Программное обеспечение для коллективной работы над морфологической разметкой корпуса

В. В. Бочаров
OpenCorpora
bocharov@opencorpora.org

Д. В. Грановский
OpenCorpora
granovsky@opencorpora.org

OpenCorpora[1] — это проект по созданию силами сообщества аннотированного корпуса текстов на русском языке, который можно легально распространять.

Аннотированные корпуса текстов необходимы для реализации многих лингвистических проектов, включая как студенческие работы, так и более зрелые исследования. В проектах, подразумевающих машинное обучение или тестирование алгоритмов обработки текста, корпус должен быть доступен целиком в машиночитаемом формате, в то время как большинство существующих на данный момент корпусов доступны только через веб-интерфейсы поиска по корпусу, выдача которых не предназначена для автоматической обработки. Одной из причин такого положения вещей являются лицензионные ограничения на тексты литературных произведений, материалы СМИ и другие источники. Распространение текстов требует согласия правообладателей, для получения которого необходимо с этими правообладателями как минимум связаться, что само по себе потребует определённых затрат времени, если речь идёт о большом числе источников.

Создание аннотированного корпуса текстов, включающего только материалы, распространение которых не запрещено условиями их лицензии, представляется нам решением этой проблемы: все тексты и аннотацию можно будет сразу помещать в свободный доступ. Это также может способствовать улучшению качества корпуса, т. к. все желающие получат возможность изучать и оценивать опубликованный материал, в том числе и методами, подразумевающими автоматическую обработку всего материала. Предоставление он-лайнового доступа к интерфейсу редактирования позволит оперативно исправлять найденные ошибки.

Материалами, распространение которых не запрещено тем или иным способом, являются:

- литературные произведения, находящиеся в общественном достоянии в связи с истечением срока охраны имущественных прав (в России исключительное право на произведение действует в течение всей жизни автора и семидесяти лет после его смерти) или переданные в общественное достояние авторами;

[1] Проект OpenCorpora // http://opencorpora.org

- «официальные документы государственных органов и органов местного самоуправления муниципальных образований, в том числе законы, другие нормативные акты, судебные решения, иные материалы законодательного, административного и судебного характера, официальные документы международных организаций, а также их официальные переводы»[2];

- тексты, опубликованные под свободными лицензиями[3].

На настоящий момент под лицензией «CreativeCommons-Attribution-ShareAlike» опубликованы все материалы проектов Фонда Викимедиа (Википедия[4], Викиновости[5], Викитека[6], ...), техническая документация на ряд программных продуктов, материалы интернет СМИ «Частный корреспондент»[7] и ряд других источников[8]. Таким образом, существует возможность составить корпус, свободному распространению которого не будут мешать лицензионные ограничения, и который будет включать тексты как минимум следующих жанров:

- художественная литература (до середины XX века);

- новости;

- юридические тексты;

- энциклопедические и словарные статьи;

- техническая документация.

Следует отметить, что существует ряд задач, для которых такой состав корпуса окажется неприемлемым. Например, задачи, предъявляющие высокие требования к репрезентативности исходных данных, где необходим корпус составленный из большого числа различных источников, или где нужна современная проза и тексты конкретных авторов. Проект OpenCorpora может быть использован в задачах, где имеет значение принадлежность текстов к вышеперечисленным жанрам и наличие аннотации, а разнообразие представленных источников не является необходимым условием.

Вычитка помещаемых в корпус текстов и создание разметки требуют больших временных затрат и определённой квалификации специалистов. Распределение этой работы среди штатных сотрудников возможно только в очень крупных организациях. Опыт Фонда Викимедиа[9] и ряда других проектов и инициатив показывает, что большие по объёму интеллектуальные продукты могут быть созданы сообществом добровольцев. При этом

[2]Гражданский кодекс Российской Федерации, глава 70, статья 1259 «Объекты авторских прав»

[3]Например, под лицензиями Creative Commons (`http://creativecommons.org/`), GNU FDL, ... (см. http://ru.wikipedia.org/wiki/Свободная лицензия)

