Автоматизация лингвистической разметки корпуса текстов

Бабина О.И.
Южно-Уральский государственный университет
Дюмин Н.Ю.
Южно-Уральский государственный университет

В современной лингвистике корпусные исследования занимают значительное место, так как корпус позволяет выявлять и статистически обосновывать лингвистические явления, дает возможность прослеживать диахронические изменения в языке. Современные корпусы текстов, как правило, покрывают несколько функциональных стилей: включая разговорную речь, художественную литературу, публицистический стиль и т.д. В зависимости от типа помимо собственно текстовой информации, корпусы могут включать в себя лингвистическую разметку на морфологическом, синтаксическом, лексико-семантическом, дискурсном уровнях. Обзор некоторых концепций разметки русскоязычных корпусов можно найти в [1].

Однако собранные коллекции не всегда отвечают нуждам исследователя. Несмотря на обширный охват различных функциональных стилей, специализированные типы текстов (деловая переписка, тексты технических отчетов, тексты патентных документов и т.п.), как правило, не включены в корпусы. Вместе с тем, имеется необходимость в автоматической обработке также и таких специальных текстов. Более того, как показывает практика, автоматические системы обработки текстов, настроенные на работу в ограниченных предметных областях, отличаются наилучшими показателями производительности (классический пример системы, практически идеально работающей с ограниченным подъязыком, является система машинного англо-французского перевода сводок погоды TAUM-METEO).

Кроме того, в основе разметки каждого корпуса лежит определенная лингвистическая теория, и любые выводы по корпусным данным могут производиться лишь в рамках этой концепции. В то же время, каждый исследователь может придерживаться иной точки зрения на моделирование языка в силу как объективных, так и субъективных причин: например, в зависимости от структуры языков различные теории формализации языка подходя в большей степени к одним языкам и неприменимы к

другим; модели языка, в большинстве случаев, не лишены изъянов и не всегда способны описать все многообразие языковых явлений даже в пределах одного языка; и т.п.

Все это обусловливает необходимость наряду с использованием имеющихся обширных корпусов текстов также создавать исследовательские корпусы, включающие интересующий лингвиста подъязык и снабженный наиболее адекватной для исследовательских целей лингвистической разметкой. В связи с этим актуален вопрос оптимизации процесса создания корпусов (как глобальных так и локальных) посредством внедрения инструментария для автоматизации отдельных этапов аннотирования [2, 3, 4 и др.].

В зависимости от целей исследования, лингвистическое аннотирование небольших исследовательских корпусов текстов может включать как глубокую синтаксическую и семантическую разметку [напр., 5], так и ограничиваться лишь морфологическим компонентом [напр., 6]. Слишком подробная лингвистическая информация, заключенная в разметке, которой снабжаются большие корпусы текстов, может быть избыточна и требовать затрат неоправданно большого количества усилий, в то время как цели исследования допускают минимизировать трудозатраты, ограничившись лишь необходимым в данном исследовании набором меток. В рамках концепции минимизации усилий [7] целесообразно, с одной стороны, включать в разметку лишь необходимую исследователю информацию, с другой стороны, строить инструменты работы с корпусом, позволяющие повторно использовать методологию автоматическую разметку текста на других исследовательских корпусах.

Следуя данной концепции, мы выделили следующие принципы построения инструментария для лингвистической разметки корпуса:

1. Программный инструментарий должен поддерживать систему кодирования символов Unicode, тем самым обеспечивая возможность описания языков, использующих диакритику и/или символы алфавитов отличных от кириллического или латинского и, при этом, помогая избежать проблем связанных с совместимостью различных систем.

2. Текстовая коллекция и соответствующая ей лингвистическая информация должны храниться в единой базе данных, к которой обеспечивается стандартизованный доступ от программных компонентов системы, реализующих различный функционал по работе с корпусом текстов.

3. Конкретные инструменты, предназначенные для обработки корпуса, должны быть отделимы от него, обеспечивая общую

универсальность системы, в том числе возможность повторного использования лингвистической базы знаний другими системами.

