Открытый корпус: принципы работы и перспективы

Д. В. Грановский
Mathlingvo
dima.granovsky@gmail.com

В. В. Бочаров
Mathlingvo
victor.bocharov@gmail.com

С. В. Бичинева
СПбГУ
sv.bichineva@gmail.com

Аннотация

Открытый корпус (OpenCorpora) — проект по созданию размеченного корпуса текстов на русском языке, доступного для исследователей в полном объеме и редактируемого пользователями, который призван решить проблему отсутствия подобных русскоязычных ресурсов. В статье описываются компоненты системы (хранилище, интерфейс разметки, подсистема экспорта), организация данных и жизненный цикл текста: добавление в корпус, автоматический разбор при помощи словаря, снятие неоднозначности пользователями. Большое внимание уделено минимизации порога вхождения в проект для новых пользователей.

1 Введение

В настоящее время проблема отсутствия корпусов текстов на русском языке, в том числе размеченных, казалось бы, не стоит. Более того, некоторые корпусы доступны в Интернете (см. обзор в [4]), что, на наш взгляд, существенно повышает их ценность для лингвистического сообщества. Насколько можно судить, под доступностью в Интернете обычно понимается наличие некоторого интерфейса, посредством которого пользователь может производить поиск в корпусе по разным параметрам. Безусловно, это открывает большие возможности для разнообразных теоретических исследований в области языка, для которых требуется анализ контекстов употребления слов, их частотности и т.д. Однако, как нам кажется, для использования корпуса в качестве ресурса для обучения или тестирования прикладных разработок — например, морфологических парсеров или систем снятия неоднозначности — этого недостаточно, поскольку в таких случаях требуется доступ непосредственно к размеченным текстам, а не к результатам поиска по ним. Получить такую разметку из существующих корпусов в принципе возможно, однако это связано со значительными трудностями — как технического, так и административного или лицензионного характера.

Существует и другая проблема. Обычно разметка корпуса производится закрытым коллективом профессиональных языковедов, что, как предпола-

гается, должно обеспечивать высокое качество разметки. На деле можно заметить, что даже лучшие корпуса не избавлены от ошибок и, что гораздо хуже, очень часто проявляют отсутствие единообразия в разметке.

Такое положение вещей послужило мотивацией для проекта «Открытый корпус» (OpenCorpora). В его рамках предполагается создание размеченного (в ближайшем будущем морфологически, впоследствии также синтаксически и семантически) корпуса русскоязычных текстов, содержимое которого доступно всем желающим под свободной лицензией, а разметка осуществляется сообществом пользователей. Эта модель редактирования уже зарекомендовала себя во многих проектах, самый известный из которых, вероятно, Википедия [3].

Настоящий проект отличается от существующих рядом особенностей, наиболее важными из которых являются:

- доступность под свободной лицензией всех имеющихся материалов (под доступностью мы понимаем не только наличие материалов в сети Интернет, но и предоставление возможности скачать эти материалы);
- веб-интерфейс редактирования, открытый неограниченному кругу потенциальных участников (для начала работы необходимо зарегистрироваться);
- низкий порог вхождения для новых участников (достаточно вносить правки только в тех случаях, в которых участник разбирается и понимает правила разметки);
- ведение истории изменений словаря и разметки с указанием авторства (позволяет вести обсуждение конкретных правок участников проекта, находить систематические ошибки и быстро их исправлять);
- возможность оценки работы каждого участника и предоставления обратной связи при помощи истории правок и функции отмены правки;
- интеграция разметки со словарем (морфологическая разметка текстоформы привязана к словарю лексем не через словарную форму слова, а через идентификатор леммы в словаре, что позже позволит отражать в разметке случаи лексической многозначности);
- наличие стандарта разметки, описывающего правила интерпретации языковых явлений средствами разметки.

Стремясь снизить порог вхождения участников в проект, мы неизбежно сталкиваемся с вопросом качества их работы и качества получаемого результата. Для решения этого вопроса мы предусмотрели следующие инструменты:

- разграничение уровней доступа к словарю и к разметке (поскольку изменения в словаре приводят к изменению всех текстов в корпусе, доступ к редактированию словаря предоставляется ограниченному числу опытных участников);
- явное указание на статус разметки каждого предложения:
  - размечено только автоматически без снятия омонимии;

– омонимия снята частично автоматическими правилами;

– омонимия снята частично участниками проекта;

– омонимия снята полностью;

– омонимия снята полностью и проверена модератором.

