Технология поиска и сбора в Интернете текстов на малых языках России*

Л.Я. Зайдельман, И.В. Крылова, Б.В. Орехов
nevmenandr@gmail.com
Москва, Национальный исследовательский университет «Высшая школа экономики»

В работе описывается созданная авторским коллективом технологическая цепочка, позволяющая обнаруживать в Интернете, проверять и скачивать тексты на малых языках России. В дальнейшем получившиеся текстовые коллекции можно использовать для построения лингвистических корпусов и разработки программных продуктов, основанных на статистических методах распределения языковых единиц в текстах. Созданная технология предполагает автоматическое обращение к поисковой машине Яндекса с определённым типом запросов, кроссвалидацию выдачи, применение чёрных списков сайтов, выкачивание страниц, освобождение текста от html-разметки и определение языка текстового абзаца. Эти действия применяются как к большому Интернету, так и к экосистеме социальной сети VK.com. Результатом работы стали текстовые коллекции на малых языках народов России, размещённые в свободном доступе в Интернете.

Ключевые слова: *информационный поиск, малые языки, лингвистические корпуса, анализ текстов на естественном языке.*

The technology of web-texts collection of Russian minor languages*

Lyudmila Zaydelman, Irina Krylova, Boris Orekhov
National Research University The Higher School of Economics, Moscow, Russia

This paper describes the process that allows to detect on the web, check and download texts in minority languages of Russia. In the future, the resulting collections can be used to build a corpus and linguistic software. Technology includes an automatic query to the search engine Yandex, crossvalidation, black lists of sites, crowling, and language identification of a text paragraph. These technology applies to the Internet and to the social network VK.com. As a result, corpora of minority languages of Russia placed in free access on the Internet.

Keywords: *information retrieval, minority languages, linguistic corpora, natural language processing.*

Введение

Самым дорогостоящим этапом создания лингвистических корпусов является оцифровка печатных текстов. Существенное время и иные ресурсы уходят на сканирование, распознавание и вычитку текстов, не переведённых в электронный вид заблаговременно. Сейчас эта проблема уже неактуальна для широко распространённых языков. В случае с русским языком, например, получить большую текстовую коллекцию для создания корпуса очень легко благодаря Интернету. Сложности остаются только в той области, где у исследователей и корпусостроителей имеются специфические требования к текстам: например, в сфере исторических корпусов (нужно получить в коллекцию много текстов определённого исторического периода) или жанровых (требуются тексты определённого жанра). Но проблема создания текстовой коллекции остаётся актуальной для малых языков. Хотя история многих миноритарных языков России знает и большую периодику, и специализированные издательства, выпускающие печатную продукцию на этом языке (татарский, башкирский, удмуртский, якутский), эти тексты не оцифрованы, а, следовательно, недоступны для создания и пополнения лингвистических корпусов, а также непригодны для разработки специфического программного обеспечения, основанного на статистике распределения языковых единиц в текстах. К такого рода программному обеспечению можно отнести спеллчекеры, разного рода системы подсказок (скажем, при наборе с клавиатуры), системы определения тональности текста и многое другое.

В этой ситуации хорошим подспорьем для развития компьютерной лингистики малых языков России был бы способ находить и скачивать все тексты на данном языке, которые размещены в Интернете: на сайтах, в блогах и в социальных сетях. Последний класс текстов даже особенно ценен, так как фиксирует особую разновидность языка, не отражённую в академических грамматиках, в большинстве своём составленных ещё в советское время.

Однако такого способа до сих пор не существовало. Некоторые большие поисковые машины при определённых настройках позволяли задавать запросы так, чтобы выдача содержала в себе тексты на определённых языках, но возможность выбора этих языков была ограничена специфическим набором крупных идиомов, среди которых, разумеется,

Работа опубликована при финансовой поддержке РФФИ, грант 15-07-20370.

отсутствовали малые языки, в том числе и малые языки России. Поэтому мы разработали технологию, которая на данном этапе позволяет находить тексты на малых языках России и формировать коллекции, пригодные для дальнейшего использования как в теоретических исследованиях, так и в инженерных разработках. Все полученные коллекции размещены в свободном доступе в Интернете[1].

