различного объема словарных характеристик.

Стоит отметить, что метод MEMM не значительно зависит от числа характеристик в рассмотренном диапазоне. Уменьшение характеристик сильно сказывается на полноте в методе CRF. Метод MaxEnt имеет низкую полноту (ниже, чем у CRF и MEMM) и высокие показатели F-меры устанавливаются за счет показателя точности.

Заключение

В работе представлен метод выявления неоднозначностей при поиске в текстах географических объектов и определения их типов. Для разрешения этих неоднозначностей использовались различные обучаемые методы, среди которых лучшим оказался метод CRF. Для метода CRF есть возможность подобрать характеристики переходов, что планируется сделать в будущих работах. Также в дальнейшем планируется провести анализ, какие характеристики лексем и переходов влияют на качество выполнения географической привязки.

При разрешении других указанных во вступлении неоднозначностей необходимы методы, учитывающие глобальный контекст и зависимости между фрагментами текста, такие методы предложены в работах [4] [8]. Описанные выше методы в том виде, в котором они используются в этой работе, не могут разрешить эти неоднозначности. В дальнейшем планируется использовать методы учета глобального контекста.

Работа показывает проблемы и возможные пути их разрешения при выполнении географической привязки текста. Использование географически привязанных текстов дает возможность совместно производить пространственный и тематический анализ и поиск данных, на что будут направлены дальнейшие исследования.

Литература

[1] Darroch J. N. and Ratcliff D. Generalized Iterative Scaling for Log-Linear Models. The Annals of Mathematical Statistics, 43(5):1470-1480, 1972.

[2] Сервис «Google Maps». http://maps.google.ru.

[3] J. Lafferty, A. McCallum, F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML, 2001.

[4] Huifeng Li, Rohini K. Srihari, Cheng Niu, and Wei Li. InfoXtract location normalization: a hybrid approach to geographic references in information extraction, NAACL 2003 Workshop on the Analysis of Geographic References

[5] Solutions for government. MetaCarta, MC2006CB4-01, 2006. (www.metacarta.com).

[6] McCallum A., Freitag D., Pereira F. Maximum entropy Markov models for information extraction and segmentation. *Proc. ICML 2000* (pp. 591–598). Stanford, California.

[7] A. Ratnaparkhi. Maximum Entropy Models for Natural Language Ambiguity resolution. PHD thesis, Univ. of Pennsylvania, 1998.

[8] Erik Rauch, Michael Bukatin, Kenneth Baker. A confidence-based framework for disambiguating geographic terms. — 2003 — HLT-NAACL 2003 Workshop: Analysis of Geographic References.

[9] RCO Fact Extractor: персональная аналитическая система для поиска фактов в тексте. Компания RCO. www.rco.ru.

[10] Srihari, Rohini, Cheng Niu, and Wei Li. A Hybrid Approach for Named Entity and Sub-Type Tagging. In Proceedings of ANLP 2000, Seattle.

[11] Процессор SynSys Semantix. Компания «Синергетические системы». www.synsys.ru.

[12] Сервис «Яндекс – Карты». http://maps.yandex.ru.

[13] Автоматическая Обработка Текста. www.aot.ru.

[14] Описание технологии ИАС Арион. Компания «САЙТЕК». www.sytech.ru.

Using the methods of information extraction in the geographic referencing of russian texts

Prokofjev Petr Alexandrovich

In this paper, a method of allocating text's fragments to geo-reference of russian texts. Set up and trained methods MaxEnt, MEMM, CRF and combined method for permits ambiguities. Methods compare of permission and recall