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

[5]Викиновости // `http://ru.wikinews.org`

[6]Викитека — свободная библиотека // `http://ru.wikisource.org`

[7]`http://www.chaskor.ru`

[8]Кто в России использует лицензии Creative Commons? // `http://creativecommons.ru/who-uses-cc-in-russia`

[9]Сайт Фонда Викимедиа // http://wikimediafoundation.org/wiki/Приёмная

качество итогового результата сопоставимо, а иногда и превосходит[10] качество аналогичных продуктов, созданных в рамках соответствующих организаций. В проекте OpenCorpora сообществу добровольцев предлагаются следующие задачи:

- вычитка и добавление текстов в корпус (включает разделение текста на слова и предложения);
- снятие морфологической омонимии, оставшейся после автоматической разметки текста при помощи словаря.

Для организации работы сообщества создаются инструменты, предназначенные для упрощения создания аннотации (уменьшения количества необходимых действий) и для проверки её качества. Под качеством понимается соответствие стандартам и правилам, предписывающим определённую интерпретацию для каждого известного типа языковых явлений. Инструментами, упрощающими работу, являются интерфейс снятия морфологической неоднозначности и полуавтоматический токенизатор, выполняющий деление предложений на слова, знаки препинания и иные сочетания знаков, интерпретируемых как единое целое (токены). Инструментами контроля качества являются морфологический словарь (используется переработанная версия словаря проекта АОТ[11]), описывающий все возможные морфологические интерпретации словоформ, и система иерархических правил сочетания грамматических помет, применимая как к словарю, так и к морфологической аннотации текста[12]. Кроме этого, имеется возможность осуществлять выборочный контроль качества разметки вручную, устанавливая для каждого предложения статус, обозначающий, что его разметка полностью проверена данным специалистом и признана правильной.

Полуавтоматический токенизатор используется на этапе подготовки текста к добавлению в корпус. Он расставляет границы токенов, которые проверяются специалистом перед тем, как подтвердить добавление текста. Ошибки, допущенные токенизатором, должны быть исправлены вручную перед добавлением. Токенизатор основан на машинном обучении и использует для этого уже включённые в корпус тексты.

Интерфейс для снятия морфологической неоднозначности[13] показывает все имеющиеся морфологические разборы для каждой словоформы в выбранном предложении. Пользователь может отметить как единственно правильный вариант разбора (при этом все остальные варианты будут удалены), так и удалить несколько вариантов, в неправильности которых он уверен, оставив остальные для последующего анализа. Возможность такого частичного снятия омонимии предусмотрена для того, чтобы позволить специалистам, занимающимся этой работой, принимать решения только в тех пределах, в которых они сами считают себя компетентными.

[10] Munro R., Bethard S., Kuperman V., Lai V.T., Melnick R., Potts C., Schnoebelen T, Tily H. Crowdsourcing and language studies: the new generation of linguistic data // Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010

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

[12] Бочаров В.В., Бичинёва С.В., Грановский Д.В., Остапук Н.А., Степанова М.Е. Инструменты контроля качества данных в проекте Открытый Корпус, 2011.

[13] Интерфейс снятия морфологической неоднозначности: `http://opencorpora.org/sentence.php?id=8`. Кнопка «Сохранить» недоступна для пользователей, не зарегистрированных на сайте.

Упомянутые инструменты размещены на сайте проекта OpenCorpora и подразумевают удалённую работу всех участников проекта. На этом же сайте находится основное хранилище всех материалов: текстов, разметки и морфологического словаря. Изменения, вносимые участниками проекта, становятся доступны для просмотра сразу же. Статистика и иная агрегированная информация обновляются с определённой периодичностью, на настоящий момент — один раз в сутки. С такой же периодичностью формируются файлы для скачивания, содержащие материалы проекта в машиночитаемом виде.

Исходный код программного обеспечения, установленного на сайте OpenCorpora.org, хранится в Subversion репозитории Google Code[14] и публикуется под лицензией GNU GPL v2[15], разрешающей изучение, распространение и модификацию кода.

[14] Домашняя страница проекта OpenCorpora на Google Code // `http://code.google.com/p/opencorpora/`

[15] Текст лицензии GNU General Public License, version 2 на английском языке // `http://www.gnu.org/licenses/old-licenses/gpl-2.0.html`