Интеллектуальная система контроля качества научно-технического текста

М.Г. Мальковский, Е.И. Большакова

Написание связного научного текста (статьи, диссертации, автореферата, отчета и проч.) — сложный многоэтапный процесс. Помимо автора в него могут быть вовлечены и другие люди: рецензенты, литературный, научный и технический редакторы, корректоры, просто заинтересованные читатели.

Решение задачи моделирования такой творческой интеллектуальной деятельности человека, как создание текстов на естественном языке, очевидно, дело не ближайшего будущего. В то же время в издательствах давно и успешно используются универсальные (не привязанные к какой-либо конкретной предметной области) коммерческие системы подготовки текстов: текстовые редакторы, издательские системы, автокорректоры; причем спектр их возможностей постоянно расширяется [1]. Сделаны и первые попытки моделирования достаточно сложных и содержательных функций литературного и научного редакторов — в рамках создания специализированных систем автоматизированной обработки естественноязычных текстов, относящихся к достаточно узким предметным областям.

К таким системам можно отнести ЛИНАР (ЛИтературно-НАучный Редактор) — систему комплексного контроля и редактирования русскоязычных текстов [2], разработанную коллективом сотрудников кафедры алгоритмических языков факультета вычислительной математики и кибернетики МГУ. Модельная версия ЛИНАР была реализована на языке Плэнер [3] для ЭВМ БЭСМ-6 и МВК "Эльбрус", а затем перенесена на ПЭВМ (Planner-PC, MS-DOS). Эта версия была ориентирована на обработку русскоязычных научно-технических текстов, например, пояснительных записок, технических заданий, отчетов о НИР, тематических

сборников в области "Архитектура и программное обеспечение многопроцессорных вычислительных комплексов обработки и интерпретации результатов наблюдений".

Пользователем системы ЛИНАР является человек, желающий проверить и оценить качество некоторого текста с позиций потенциального читателя, литературного или научного редактора и при необходимости внести в этот текст исправления. Пользователем может быть не только сам автор анализируемого текста, но и, например, редактор сборника научных статей, обеспокоенный терминологическими и стилистическими неувязками в текстах, подготовленных разными авторами.

Обычно обработка с помощью системы ЛИНАР (уже введенного в ЭВМ) текста осуществляется за несколько сеансов работы с системой. Сеанс работы состоит из серии одноаспектных проверок текста в целом или его фрагментов (каждая одноаспектная проверка — суть проверка одного определенного свойства/характеристики текста, например, анализ использования аббревиатур). В начале каждого сеанса пользователь формирует задание на обработку текста, для выполнения которого система загружает необходимые программные и информационные модули. В результате проверок система фиксирует ошибки и выдает замечания по тексту, а пользователь просматривает замечания и при желании вносит необходимые изменения в текст с помощью текстового редактора. Измененная версия текста может быть объектом обработки в следующем сеансе.

В целом система ЛИНАР берет на себя выполнение ряда нетривиальных функций по контролю качества текста. Ее возможности базируются на совокупности процедурных и декларативных знаний, обеспечивающих возможность оценить сложность восприятия текста, его соответствие общеязыковым нормам и общепринятым правилам оформления и изложения материала в научно-технических документах, а также семантическую корректность текста (его соответствие понятийной модели предметной области) [4].

Процедурная часть знаний системы представлена программами одноаспектного контроля (их несколько десятков). Каждая программа проверяет одно заданное свойство/характеристику текста или фиксирует определенный дефект текста. В зависимости от анализируемого аспекта текста совокупность программ контроля можно разбить на несколько групп:

- Программы орфографического контроля могут обнаружить грамматические ошибки в склонении и спряжении слов, а также "клавиатурные"ошибки (например, пропуск буквы или пробела).

- Программы анализа лексического состава текста подсчитывают частоту употребления заданных слов и словосочетаний, выявляют слова с нежелательной стилистической окраской (канцеляризмы, жаргонизмы и др.), проверяют правильность использования аббревиатур.

