АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ТЕМАТИКИ СВЕРХКОРОТКИХ ТЕКСТОВ

AUTOMATIC CLASSIFICATION OF VERY SHORT TEXTS

Белов А.А. *(abelov@ashmanov.com),* ***Волович М.М.*** *(mv@ashmanov.com)*
«Ашманов и Партнеры», «Поисковые технологии», Москва

Подход, реализованный компаниями «Ашманов и Партнеры» и «Поисковые технологии», позволяет эффективно распознавать тематику поисковых запросов, заголовков и других сверхкоротких текстов при помощи единой базы терминов, которая используется и для распознавания тематики обычных текстов.

Распознавание тематики обычных текстов

Компаниями «Ашманов и Партнеры» и «Поисковые технологии» разработана и уже более двух лет успешно эксплуатируется система автоматического распознавания тематики текстов. Она предназначена для определения тематики веб-страниц и, шире, любых текстов на русском языке – например, рекламных объявлений. В действии систему можно увидеть, в частности, на новостном поисковике «Новотека» (http://www.novoteka.ru), где с ее помощью автоматически рубрицируются новости и новостные сюжеты.

Данная система распознавания тематики разработана в рамках инженерного подхода, или, иначе, подхода, основанного на знаниях: соответствие текста рубрике определяется в ней по факту вхождения в текст терминов – слов и словосочетаний, – заранее приписанных этой рубрике составителями. Такой подход требует существенных трудозатрат по наполнению словаря системы, однако обеспечивает более надежное распознавание по сравнению с технологиями, основанными на методах машинного обучения.

Кроме того, выбор инженерного подхода связан с тем, что для решения наших задач необходим универсальный, большой рубрикатор, создание которого методами машинного обучения пока что слабоосуществимо [1]. На сегодняшний день рубрикатор нашей системы содержит более полутора тысяч предметных и более трехсот географических рубрик, которые описываются набором более чем из двухсот тысяч терминов.

Описание рубрики в общем виде представляет собою список терминов, тематически принадлежащих данной рубрике (отметим, что булевские формулы для описания рубрик в нашей системе не используются). Для того, чтобы тексту была присвоена та или иная рубрика, он должен набрать достаточный вес по этой рубрике. При подсчете веса текста по рубрике учитываются:

- веса входящих в текст терминов (характерные термины получают более высокие веса в описании рубрики);
- длина терминов (в словах);
- количество вхождений каждого термина в текст;
- местоположение терминов (в заголовке или в основном тексте);
- длина текста, а также ряд других факторов.

Разумеется, тексту может быть одновременно присвоено и несколько рубрик; в этом случае можно сравнить веса различных рубрик для данного текста и, если необходимо, выбрать для него основную рубрику.

Распознавание тематики запросов и заголовков

Для сверхкоротких текстов – таких, например, как заголовки новостей или запросы к поисковым системам – методика распознавания, основанная на подсчете весов, неприменима. Даже если в запрос или заголовок и входят какие-либо термины из словаря системы (что, учитывая большой объем словаря, случается довольно часто), у них практически нет шансов набрать нужный для присвоения рубрики вес. (В числе редких исключений – случаи, когда в сверхкоротком тексте встречается достаточно длинный термин с высоким весом.)

С другой стороны, именно краткость таких текстов, как заголовки и поисковые запросы, практически исключает **случайное** попадание в них тематически маркированных слов и особенно словосочетаний. Если, например, на странице сайта есть слово *Ялта*, то это еще не означает, что сайт посвящен Крыму: упоминание может быть случайным, принадлежащим рекламному объявлению и т. п. Однако если слово *Ялта* встретилось

в заголовке или в запросе, можно почти с полной уверенностью заключить, что речь в них идет именно о Крыме. Таким образом, вхождение в сверхкороткий текст **единственного** термина, принадлежащего к некоторой рубрике, может быть достаточным основанием для того, чтобы эту рубрику ему приписать.

Особый случай – когда термин, соотносимый с той или иной рубрикой в изолированном виде, в определенных контекстах перестает иметь к ней отношение. Например, слово *обои* не должно быть использовано для того, чтобы приписать запросу *эротические обои для рабочего стола* рубрику «Отделочные материалы», поскольку в этом контексте с ремонтом оно уже не связано.

