А.С. Гребеньков

ИНТЕРНЕТ-РЕСУРС WWW.WORDFORM.RU КАК ИНСТРУМЕНТ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ ТЕКСТОВ

Одной из ключевых проблем в подготовке текстового корпуса является лингвистическая разметка текстов. В этой статье будет предложено практическое решение проблемы *морфологической разметки* текстов, а также описан процесс использование Интернет-ресурса **www.wordform.ru**, лежащего в основе предлагаемого метода.

Под морфологической разметкой будем понимать определение по заданной словоформе ее словарной леммы, частеречной принадлежности, а также определение ее морфологических характеристик (род, число, время и т.п.) – граммем.

Проблема морфологического анализа затратна по времени и ресурсам, Но без преодоления этого этапа создание полноценного корпуса представляется авторам невозможным. Таким образом, подготовка корпуса предполагает либо создание собственного морфоанализатора, либо использование уже существующих решений.

На сегодняшний день исследователь ограничен в выборе доступных морфоанализаторов[1]. Существующие решения можно разделить на платные и бесплатные продукты.

Если рассматривать бесплатные морфоанализаторы (скомпилированные модули или модули в исходном коде), то, с точки зрения авторов, основными причинами, снижающими удобство,

[1] *Коваль С.А.* К унификации представления русской морфологии в системах обработки текстовой информации // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2002». М., 2002. // URL: **http://www.dialog-21.ru/materials/archive.asp?id=7570&y=2002&vol=6078**

надежность и доступность этих решений, являются следующие моменты:

1) закрытость морфомодулей (в большинстве случаев, они поставляются как «черные» ящики, в которых очень сложно что-либо поменять);

2) ориентированность морфомодулей на настольные приложения;

3) сложность задачи («доведение» морфологии до серьезного уровня требует много времени и ресурсов);

4) сложность использования (процесс «прикручивания» этих модулей к проекту зачастую довольно сложен и требует определенной квалификации).

Эти и другие проблемы[1], а также отсутствие доступного решения, которое можно было бы легко и быстро адаптировать к поставленной задаче (в данном случае, разметки корпуса), привело авторов к разработке иного технологического подхода, который представлен Интернет-ресурсом **www.wordform.ru**.

По сути дела, этот ресурс является *централизованной базой данных порожденных словоформ русского языка в Интернете*.

В основе этого технологического подхода лежат несколько положений, которые, как представляется авторам, определяют новизну такого решения задачи морфологического анализа.

1. Почему использовать уже *порожденные словоформы*, а не распознавать каждую словоформу на входе как неизвестную?

Морфологический анализ, по своей сути, является одноразовой задачей. Если на вход подается словоформа *книгами*, то на выходе пользователь должен всегда получать ответ: {лемма: **книга**; часть речи: **сущ**; граммемы: **мн. ч., тв. п., ж. р.**}. Таким

[1] *Гребеньков А.С.* Использование словаря порожденных словоформ для централизованного решения проблемы морфологического анализа. Интернет-ресурс **www.wordform.ru** // Материалы XXXVII Международной филологической конференции. СПб, 2008. [в печати].

образом, промежуточные операции между запросом на входе и ответом на выходе являются, по сути, избыточными, так как всякий раз по запросу определенной словоформы существует строго определенные ответы. Следовательно, минимальное количество операций между входом и выходом будет иметь место при использовании словаря типа *hash*, когда по ключу выдается строго определенный ответ. В данном случае подобным словарем будет словарь порожденных словоформ.

книгами → **книга**
книгою → **книга**
…
плавал → **плавать**
плавать → **плавать**

При использовании подобного словаря процедура анализа словоформы (что, собственно, и происходит в классическом морфоанализаторе) заменяется процедурой поиска в словаре.

Необходимо отметить, что некоторые морфоанализаторы предлагают механизм предсказания: если запрошенная словоформа не может быть строго приведена ни к одной из лемм, то строится гипотеза, которая может быть как верной, так и ложной. Качественные морфоанализаторы обладают высокие показатели в предсказании характеристик слова (около 90%[1]). Тем не менее, существует определенный процент слов с неправильными парадигмами и исключениями (типичными примерами являются словоформы *мечт* и *победю*, которые могут быть распознаны как правильные). Этот процент «неправильных» словоформ всегда будет оставаться проблемным местом для автоматических морфоанализаторов. Надежным решением этой проблемы, исклю-

[1] *Сокирко А.В.* Морфологические модули на сайте **www.aot.ru** // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2004». М., 2004. // URL: **http://www.dialog-21.ru/Archive/2004/Sokirko.htm**

чающим сам механизм гипотез, является опять-таки использование списка уже порожденных словоформ.

2. Почему необходим *централизованный* подход в решении морфологического анализа?

Морфоанализатор в виде конечного автомата изначально строится для фиксированного количества лемм с их словоформами. Для распознавания большего числа слов необходимо пополнение словаря (либо использование механизма предсказания, который не дает полного контроля над анализом). Разные группы исследователей занимаются пополнением словарей своих морфоанализаторов децентрализовано. Очевидно, что наиболее эффективной процедурой пополнения словаря словоформ будет централизованное решение, когда различные проекты пользуются общим словарем и пополняют его.

3. Почему наилучшей формой хранения словаря словоформ является *база данных в Интернете*?