- Программы стилистической проверки выявляют лексические повторы и другие виды тавтологий, фразы чрезмерной сложности (по длине фразы, количеству знаков препинания и простых предложений в составе сложного).

- Программы анализа композиционной структуры текста проверяют выполнение правил структуризации текста (наличие определенных разделов, порядок их следования, правильность нумерации и др.).

- Программы синтаксического контроля могут выявить нарушения обычного порядка слов, ошибки согласования и управления.

- Программы семантического анализа кроме неоднозначности смысловой интерпретации фразы или ее фрагмента могут обнаружить некоторые виды семантических противоречий.

ЛИНАР не только обнаруживает неточности, ошибки, но и может объяснить пользователю суть своих замечаний, а также предложить способы устранения ошибок. Так, например, в случае орфографической ошибки система предлагает свой вариант исправления слова, в случае нарушения нейтрального порядка слов — рекомендуемые перестановки и т.д.

Отметим, что программы одноаспектного контроля различаются по характеру получаемого ими результата. Большинство программ лексического уровня — в отличие от орфографического, синтаксического и семантического — не выявляют нарушения и ошибки в тексте, а собирают глобальную вспомогательную (статистическую) информацию об используемой в исследуемом фрагменте текста лексике.

Программы орфографического, синтаксического и семантического контроля используют мощные средства анализа текстов лингвистического

процессора АДАМАНТ [5]. Основу декларативных знаний ЛИНАР составляет лингвистическая база знаний процессора АДАМАНТ, включающая:

- словари неизменяемых слов, основ слов и аффиксов;
- совокупность грамматических правил и шаблонов анализа;
- словарь слов-предикатов (с описанием моделей управления);
- базовый семантический словарь (тезаурус);
- словари синонимов, гипонимов, гиперонимов и омонимов;
- словари-справочники типа: "Трудности русского языка", "Слитно/раздельно"и др.

Кроме того, ЛИНАР использует тематический словарь и тезаурус той предметной области, к которой относятся обрабатываемые тексты, а также описания нормативных требований, предъявляемых к текстам. Соответствующие информационные массивы создаются на основе общеязыковых и предметно-ориентированных словарей и справочников, Государственных стандартов и отраслевых инструкций по оформлению текстовых документов.

Эксперименты с системой ЛИНАР показали сильные и слабые места системы, пути дальнейшего развития систем такого типа, усиления их мощности. Одно из самых важных и наименее изученных направлений развития ЛИНАР — создание программ глобального содержательного анализа текста. Большинство программ одноаспектного контроля, проводящих глубокий синтаксический и семантический анализ, имеют локальную область действия — они обрабатывают мелкие фрагменты текста: отдельные словосочетания, предложения, реже — несколько соседних предложений в рамках одного абзаца. А программы, имеющие обширную область действия (несколько абзацев, пункт, раздел, весь текст целиком), например, программы анализа композиционной структуры текста, осуществляют лишь формальный, поверхностный контроль, без учета смысловых связей отдельных фрагментов текста. Для разработки программ глобального содержательного анализа текста в идеале необходима формальная модель связного текста (в лингвистике лишь

намечаются пути ее разработки), включающая описание различных видов смысловых связей между крупными смысловыми единицами текста: абзацами, пунктами и подпунктами, разделами и т.п. Такая модель поможет более точно ответить на вопрос "Что такое хороший, качественный текст на естественном языке?".

Программы одноаспектных проверок системы ЛИНАР применимы для обработки научно-технических текстов любых жанров (статья, реферат и т.п.). Однако в системе отсутствуют средства проверки параметров, специфических для текстов некоторых жанров, например, для дипломных работ. Очевидно, что включение в систему новых специфических проверок сделало бы ее чрезмерно громоздкой, с плохо обозримым для пользователя спектром возможностей.