Разграничение доступа и возможность отката изменений позволяет поддерживать целостность данных, а указание статуса разметки дает возможность пользователю выбирать необходимый для его задач уровень качества материала. При этом сохраняется возможность быстрого вхождения в проект новых участников.

2 Компоненты системы и техническая реализация

Очевидно, центральная часть любого корпуса текстов — это хранилище, в котором содержатся исходные тексты, разметка и, возможно, что-то еще (в нашем случае это словарь). Минимальные требования к хранилищу — по-видимому, возможность добавления новых данных, хранение имеющихся данных без потерь и выдача данных по каким-либо запросам. Кроме того, вовлечение большого количества участников в работу по созданию корпуса влечет необходимость создания эффективных механизмов управления вносимыми в материал изменениями: ведения истории правок и функции отката, а также группировки изменений в пакеты.

Второй главный компонент — пользовательский интерфейс. В нашей модели редактирования корпуса, очевидно, существует два вида пользователей — «редакторы» и «потребители», — и необходимо учитывать интересы обеих групп. Интерфейс для «потребителей» в сущности должен обеспечивать удобную навигацию и поиск. Основные требования к интерфейсу редактирования разметки — интуитивно понятная форма представления лингвистической разметки предложения и удобство редактирования. Это действительно важно, поскольку для формирования сообщества требуется, чтобы порог вхождения участников в проект был по возможности невысок. Очевидно, что простейший способ организовать совместное редактирование — использование Интернета, поэтому интерфейс должен быть веб-интерфейсом.

В качестве еще одного компонента можно выделить систему экспорта данных. Экспорт (или выгрузка) данных отличается от поиска, во-первых, тем, что такие данные предназначаются для скачивания и последующей (скорее всего, машинной) обработки, а не для просмотра, а во-вторых, потенциальным отсутствием поискового запроса, т. е. в выгрузку попадают все данные (хотя, возможно, и отфильтрованные по некоторому признаку).

На первом этапе мы выделили 4 задачи, которые система должна решать:

1. доступ к словарю (чтение, редактирование, экспорт),

2. доступ к добавлению новых текстов в корпус и редактированию имеющихся,

3. автоматический морфологический разбор новых текстов при помощи словаря,

4. поддержка интерфейса для ручного снятия морфологической неоднозначности.

В качестве инструментов для решения этих задач мы рассматривали несколько «готовых» программных продуктов, объединяющих в себе все требуемые компоненты. В частности, предполагалось использовать механизм MediaWiki, на базе которого работает Википедия. MediaWiki включает в себя систему работы с SQL-хранилищем, в том числе историю правок, внутренний язык разметки (вики-разметка) и многое другое.

Однако позже мы были вынуждены отказаться от этой идеи и начать разработку собственной системы. Это было продиктовано следующими связанными соображениями:

- в MediaWiki без дополнительной разработки можно хранить только тексты в вики-разметке, связанные между собой ссылками или путем категоризации, а большинство элементов корпуса хранить в таком виде нецелесообразно с точки зрения последующей автоматической обработки;

- все равно требуется разрабатывать пользовательский интерфейс для редактирования словаря, иначе слишком высок порог вхождения;

- внесение изменений в MediaWiki весьма трудоемко.

Как и MediaWiki, наша система написана на языке PHP, в качестве СУБД используется MySQL.

3 Организация данных в корпусе

Размечаемые тексты в корпусе имеют определенную структуру. Самая крупная единица этой структуры — книга. Это название в большой степени условно, тем более что книги образуют собственную иерархию: например, в книгу «Война и мир» войдет еще 4 книги — по одной на каждый том; каждую статью Википедии мы тоже считаем отдельной «книгой».

Книга делится на абзацы, абзацы — на предложения, а предложения — на токены.

Деление на абзацы берется из источника и введено в основном для удобства добавления текстов, хотя нельзя исключать, что эта информация будет полезна и для исследователя.

Деление абзаца на предложения производится автоматически и проверяется человеком. Заголовок (например, главы), если он есть, считается входящим в ближайший абзац в качестве первого предложения; в его конце знаки препинания не восстанавливаются.