4. Каждый компонент системы должен выполнять одну отдельную лингвистическую задачу, что обеспечивает модульную организацию системы.

5. Текстовая репрезентация языкового материала первична; вся производная лингвистическая информация (в частности, вхождения лексикона) привязана к позициям в текстовом корпусе. Принцип первичности текстовой репрезентации позволяет получать быстрый доступ к лексическому и грамматическому контексту для различных словоформ и словосочетаний из корпуса. Привязка лексических единиц к тексту также обеспечивает возможность извлечения набора допустимых морфологических меток слова или словосочетания для омонимичных форм.

Разрабатываемый нами комплекс автоматизации разметки корпуса текстов ориентирован на дальнейшую задачу извлечения лексико-грамматической информации о функционировании языковых единиц. В связи с этим разметка корпуса включает морфологический и синтаксический уровень.

Программный инструментарий разрабатываемого комплекса автоматизации текстовой разметки включает в себя: 1) модуль управления корпусом (CorpusManager), 2) модуль автоматической морфологической разметки (AutoPOSTagger), 3) модуль автоматизированной коррекции морфологической разметки (Corrector), 4) модуль автоматизации синтаксической разметки (SynTagger). Следует отметить, что система представляет возможности многосторонней обработки текста, при этом отдельные функции представлены различными инструментами (некоторые функции в отдельных приложениях системы дублируются – это сделано для удобства исследователя и не нарушает общего принципа модульности и универсальности).

Концептуальная схема информационных потоков и взаимодействия модулей системы для автоматизации разметки корпуса текстов, разработанного в соответствии с сформулированными принципами в лаборатории инженерной лингвистики ЮУрГУ, представлена на рисунке 1.

Модуль управления корпусом (CorpusManager) предназначен для загрузки текстов в корпус. Перед загрузкой текст проходит предобработку – разбивается на отрывки согласно пунктуации. В следующей фазе предобработки пунктуация и специальные символы удаляются, а отрывки разбиваются на n-грамы размерностью 1,2,3 и 4.

