А.Ю. Станкевич (Гродно, ГрГУ)

ТЕХНОЛОГИЯ СБОРА И СИСТЕМАТИЗАЦИИ ЭЛЕКТРОННОГО КОНТЕНТА ДЛЯ КОРПУСА РУССКОЯЗЫЧНЫХ СМИ ГРОДНЕНЩИНЫ

Иллюстративный корпус русскоязычных СМИ Гродненщины создан как инструмент выявления лексико-семантической и социокультурной специфики русской речи Гродненщины. В статье описаны сбор и систематизация электронного контента, то есть этапы, предшествовавшие металингвистической и собственно лингвистической разметке коллекции текстов и формированию ß-версии корпуса.

Сбор электронного контента осуществлялся в автоматическом режиме при помощи менеджера загрузок TeleportPro [1]. Среди доступных функций есть загрузка html-станиц по ключевому слову (текстовому фрагменту, определяемому через маску с использованием спецсимволов «?» со значением ‘один символ’ и/или «*» со значением ‘нуль и более символов’) и фильтрация загружаемых страниц по маске имени. Условно-бесплатная версия позволяет загружать не более 500 файлов с исходного адреса. С учетом этого ограничения и того факта, что годовой архив издания часто превышает порог 500 файлов, было решено загружать файлы частями за 1–4, 5–6 и 7–8 месяцы.

Была определена следующая стратегия загрузки электронных архивов: шаг 1) предварительная загрузка архива, определение маски для даты публикации, определение фильтра имени файла для исключения из загрузки страниц с анонсами статей и других служебных страниц; шаг 2) собственно загрузка частей архива; шаг 3) проверка ожидаемого и реального объема загруженного контента; шаг 4) выборочная проверка загруженных страниц малого веса и (редко) отсев таких страниц.

Приведем примеры масок для 1–4 месяцев и фильтров: маска: ??-04-2012;??-03-2012;??-02-2012;??-01-2012, фильтр: *index*.htm («Свислочская газета»); маска: ??.04.2012 ??:??.;??.03.2012 ??:??.;??.02.2012 ??:??.;??.01.2012 ??:??; фильтр: -m=*.htm («Берестовицкая газета»), маска: ??-04-2012, ??:??;??-03-2012, ??:??;??-02-2012, ??:??;??-01-2012, ??:??; фильтр: index*.htm; page,*.html («Праца»).

Ниже опишем этапы систематизации электронного контента.

Этап 1. Очистка загруженных архивов от элементов html-«обертки» и формирование файла, подготовленного для xml-представления в формате НКРЯ.

Основным инструментом стала программа ChainReplacer (разработчик: Лев Алексеевский), предоставленная российской стороной. Назначением программы является очистка текста от html-кода и извлечение доступной метаинформации по заданным шаблонам. Программа поддерживает пакетную обработку файлов. Среди доступных встроенных шаблонов: RemoveTags (удаляет все теги и их атрибуты, оставляя содержимое); blank —

пустой шаблон, задает отбивку после фрагмента текста; Trim (отсекает пробельные символы в начале и конце строки); PlainText (осуществляет сложное преобразование html документа в простой текст: удаляет теги, стили, скрипты, заменяет escape-последовательности типа «<» -> «<« и т.п.), шаблоны для работы с регистром и др. Пользовательский шаблон представляет собой пару [шаблон поиска]->[шаблон замены] и определяется через встроенные шаблоны и/или регулярные выражения. Уникальная для каждой газеты цепочка шаблонов, разделенных ограничителями «++», сохраняется в ptn-файле и используется для пакетной обработки файлов этой газеты.