Запросы к поисковой системе

Чтобы найти тексты на каком-либо языке в Интернете, в идельном случае следовало бы сконструировать робота, который обошёл бы все страницы в Сети, проанализировал размещённые на них тексты и определил, какие страницы содержат слова на нужном языке. Однако создание такого сервиса в современных условиях доступно только крупным технологическим компаниям. Интернет стал слишком большим и краулер без дорогостоящей инфраструктуры серверов не справится с этой задачей за разумное время. Поэтому уместнее использовать индексы, которые уже составлены большими поисковыми компаниями в процессе обхода Интернета их роботами. Конечно, мы не можем рассчитывать на прямой доступ к индексным базам поисковиков, поэтому единственный выход — это обращение к выдаче поисковой машины через программный интерфейс. В нашем случае это Яндекс.XML[2]. Этот программный интерфейс имеет ограниение в виде 1000 запросов в сутки. Показ каждой новой страницы выдачи, содержащей по 10 пунктов) считается запросом, так что на обслуживание поиска всех страниц Интернете на всех малых языках России у нас ушло около 240 суток.

Поисковый движок принимает в качестве запроса слова (а не отдельные символы и не их сочетания), так что предварительно необходимо сформировать список слов для данного языка, каждое из которых поданное в качестве запроса к поисковику, позволило бы получить выдачу из максимального числа страниц, содержащих текст на данном языке. При этом тексты на других языках должны считаться «шумом», и их число в выдаче должно быть минимизировано.

Разумным было бы составление частотного списка слов для каждого языка и их пересечение с такими же списками других языков. В этом случае наиболее частотные слова из получившихся списков, не совпадающие со словами из других списков, стали бы хорошими кандидатами для того, чтобы находить по ним тексты. Однако не следует забывать, что у нас отсутствуют текстовые коллекции, которые позволили бы предпринять такую операцию. Следовательно, подбор поисковых терминов должен производиться вручную на основе лингвистической литературы (словари и грамматики). В слова-маркеры для каждого языка должны войти служебные слова: они достаточно частотны в любом языке, но при этом подойдут не любые служебные слова, а достаточно длинные в смысле числа букв, чтобы не совпасть со служебными словами или аббревиатурами в других языках.

К сожалению, не может помочь в составлении списков слов-маркеров и Википедия, так как все языковые разделы на малых языках России в большой степени сгенерированы автоматически и естественное частотное распределение слов в этих текстовых коллекциях нарушено [1].

Составленные списки слов-маркеров отправлялись в качестве запросов к Яндекс.XML. При этом были выставлены дополнительные настройки, заставляющие поиск воспринимать запрос как точный, потому что в противном случае Яндекс стремился понять слово малого языка как опечатку в написании русского слова и исправить. Это приводило к нерелевантным результатам поиска.

Несмотря на то, что большинство слов-маркеров позволяли получить нужную выдачу, эта выдача нуждалась в перепроверке. Первый этап перепроверки: выяснить, находятся ли те же самые сайты по другому слову-маркеру для этого языка. Второй этап перепроверки: находятся ли те же самые сайты по слову-маркеру другого языка. Эти сравнения выдачи позволяли выделить как «шумные маркеры», то есть слова, которые имеются не в одном, а сразу в нескольких малых языках (частый случай для этимологически близких языков), так и «шумящие сайты», то есть страницы, на которых имеются тексты сразу на нескольких языках.

Запросы производились таким образом, чтобы обнаружить не только страницы в Интернете вообще, но и специально на сайте vk.com. В этой социальной сети мы искали сообщества, в которых имелись бы тексты на интересующих нас языках.