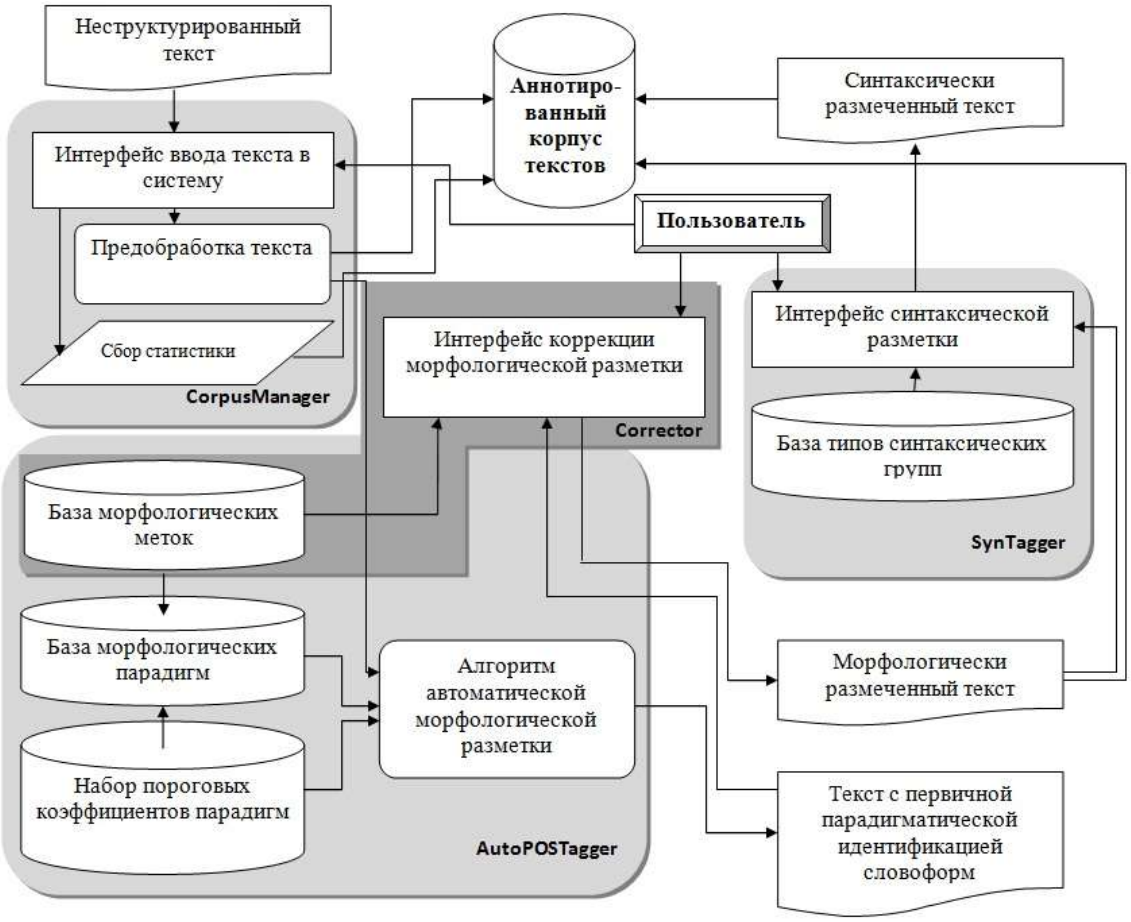

Рис. 1. *Схема информационных потоков и взаимодействия модулей системы*

Модуль автоматической морфологической разметки (AutoPOSTagger) и коррекции автоматической разметки (морфологические тэггеры) предназначены для присваивания меток 1-грамам. Морфологическая разметка является двухступенчатой – в каждом слове определенной меткой помечено его 1) частеречное и 2) грамматическое значение. Поскольку набор грамматических категорий и парадигм уникален для каждого языка, с целью обеспечения возможности повторного применения инструментария к корпусам на различных языках необходимым элементом модуля разметки является отделимая от программной реализации его интерфейса база морфологических меток, соответствующая рассматриваемому в пределах данного языка набору грамматических категорий. Так, например, база меток, репрезентирующих грамматические категории для некоторых открытых классов частей речи и используемых для аннотации корпуса патентных текстов на русском языке, представлена в таблице 1.

Таблица 1

Частеречное значение	*Грамматическое значение*		
N – сущ.	Nom – Им.п. Gen – Род.п. Dat – Дат.п. Acc –Вин.п. Inst – Тв.п. Prep – Пр.п.	sg – ед. ч. pl – мн.ч.	masc – м.р. fem – ж.р. neu – ср.ср.
ADJ – прил.			
V – глаг.	1p – 1 л. 2p – 2 л. 3p – 3 л.		
PTG – прич.			

Аннотирование корпуса текстов с применением хранящегося в базе набора меток может проходить в полностью автоматическом или автоматизированном режиме. Автоматический режим включает разметку с помощью алгоритма, основанного на использовании построенной вручную базы парадигм. Каждая парадигма представлена набором квазисуффиксов (рис.2, вкладка «парадигмы»). Работа алгоритма заключается в построении дерева гипотез относительно деления словоформы на основу и флексию (рис. 2, второй столбец). Далее для каждой словоформы в процессе применения эмпирически сформированного набора пороговых коэффициентов частотности парадигм из списка гипотез (рис. 2, третий столбец) отбирается наиболее вероятная гипотеза морфологического членения этой словоформы, и определяется принадлежность гнезд словоформ с одинаковой квазиосновой к одной из парадигм [18]. Данный метод оказывается особенно эффективным на большом корпусе текстов.