Для соблюдения условия централизованности словарь не должен загружаться в виде копии пользователям, но должен постоянно находиться в Интернете. Наиболее простым и прозрачным решением является использование баз данных. Естественно, если говорить в терминах низкоуровневой реализации, морфоанализатор обычно представляет собой конечный автомат, который, в свою очередь, можно считать «сложным» словарем. Но концептуально эти два подхода (морфоанализатор в виде конечного автомата и использование словаря порожденных словоформ) представляются авторам существенно отличающимися. Конечный автомат после построения является «черным ящиком», словарь же в виде базы данных является «прозрачной» во всех отношениях структурой, с которой комфортно работать и человеку, и машине.

Эти положения определяют технический формат предлагаемого решения. Из соображений «прозрачности» и простоты была выбрана система баз данных MySQL, которая является де-

факто одной из наиболее распространенных систем баз данных в мире, а язык запросов SQL идеально подходит для хранения и извлечения этой лингвистической информации.

Для реализации этой идеи был создан Интернет-ресурс **www.wordform.ru**. Технически он представляет собой информационный сайт, открытую базу данных порожденных словоформ (MySQL) и набор интерфейсов для работы с базой данных на различных языках программирования (PHP, Java, Python и др.). На сегодняшний момент завершены не все пункты проекта, но основная часть – словарь словоформ и база данных готовы и функционируют.

На данный момент словарь, размещенный на **www.wordform.ru** содержит 45 000 слов изменяемых частей речи (NOUN, VERB, ADJ), что дает около 2 500 000 словоформ:

- 28 000 существительных (105 классов словоизменения), порождается ~12 словоформ на каждую лемму;
- 10 000 глаголов (120 классов словоизменения), порождается ~104 словоформы на каждую лемму;
- 7000 прилагательных (34 класса словоизменения), порождается ~44 словоформы на каждую лемму;
- словарь имен собственных (около 2000);
- словарь неизменяемых частей речи (около 2000).

Предполагается постоянное пополнение словаря за счет словоформ, запрошенных пользователями и не найденных в существующем словаре.

Интерфейсы работы с MySQL представляют собой наборы функций. Например, на Java основными функциями являются:

String[] getLemma(String wordform);
(возвращает набор лемм, соответствующих словоформе),

String getPOS(String lemma);
(возвращает часть речи данной леммы), и т.д.

В дальнейшем предполагается открытие баз данных для прямого обращения к словарям.

Работа с ресурсом строится следующим образом. Наиболее типичной задачей является распознавание словоформ в потоке текста (могут осуществляться и другие менее тривиальные задачи). Пользователь (будь то программа или человек) запрашивает базу данных словоформ через стандартные интерфейсы. На вход подается текст (по сути, набор словоформ). В ответ пользователь получает размеченный текст с выраженными в явном виде леммой, частеречной принадлежностью, морфологическими показателями каждой словоформы. Если словоформы в словаре нет, то она поступает на обработку лингвисту, после чего в словарь добавляется соответствующая ей лемма, которая пропускается через морфогенератор для порождения словоформ. Чем больше текстов будет «пропущено» через ресурс, чем больше словоформ будет запрошено, тем быстрее словарь будет пополняться и тем меньше будет выдаваться нулевых результатов.

Например, при запросе фразы *На берегу пустынных волн стоял он, дум высоких полн*, будет выдан следующий результат:

На [**на,** PREP]
берегу [**берег,** NOUN, m, L, sg]
пустынных [**пустынный,** ADJ, G, pl]
волн [**волна,** NOUN, f, G, pl]
стоял [**стоять,** VERB, Dm (past masculine)]
он [**он,** PRONOUN]
дум [**дума,** NOUN, f, G, pl]
высоких [**высокий,** ADJ, G, pl]
полн [**полный,** ADJ, m, S (short form)]

Задача пользователя – интерпретировать и использовать полученную информацию в своих целях.

Как уже было сказано, можно выделить два основных вида использования этого ресурса:

I. Встраивать морфологию в качестве промежуточного модуля в проект автоматической обработки текста. Когда обращение к ресурсу происходит в компьютерной программе через программные интерфейсы:

Input: данные (слово, предложение, текст и т.п.)
1. Operation *A*
2. Operation *B*
…
n. Operation: обращение к **www.wordform.ru** (морф. анализ)
…
n+k. Operation *N+K*
Output

II. Обрабатывать тексты непосредственно на сайте. Для этого необходимо загрузить на сайте нужный текст и, дождавшись результата обработки, скачать к себе на компьютер размеченный текст.

В обоих случаях можно выбрать или самостоятельно задать предпочтительный формат разметки (например, xml или линейный текст).

Следует отметить, что в настоящий момент в проекте никоим образом не решены проблемы снятия омонимии: в ответ на неоднозначную по анализу словоформу будут выданы все найденные варианты анализа (например, *печь* как существительное и как глагол; *собаки* – мн.ч. им.п., ед.ч. род.п.). Ответственность за интерпретацию ложится на пользователя.

Практическая ценность ресурса **www.wordform.ru** в том, что он является готовым механизмом для быстрого прохождения морфологического этапа, так как позволяет простейшим образом интегрировать в программу уже готовое решение. Соответственно, этот ресурс может быть использован для решения многих практических задач в области автоматической обработки текстов, в том числе и в корпусной лингвистике.

Для русского языка, насколько известно автору, аналогов подобному ресурсу не существует. Мы надеемся, что этот ресурс заинтересует исследователей, и в дальнейшем будет развиваться и совершенствоваться.