Полученные списки доменов распределялись по нескольким группам в зависимости от того, много или мало страниц, содержащих текст на интересующем нас языке, на этом сайте обнаруживалось. Так, понятно, что на сайте youtube.com тексты на удмуртском языке не составляют основную часть текстовой информации, хотя несколько роликов и комментариев на удмуртском языке на сайте присутствуют. Кроме того, в Сети нашлось достаточно много страниц, не содержащих полноценных текстов на малых языках России, но содержащих отдельные слова этих языков. Среди них хостинги

[1] http://web-corpora.net/minorlangs/

[2] https://xml.yandex.ru/

музыки, сайты рефератов (в том числе рефератов лингвистических работ, посвящённым малым языкам России) и т.д. Для таких сайтов мы формировали чёрные списки, которые позднее применяли к выдаче поисковика.

Бытовая система письма

При составлении списка слов-маркеров, которые должны стать поисковыми терминами при запросе к поисковику, нужно учитывать такую особенность, как бытовое написание слов на малых языках. Многие языки содержат в своих алфавитах знаки, отличные от русского варианта кириллицы и, следовательно, отсутствующие на стандартной раскладке русской клавиатуры.

С одной стороны, наличие в слове-маркере специальной графемы, специфичной для орфографии данного языка и отсутствующей в русском алфавите, повышает вероятность того, что по этому слову будут находиться веб-страницы с текстами именно на данном языке. С другой стороны, это понижает полноту найденного, потому что в случае с малыми языками велико число случаев, когда текст в интернете написан в бытовой системе письма (термин введён А.А. Зализняком в применении к древненовгородскому диалекту [2]), в которой не воспроизводятся специфические для алфавита графемы, и пишущий обходится только теми знаками, которые присутствуют в стандартной кириллической или латинской раскладке клавиатуры.

Отличие от бытовой системы письма, описанной Зализняком, в том, что для древненовгородского диалекта характерно наличие некоторого класса графем, каждая из которых может свободно заменяться другой, относящейся к тому же классу. Например, графемы «ять», ь, е могут заменять друг друга на письме. В случае с бытовой системой письма в национальных интернетах имеет место только односторонняя замена специфических графем на графемы русского или латинского алфавита.

Так, чувашское çук «нет, не имеется» (исторически соответствует всем известному йок в башкирском и др. тюркских языках) — хорошее слово-маркер. Но оно содержит специфическую графему, и тех текстов, где пользователи пишут «щук», мы по нему не найдем.

Определение языка

Задача автоматического определения языка, на котором написан текст, в принципе, известна и большей частью решена. Однако имеющиеся решения, в основном, основаны на случаях, когда у исследователя уже есть коллекция текстов на данном языке достаточного объёма. Поэтому в нашем случае следовало выстроить решение по определению языка иначе. Мы исходили из условия, что на каждой странице, которую мы анализируем, имеется либо текст на интересующем нас языке, либо текст, демонстрирующий переключение кодов, то есть такой, в котором какая-то часть является строкой на русском языке, а другая часть — на заранее известном нам малом языке. То есть задача сводилась к отличению русского текста (статистические характеристики его известны) от нерусского с помощью распределения буквенных триграм. Последняя часть как раз и считалась текстом на малом языке.

Заключение

В процессе выполнения проекта нами вручную составлены списки поисковых терминов (слов-маркеров для языков России), которые отправлялись в качестве запроса к Яндекс.XML. Полученная выдача проверялась таким образом, чтобы удостовериться, что сайты, на которых нашлись страницы, действительно содержат тексты на интересующих нас языках (в частности, к этим сайтам задавались дополнительные запросы с другими словами-маркерами). Позднее все обнаруженные таким образом сайты были скачены, очищены от html-разметки и размещены в свободном доступе в Интернете.

Благодаря описанной работе на данном этапе мы можем строить технологию поиска и определения языка иначе, основываясь на статистической обработке текстовых коллекций, а не на ручном труде исследователей.

Литература

[1] Орехов Б. В., Решетников К. Ю. К оценке Википедии как лингвистического источника: сравнительное исследование // Современный русский язык в интернете. — М.: Языки славянской культуры, 2014. — С. 310—321.

[2] Зализняк А. А. Древненовгородский диалект / Изд. 2-е, переработанное с учётом материала находок 1995—2003 гг. — М.: Школа «Языки русской культуры», 2004. — С. 21-23 и далее.