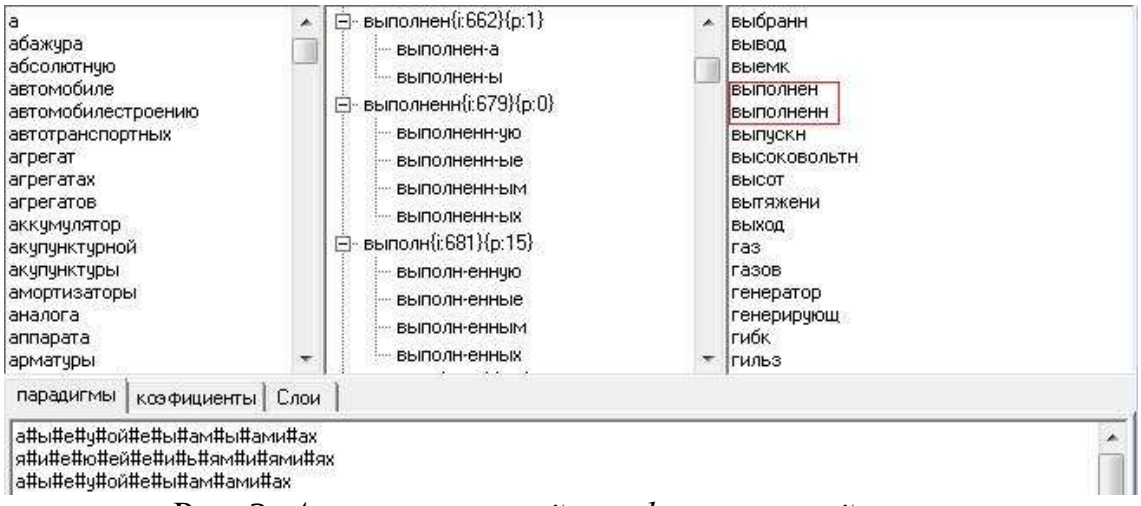

Рис. 2. *Автоматический морфологический анализ*

Автоматизированный режим включает автоматическую разметку с последующим пост-редактированием посредством *модуля коррекции морфологической разметки* (Corrector). В автоматизированном режиме пользователю предлагаются варианты меток исходя из ранее размеченных элементов корпуса.

Результатом применения модулей AutoPOSTagger и Corrector является морфологически размеченный корпус текстов, пример которого представлен на рис. 3.

Радиодальномер_N#sg#masc#Nom ***относится***_V#sg#3p
к_PREP ***радиотехнике***_N#sg#fem#Dat ***и***_CONJ
предназначен_PTG#sg#masc ***для***_PREP
прецизионного_ADJ #sg#neu#Gen
определения_N#sg#neu#Gen ***расстояния***_N#sg#neu#Gen...

Рис. 3. *Пример морфологической разметки корпуса текстов*

Все 1-грамы размечены таким образом и составляют словарную базу корпуса. Простые элементы n-грам содержат ссылки на соответствующие 1-грамы, и, таким образом, разметка словосочетаний словаря корпуса осуществляется опосредованно (как совокупность меток словоформ, составляющих n-грам).

Модуль автоматизации синтаксической разметки (SynTagger) в описываемом нами комплексе (рис. 4) включает объединение синтаксически зависимых друг от друга лексических единиц текста посредством скобочной записи.