Итак, система распознавания тематики сверхкоротких текстов была настроена таким образом, чтобы для приписывания тексту рубрики хватало бы наличия в нем одного любого термина из этой рубрики. Однако, если в тексте встречаются два термина, один из которых полностью входит в состав другого, то срабатывает только более длинный термин, а более короткий игнорируется. Например, для запроса *эротические обои для рабочего стола* срабатывают термины *эротические обои* («Эротика»), и *обои для рабочего стола* («Скринсейверы, обои»), а термин *обои* («Отделочные материалы») игнорируется.

Адаптация базы терминов к работе со сверхкороткими текстами

Принципиально, что для распознавания тематики сверхкоротких текстов были использованы тот же рубрикатор и **та же база терминов**, которые используются нами и для распознавания тематики обычных текстов. Бинарные модули, осуществляющие распознавание, – разные, однако в их основе лежит один и тот же набор исходных данных. Основная сложность в адаптации этих данных к новым задачам была связана с повышением требований к наполнению базы терминов и необходимостью частично пересмотреть ее состав.

Во-первых, пришлось существенно почистить базу – найти и удалить ошибки, термины, недостаточно надежно связанные со своими рубриками и т. п. Работа со сверхкороткими текстами не только сделала такую чистку необходимой, но и помогла ее осуществить.

При распознавании тематики обычного текста нет строгой необходимости в том, чтобы все термины точно соответствовали рубрикам. Небольшой процент «шума» не сказывается критическим образом на качестве распознавания, а иногда даже может его улучшить (особенно если «лишние» термины относительно близки по тематике): у рубрики появляются дополнительные шансы набрать нужный вес. И как раз в силу того, что подобные ошибки мало влияют на результат распознавания, заметить их достаточно трудно.

Однако при работе со сверхкороткими текстами точное соответствие термина своей рубрике становится решающим фактором. Работа системы с короткими текстами сама по себе оказалась эффективным средством тестирования и чистки базы терминов. В этом режиме ошибочный термин сразу приводит к присвоению неправильной рубрики – а значит, ошибку легко заметить и исправить.

Во-вторых, большего, чем раньше, внимания к себе потребовали омонимичные и многозначные термины.

Например, термин *вирус* присутствует как в рубрике «Инфекционные заболевания», так и в рубрике «Компьютерная безопасность». Разумеется, в обычном случае это не приводит к тому, что компьютерные тексты получают медицинскую рубрику, а медицинские – компьютерную: одного термина, тем более однословного, для этого заведомо недостаточно. Но так как при распознавании сверхкоротких текстов одиночные термины срабатывают, нам пришлось: а) пометить термин *вирус* в рубрике «компьютерная безопасность» особым знаком, блокирующим его при сборке бинарной базы для сверхкоротких текстов, и б) проследить, чтобы те сочетания, в которых этот термин может встретиться в «компьютерных» запросах и заголовках (например, такие сочетания, как *вирусы скачать* или *мобильные вирусы*), попали в базу терминов.

В-третьих, появилась специальная рубрика «Исключения», которая содержит словосочетания, которые сами по себе не отнесены к содержательным рубрикам, но должны блокировать срабатывание более коротких терминов. Например, в нее включен термин *кодекс чести* – для того, чтобы запросу, содержащему такое словосочетание, не была приписана рубрика «Законодательство» (аналогичный прием используется, например, в системе LexisNexis, см. [2]).

Эффективность распознавания

Целенаправленная работа над базой терминов, которая с октября 2006 года ведется нами с ориентацией прежде всего на распознавание тематики сверхкоротких текстов, позволила достичь довольно высоких показателей. При этом качество распознавания улучшилось не только для запросов, заголовков и т. п., но и для обычных текстов.