Более целесообразно создание на основе ядра системы ЛИНАР мобильных систем контроля качества текста, ориентированных на один или несколько жанров и обладающих возможностями как общих, универсальных проверок текста (наследуемыми от ЛИНАР), так и специфических для рассматриваемых жанров. Именно так создавалась система КОНУТ (КОНтроль Учебных Текстов), предназначенная для проверки и редактирования учебно-научных русскоязычных текстов: рефератов, дипломных и курсовых работ. Полезность такой системы для пользователя-студента (а, возможно, и преподавателя) определяется, во-первых, скромным уровнем грамотности среднего студента, а во-вторых, неумением большинства студентов правильно оформлять свои работы (этому специально нигде не учат, и умение это приходит с опытом).

Первая версия системы КОНУТ была разработана в 1996 г. (язык C++, Windows NT) студентами кафедры. Первоначально ставилась задача реализации в системе как можно большего числа проверок правил оформления учебных текстов без привлечения обширных машинных словарей русского языка (словарей основ слов, слов-предикатов и т.п.).

В системе КОНУТ использован тот же технологический принцип, что и в ЛИНАР — средства контроля текста реализованы как набор программ одноаспектных проверок текста, что облегчает как использование системы, так и ее расширение. В системе есть свой встроенный редактор, в котором можно вводить и исправлять текст. В то же время система позволяет работать с текстами, созданными в других редакторах.

Основная функция системы КОНУТ — проверка правил оформления учебно-научных работ: контроль заголовков и их нумерации, проверка

титульного листа, библиографии (как списка литературы, так и правильности библиографических ссылок в тексте) и др.

Кроме этого система может провести анализ и оценку стиля и композиции текста — проверяется соблюдение принципа соразмерности (по объему) всех структурных единиц текста (от предложения и абзаца до разделов), оценивается сложность восприятия этих единиц потенциальным читателем. В системе имеется также информационный модуль-справочник, в котором собраны воедино все подробные сведения о формальных и неформальных требованиях к учебно-научным текстам. В модуле используется гипертекстовая технология: вся справочная информация представлена как гипертекст, и студенту-пользователю предоставляется возможность как свободной навигации по нему, так и прохода основных узлов гипертекста в строго определенной (обучающей) последовательности. Кроме такого независимого просмотра справочника система дает отсылки к соответствующим разделам справочника при обнаружении ошибок в тексте в ходе его контроля.

Хотя в нынешней версии системы КОНУТ нет некоторых важных проверок (орфографии, синтаксиса и семантики), в ней реализован более широкий по сравнению с ЛИНАР спектр формальных проверок, необходимых именно для учебно-научных текстов. В дальнейших версиях системы предполагается усилить ее мощность именно за счет подключения синтактико-семантического анализа. Эта работа осуществляется по мере переноса на язык C++ основных компонентов лингвистического процессора АДАМАНТ.

Литература

1. Ашманов И. Грамматический и стилистический корректор для текстов на русском языке // Тр. Межд. семинара Диалог'95: компьютерная лингвистика и ее приложения — Казань, 1995. — С. 39–42.

2. Мальковский М.Г., Большакова Е.И., Волкова И.А. и др. Эксперименты с системой ЛИНАР // Тр. Машинного фонда русского языка, т. 1 — М., 1991. — С. 51–71.

3. Пильщиков В.Н. Язык плэнер — М., 1983. — 208 с.

4. Малютин П.Г., Мальковский М.Г. Работа с противоречиями в модели предметной области "Нормативные документы в строительстве"// Тр. Межд. семинара Диалог'96 по компьютерной лингвистике и ее приложениям — М., 1996. — С. 151–152.

5. Мальковский М.Г. Программно-информационное обеспечение адап-

тивных систем общения с ЭВМ на естественном языке — Дисс. ... докт. физ-мат. наук, — МГУ, 1990.