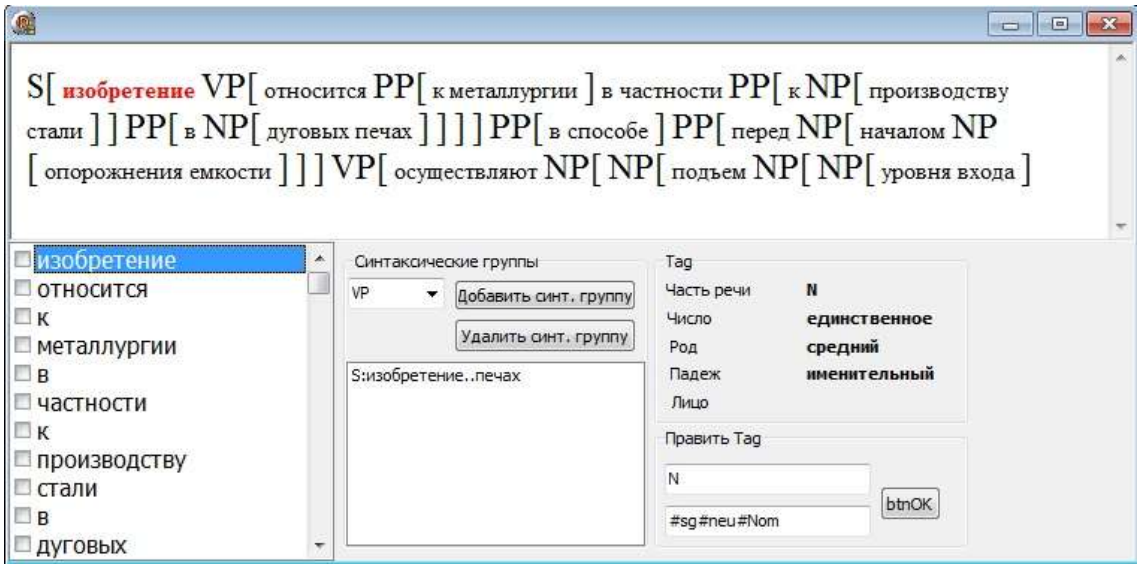

Рис. 4. *Синтаксическая разметка корпуса текстов*

Пользователю предлагается определить начало и конец синтаксического блока и выбрать его тип (именная, предложная, глагольная, адъективная, адвербиальная группа, группа числительного или предложение). Модуль SynTagger позволяет, при условии наличия морфологической разметки, автоматически выявлять структуру синтаксических групп разного типа, что является полезным при исследовании синтаксических особенностей различных функциональных стилей или подъязыков.

Описываемая система обладает распределенной архитектурой, что делает возможной одновременную работу с корпусом для нескольких исследователей, в том числе и по сети. Кроме того, реализация базы знаний в виде сетевой базы данных позволяет значительно ускорить процесс обработки запросов за счет вычислительных ресурсов сервера, на котором расположена база.

Литература

1. 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.: Изд-во «Индрик», 2005. – С. 31–61.
2. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации / И.М. Богуславский, Н.В. Григорьев, С.А. Григорьева, Л.Л. Иомдин, Л.Г. Крейдлин, В.З. Санников, Н.Е. Фрид // Труды международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2000» / под ред. А.С. Нариньяни. – Протвино, 2000. – Т. 2. – С. 41-47.
3. Семеренко, В.Р. Автоматизация морфологической разметки текстов Национального корпуса украинского языка / В.Р. Семеренко // Искусственный интеллект. – 2005. – №4. – С. 640-645.
4. Баранов, А.Г. Моделирование применения корпусных методов для локальных лингвистических исследований // Материалы международной конференции «Диалог-2010». Режим доступа: http://www.dialog-21.ru/dialog2010/materials/pdf/Baranov.pdf (29.09.2010).
5. Бабина, О.И. Аннотирование корпусов текстов / О.И. Бабина // Наука ЮУрГУ: Материалы 60-й юбилейной научной конференции. Секции естественно-научных и гуманитарных наук. – Челябинск: Изд-во ЮУрГУ, 2008. – Т. 2. – С. 3-6.
6. Baroni, Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the *La Reppublica* Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In Proceedings of LREC 2004. Lisbon: ELDA. Pp. 1771–1774.
7. Шереметьева, С.О. Методология минимизации усилий в инженерной лингвистике: дис. … д-ра филол. наук / С.О. Шереметьева. – СПб., 1997. – 288 с.
8. Бабина, О.И. Автоматический морфологический анализ флективных языков / О.И. Бабина, Н.Ю. Дюмин // Наука ЮУрГУ: материалы 62-й научной конференции. Секции естественно-научных и гуманитарных наук. – Челябинск: Издательский центр ЮУрГУ, 2010. – Т. 2. (*в печати*)