Метаинформация, извлекаемая из html-кода, в общем случае была такова: название газеты, автор материала, заголовок, дата создания материала, рубрика, url материала, язык материала. Ниже приведено содержимое ptn-файла для «Берестовицкой газеты», встроенные шаблоны в коде выделены полужирным шрифтом:

<title>(.*)</title>->\1++«->“++»->”++–->–++…->…++->\1
blank
->@au
<h1>(.*)</h1>->\1++RemoveTags++«->“++»->”++–->–++…->…++^(.*)$->@ti \1
.*([0-9]{1,2}\.[0-9]{1,2}\.[0-9]{4}).* ->@da \1
<div class="post-date-single».*Рубрика:.*rel="category">(.*)\ ->\1 ++-> | ++PlainText++Trim++\|$->++Trim++\| ,->|++Tags: ->++\| ->| ++\| ->| ++->@topic \1
<h1>.*tppabs="(.*)"->@url \1
blank
(.*)<div style->\1++PlainText

Для всех газет цепочка шаблонов для извлечения названия газеты размещается в абсолютном начале ptn-файла.

В ptn-файле для «Берестовицкой газеты» метка автора (@au) не определена, так как отсутствуют устойчивые способы оформления позиции автора в исходном html-коде страниц этого издания; позиция заполняется вручную. Метки названия статьи (@ti), рубрики (@topic), url статьи (@url) определены цепочкой шаблонов и заполняются автоматически. В приведенном выше коде не определена метка языка. Метка языка (@lang) также заполнятся автоматически, но на вход идет не html-файл, а файл, уже очищенный от html-«обертки». Метка ставится только для текстов, которые содержат символ «ў» в теле статьи (а не в названии газеты). Ниже приведено содержимое ptn-файла для установки метки языка:

```
^(.*@url.*)\n->\1
@url.*ў.*->@lang be
blank
@url.*\n(.*)$->\1
```

Заметим, что в некоторых случаях метка автора могла быть определена форматом: (.*) ->\1++Removetags++&#[0-9]+;->++\n->++Trim++, -> | ++\.$->++->@au \1. Этот фрагмент кода использовался в ptn-файле для «Островецкой правды», поскольку в этой газете имя автора хотя и не размечалось специальным html-тегом (например, <span

class="createby">…), но регулярно набиралось с выравниванием по правому краю. В приведенном фрагменте помимо удаления escape-последовательностей (&#[0-9]+;->), задается замена запятых на знаки вертикального слеша и удаление конечной точки. Необходимо учитывать, что если в исходном html-файле с выравниванием по правому краю и через запятую были набраны имена соавторов, то результат обработки не требует коррекции, если же, к примеру, с таким выравниванием было набрано имя автора и (через запятую) его должность — результат коррекции требует.

Этап 2. Постобработка выходных файлов ChainReplacer.

Постобработка осуществлялась в ручном, автоматизированном, автоматическом режимах. В ручном режиме проводилось:

1) редактирование зоны меток: а) метки автора: заполнение метки автора, если содержимое не извлечено автоматически по шаблону; б) метки языка: снятие метки @lang be в преимущественно русскоязычном тексте (в этих случаях метка языка была установлена автоматически по неравному тексту/микротексту фрагменту, например: названию, цитате);

2) редактирование зоны основного текста (простановка точек в конце заголовков, не размеченных спецтегом <h…> в исходном html);

3) разбивка на части файлов, состоящих из текстов на разных языках и/или принадлежащих разным авторам (в случае нерегулярной верстки html-страницы).

В автоматизированном режиме проводились:

1) постпроверки по регулярным выражениям AntConc [2], целью постпроверок было выявление фрагментов, некорректно обработанных шаблоном из-за нерегулярной верстки html-страницы;

2) исправление неверной капитализации в имени автора (то есть записей типа *Мария ИВАНОВА* вместо *Мария Иванова*) Инструмент: python-модуль;

С учетом того, что в электронных газетах Гродненщины в html-метке автора часто индексируется название организации и / или аббревиатура, обрабатывать содержимое этой метки доступными в программе ChainReplacer функциями автоматического изменения регистра нецелесообразно. Принятая стратегия обработки такова: шаг 1) автоматически создается текстовый файл с данными «метка автора — имя файла»; шаг 2) через автозамену ставится верная капитализация в фамилиях (поскольку в газете имена корреспондентов повторяются, процесс невремяемкий); шаг-3) данные «исправленная метка автора — имя файла» идут на вход python-модуля, содержимое меток автора в соответствующих файлах заменяется верно капитализированной записью.