Последнее тестирование системы было проведено на тридцати тысячах запросов, скачанных из «Прямого эфира» поисковой машины «Яндекс» (http://stat.yandex.ru/queries/last20.xml) 13 марта 2007-го года в 15:20 по

московскому времени. Предварительно база запросов прошла техническую чистку, в результате которой из нее были удалены:

- навигационные запросы (названия доменов): 1074 запроса;
- цифровые запросы (номера телефонов, почтовые индексы и т. п.): 228 запросов;
- запросы, на которые «Яндекс» выдает менее десяти страниц[1] (часто это либо запросы, набранные при ошибочной раскладке клавиатуры, либо запросы с серьезными опечатками): 805 запросов.

При тестировании на очищенной таким образом базе запросов – со средней длиной запроса в 3 слова – была показана полнота распознавания в 60 % (полнота здесь понимается как отношение распознанных запросов к общему числу запросов в базе). Отметим, что 60 % – это существенно ниже аналогичного показателя, получаемого, например, при стандартном распознавании новостных сообщений (90–93 %), однако стоит подчеркнуть, что данный уровень полноты был достигнут именно на живом запросном материале, прошедшем лишь сравнительно небольшую предварительную очистку.

Также было проведено тестирование системы на неоднословных запросах – как потенциально более содержательных. Общее число неоднословных запросов составило 22828 (отбор запросов проводился уже из предварительно очищенной базы), средняя длина запроса – 3,6 слова. Полнота распознавания на неоднословных запросах, как и предполагалось, оказалась несколько выше, чем на базе в целом – 66 %. Разумеется, по мере дальнейшей работы над рубрикатором показатель полноты будет еще расти.

Что касается точности распознавания, то по нашим оценкам доля ложных срабатываний в тестах не превышала 2–3%, т. е. точность была весьма высока. Такой показатель точности обусловлен осторожным подходом к отбору терминов: мы придерживаемся той позиции, что лучше не приписать запросу ни одной рубрики, чем случайно приписать ему нерелевантную рубрику.

Возможные применения

Среди возможных практических применений системы автоматического распознавания тематики сверхкоротких текстов можно назвать следующие:

Увеличение полноты рубрикации новостей и веб-страниц за счет распознавания тематики их заголовков. Иногда содержащихся в тексте характерных слов и словосочетаний недостаточно для того, чтобы можно было определить его тематику. Такое особенно часто случается с короткими текстами, состоящими из одного-двух абзацев. В подобных случаях тематику текста нередко можно установить, распознав тематику его заголовка.

Уточнение области поиска в поисковых системах. При выдаче результатов поиска имеет смысл повышать вес сайтов (страниц), тематика которых соответствует тематике запроса. Для подобных сайтов меньше вероятность того, что слова из запроса встретились на них случайно.

Показ контекстной рекламы в поисковых системах. Подбор рекламных объявлений в соответствии с распознанной тематикой запроса может быть более эффективен, чем на основе конкретных слов запроса. При таком подходе, с одной стороны, начинают приносить прибыль и такие запросы, слова из которых никто не купил, – при условии, что для них удалось определить тематику. С другой стороны, меньше вероятность того, что, например, по запросу *коксовый газ* будут выданы объявления типа *запчасти для автомобилей Волга*.

Показ контекстной рекламы посетителям, пришедшим с поисковых систем. В большинстве случаев известно, по какому запросу посетитель перешел с поисковой системы на сайт, размещающий контекстную рекламу. Подбор объявлений в соответствии с поисковым запросом может оказаться более эффективной стратегией, чем их подбор в зависимости от тематики самого сайта, поскольку запрос, как правило, более точно отражает интересы посетителя.

Список литературы

1. Dumais S., Lewis D., Sebastiani F. Report on the Workshop on Operational Text Classification Systems (OTC-02) // SIGIR-2002. Tampere, Finland, 2002. http://www.sigir.org/forum/F2002/sebastiani.pdf
2. Wasson M. Classification Technology at LexisNexis // SIGIR 2001. Workshop on Operational Text Classification. http://www.daviddlewis.com/events/otc2001/presentations/otc01-wasson-paper.txt

[1] Каждый запрос в «Прямом эфире» снабжен информацией о числе страниц, которое выводится «Яндексом» в ответ на данный запрос. Эта информация и была использована нами при фильтрации.