Деление предложения на токены также производится автоматически и проверяется человеком. Токеном мы считаем минимальную значимую последовательность символов без пробелов. Порядок абзацев, предложений

в пределах абзаца и токенов в пределах предложения сохраняется. Можно разделить все токены на словарные (формы, присутствующие в словаре) и несловарные. К несловарным относятся знаки препинания, Интернет-адреса, формулы (химические, математические и прочие) и другие сочетания знаков, которые нет смысла помещать в словарь, например, ввиду их неограниченного количества.

Размечаемой единицей корпуса является токен. Разметка токена состоит из одной или нескольких (в случае омонимии) интерпретаций. Каждая интерпретация обязательно содержит указание на класс токена (словарный, несловарный). Для словарных токенов интерпретация также включает:

- идентификатор леммы из словаря,
- часть речи,
- набор значений обязательных для данной части речи грамматических категорий (например, число для имен существительных),
- набор меток, обозначающих особенности конкретного употребления словоформы в тексте (например, «опечатка», «безличное употребление глагола»).

4 Жизненный цикл текста в корпусе

Поскольку открытый корпус задуман как постоянно пополняемый и развивающийся проект, в работе над которым участвует теоретически неограниченное количество людей, особый интерес представляет жизненный цикл материалов, включенных в корпус.

В корпус могут быть включены тексты, опубликованные под лицензией, совместимой с CC-BY-SA [1]. В качестве одного из атрибутов текста в корпусе указывается источник, из которого был получен текст. Перед добавлением в корпус новый текст проходит вычитку, чтобы избежать копирования опечаток. Расстановка границ предложений и абзацев производится на данный момент вручную и должна повторять положение этих границ в исходном тексте. После этого текст помещается в форму и отправляется на сервер для добавления. При отправке все токены проверяются по словарю. Если в тексте обнаруживаются несловарные токены, являющиеся при этом цепочками кириллических букв, то их список будет показан пользователю, а текст не будет добавлен. Для добавления такого текста в корпус необходимо будет отредактировать словарь так, чтобы все эти токены могли получить словарную интерпретацию.

Морфологический словарь взят из проекта AOT/Dialing [2]. Морфологический стандарт корпуса отличается от стандарта, принятого в AOT, поэтому потребовалось изменить грамматическую интерпретацию слов и разделение их на парадигмы. Эта работа выполняется автоматически на основе правил, имеющих вид «заменить некоторый набор граммем на другой при данных условиях» или «разбить некоторую парадигму на несколько по данным критериям», и не завершена на настоящий момент.

Добавление текста сопровождается его автоматическим разбором, при этом для каждой текстоформы генерируются все возможные варианты разбора (в том числе, ни одного, если, по мнению системы, это не слово). После

этого на основе эвристик системой разбираются некоторые текстоформы, изначально не получившие привязки к словарю (например, сокращения). Затем может следовать этап полуавтоматического снятия простых случаев неоднозначности (не реализованный на настоящий момент). Наконец, в дальнейшем редактирование осуществляется пользователями, которые могут удалять неверные разборы или добавлять отсутствующие. Подразумевается, что в большинстве случаев (не во всех) каждая текстоформа должна быть разобрана единственным образом.

Все тексты, добавленные в корпус, становятся сразу доступны для просмотра и редактирования средствами веб-интерфейса. Кроме этого, все тексты вместе с их разметкой включаются в файлы экспорта корпуса, доступные для загрузки. Файлы экспорта пересоздаются заново на регулярной основе и, таким образом, отражают актуальное состояние данных в корпусе. Экспортируемые данные предназначены для разработки и тестирования программного обеспечения, обрабатывающего текст на естественном языке, корпусных исследований и могут быть использованы в других задачах, где нужны морфологически размеченные тексты.

5 Заключение

Мы считаем, что открытость лингвистических баз данных является существенным стимулом для развития науки о языке и для создания прикладных систем обработки текста. Вместе с тем, создание таких баз требует больших трудозатрат, что существенным образом влияет на рентабельность разработки ПО и на трудоемкость исследований. В объединении усилий сообщества для создания открытого размеченного корпуса мы и видим решение вышеозначенной проблемы.

Список литературы

[1] Creative Commons — Attribution Share-Alike 3.0 Unported. http://creativecommons.org/licenses/by-sa/3.0/.

[2] АОТ. Автоматическая обработка текста. http://www.aot.ru.

[3] Википедия — свободная энциклопедия. http://ru.wikipedia.org.

[4] *Резникова, Т. И.* Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) / Т. И. Резникова, М. В. Копотев // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. — М.: Индрик, 2005.