3) исправление неверной капитализации в основном тексте (такая неверная запись встречалась в исходном массиве текстов в случаях, когда прописные буквы использовались как средство выделения части текста); при этом сохранялась капитализация в следующих случаях: а) капитализация в аббревиатурах; б) капитализация в надписях: *Вверху по кругу надпись: «БЕЛАРУСКАЯ ЧЫГУНКА»;* в) капитализация в названиях

фирм/механизмов (поскольку они потенциально могут быть аббревиатурами): *Еще одним объектом, который с приятной миссией награждения был посещен 6 августа, стала зерносушилка «МЕКМАР» в ЧП «Новый Двор-Агро»;* г) капитализация как средство языковой игры: *Если вы талантливЫ*;

В автоматическом режиме проводились:

1) очистка текста от двойных абзацев и двойных пробелов; инструмент: разработанный нами python-модуль;

2) удаление дублей статей (дублем называем материал, по той или иной причине размещенный на сайте газеты под разными именами); инструмент: разработанный нами python-модуль.

Типичные случаи постпроверок, проводимых с помощью AntConc, даны в таблице.

Таблица

Описание типичных постпроверок

Регулярное выражение	Значение. Комментарий
[A-z][A-яЁёЎІў] [A-яЁёЎІў][A-z]	Смесь кириллица-латиница, смесь латиница-кириллица
\b([A-яёЁўіЎІ]\s){3,}	Граница слова + 3 и более сочетания «буква кириллицы и пробел». Цель: выявление разрядок, заданных через пробелы (по аналогии задан шаблон для латиницы)
[^.]\.{2,}\s*\n	Не точка + 2 и более точки + нуль и более пробелов + абзац. Цель: проверки лишних точек.
[!?,:;…]\.\s*\n	Любой знак из набора !?,:;… + точка + нуль и более пробелов + абзац. Цель: проверки лишних точек.
[A-zA-яеЕўіЎІ]\s{1,} \n	Любая буква из набора A-zA-яеЕўіЎІ + один и более пробелов + абзац. Цель: проверка абзаца без знака препинания на конце.
[a-eўі][^.?!…:;]\s*\n[^@\s] Доп.: [a-яеўі]\s*\n\s+[^@]	Любая буква из набора a-eўі + любой знак не из набора .?!…:; + нуль и более пробелов + абзац + не пробел/не @. Цель: проверка оборванных абзацев
\b[A-ЯЎІ]{3,}	Граница слова + 3 и более буквы кириллицы в верхнем регистре. Цель: проверка неверной капитализации (только после исправления капитализации в метке автора)

В результате реализации второго этапа был получен массив текстов, подготовленных к переводу в xml.

Этап 3. Конвертация массива очищенных от html-обертки текстов в xml.

Конвертация массива очищенных текстов в формат xml проводилась с помощью предоставленного российской стороной целевого шаблона программы ChainReplacer в автоматическом режиме.

Этап 4. Сортировка xml-файлов по каталогам заданной структуры.

Сортировка проводилась в автоматическом режиме, раздельно для файлов с меткой @lang be и файлов без таковой метки, средствами написанного нами Python-модуля. Принцип сортировки: в папке с именем

газеты размещаются подчиненные папки по месяцу создания материала с соответствующим содержимым.

Таким образом, нам удалось частично автоматизировать процесс подготовки массива текстов для последующей метаразметки.

Подготовлено в рамках проекта, реализуемого при поддержке БРФФИ (договор № Г13Р-050 от 16.04.2013).

ЛИТЕРАТУРА

1. TeleportPro [Electronic resource]. – Mode of access: http://www.tenmax.com/teleport/pro/download.htm. – Date of access: 10.03.2014.

2. AntConc [Electronic resource]. – Mode of access: http://www.antlab.sci.waseda.ac.jp/antconc_index.html. – Date of access: 10.03.2014.