ИЗВЕСТИЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ
ГЕОЛОГИЯ И РАЗВЕДКА
2003, № 5

ДИСКУССИИ

УДК 550.9

В.Н. ДУДЕЦКИЙ

ОПЫТ ПРИМЕНЕНИЯ СИСТЕМЫ ПОНИМАНИЯ ТЕКСТА К АНАЛИЗУ ГЕОЛОГИЧЕСКИХ ТЕКСТОВ

Рассмотрены теоретические и практические вопросы построения терминологических систем в геологии. Приведены примеры анализа текстов по разным направлениям геологической науки.

Разработанную систему понимания текста на естественном языке [1, 3, 4] можно применять при построении терминосистем по отдельным областям научного знания. *Терминосистема* представляет собой упорядоченную терминологию. В отличие от терминосистемы терминология, как правило, не отличается полнотой и имеет нечеткое соотношение между системами понятий и терминов [5, 6]. Особенно актуальна задача создания геологических терминосистем, так как геологический язык обладает многозначностью не только лексического, но синтаксического и стилевого планов.

При построении терминосистем в геологии нужно выполнять ряд действий согласно схеме (рис. 1). Автоматизация данной схемы позволяет перейти на новый уровень построения и понимания терминосистем как к некоторой научной проблеме, т. е. перейти к *терминологике* — научному направлению исследования терминосистем. *Цель геологической терминологики* — построение терминосистем по различным направлениям геологии (рис. 1).

Предмет геологической терминологики (рис. 1) составляют исходные неопределяемые понятия и сложные производные понятия геологии, логические схемы, связывающие эти понятия, и алгоритмы обработки данных для решения различных геологических задач. Все эти составляющие предмета геологической терминологики реализуются посредством естественного геологического языка. В составе его лексики можно выделить несколько групп слов. Специфику геологического языка составляют прежде всего геологические термины и геологическая номенклатура.

К *средствам геологической терминологики* относится система понимания текста. В ее состав входят лингвистический процессор, процессор логических выводов, прагматический процессор и база знаний [2]. Лингвистический процессор системы проводит морфологический, синтаксический, семантический, прагматический и логический анализы геологических текстов. Он определяет семантическую емкость текста и пополняет базу знаний. Процессор логических выводов производит анализ геологических текстов и базы знаний на непротиворечивость. Прагматический процессор анализирует базы геологических знаний на полноту.

Непротиворечивая и полная база геологических знаний — основа создания терминосистем для различных направлений геологической науки.

Система понимания текста на естественном языке разработана в среде операционной системы Windows-2000. Требования к конфигурации вычислительных средств: процессор P-IV, оперативная память 256 Мб, внешняя память прямого доступа 40 Гб.

При автоматизации процесса построения терминосистем в геологии необходимо, чтобы в геологической терминологике были реализованы следующие процедуры: 1) выявление исходных неопределяемых понятий; 2) выявление сложных производных понятий; 3) проведение проверки системы понятий на полноту; 4) проведение проверки определений понятий на непротиворечивость; 5) построение полной терминологической системы; 6) разработка формального представления информации.

Для выполнения пунктов 1, 2 данного алгоритма пользователь должен подготовить семантически емкие тексты из заданного информационного поля. Текст — семантически емкий, если сумма количеств предложений семантического типа «именование», «коррекция», «уточнение» и «свидетельство», а также предложений, содержащих логические выводы, составляет более половины вводимых предложений [4].

Рис. 1. Структура геологической терминологики

После ввода исходной информации, пункты 1, 2 выполняются системой автоматически, пункты 3, 4 — автоматически, а если необходимо с помощью пользователя в интерактивном режиме. Пункты 5, 6 алгоритма автором не рассматривались, однако нет данных, подтверждающих то, что структура системы противоречит введению этих пунктов в ее состав.

Для проверки готовности системы понимания текста к реализации поставленной задачи сформулирован и построен контрольный пример, состоящий в следующем. Были выбраны четыре текста, представляющие разные направления геологической науки:

1. Основы региональной геологии СССР: Учебник для вузов. / В.М. Цейслер, В.Б. Караулов, Е.А. Успенская, Е.С. Чернова. М.: Недра, 1984. С. 29—87.

2. Перельман А.И. Геохимия: Учеб. для геол. спец. вузов. 2-е изд., перераб. и доп. М.: Высш. шк., 1989. С. 44—53.

3. Вахромеев Г.С. Основы методологии комплексирования геофизических исследований при поисках рудных месторождений. М.: Недра ,1978. С. 42—93.

4. Ванярхо М.А. Технология импорта геолого-геофизических данных в интерпретационные программные комплексы. Автореф. дис. ... канд. техн. наук М., 2002. 18 с.

Глава III «Основ региональной геологии» (Восточно-Европейская древняя платформа) содержит 135 915 байт текста и состоит из 235 фрагментов (абзацев) и 1062 предложений (среднее количество предложений в фрагменте 4,5). Средняя длина фрагмента 578 байт, средняя длина предложения 128 байт.

Глава 3 «Геохимии» (Геохимия планет земной группы и космохимия) включает 16 817 байт текста и состоит из 32 фрагментов и 172 предложений (среднее количество предложений в фрагменте 5,4). Средняя длина фрагмента 525 байт, средняя длина предложения 98 байт.

Глава III «Основ методологии» (Выбор рационального комплекса геофизических методов и обоснование элементов его методики) содержит 65 607 байт текста, 128 фрагментов и 364 предложений (среднее количество предложений в фрагменте 2,8). Средняя длина фрагмента 513 байт, средняя длина предложения 180 байт.

Автореферат диссертации — 21 801 байт текста, 44 фрагмента и 112 предложений (среднее количество предложений в фрагменте 2,6). Средняя длина фрагмента 495 байт, средняя длина предложения 195 байт.

Результаты синтаксического, семантического и логического анализов текстов приведены в таблице.

Существенные различия между текстами выявляются на этапе синтаксического анализа: доля глагольных предложений в учебнике по региональной геологии 4%, в геофизической монографии 21%, в учебнике по геохимии 8% и в автореферате по геоинформатике 42%. Известно, что чем

Результаты синтаксического, семантического и логического анализов

Номер текста	Синтаксический анализ предложений	Семантический и логический анализы предложений
1	Простых 613 (58%), в том числе глагольных 42 (4%); сложных 449 (42%) Σ 1062	Семантически емких 449 (42%), в том числе: именований 43 (4%), уточнений 85 (8%), свидетельств 163 (15%), именований со ссылкой на источник 85 (8%), логических высказываний 73 (7%)
2	Простых 96 (56%), в том числе глагольных 7(4%); сложных 76 (44%), в том числе с придаточными глагольными 7(4%) Σ 172	Семантически емких 110 (64%), в том числе: уточнений 89 (52%), простых логических высказываний 21 (12%)
3	Простых 189 (52%), в том числе глагольных 76 (21%); сложных — 175 (48%) Σ 364	Семантически емких 245 (67%), в том числе: именований 8 (2%), свидетельств 127 (35%),простых логических высказываний 43 (12%), сложных логических высказываний 67 (18%)
4	Простых 52 (46%), в том числе глагольных 30(27%); сложных 60 (54%), в том числе с придаточными глагольными 17 (15%) Σ 112	Семантически емких 91 (81%), в том числе: именований 4 (4%), уточнений 79 (70%), простых логических высказываний 8 (7%)

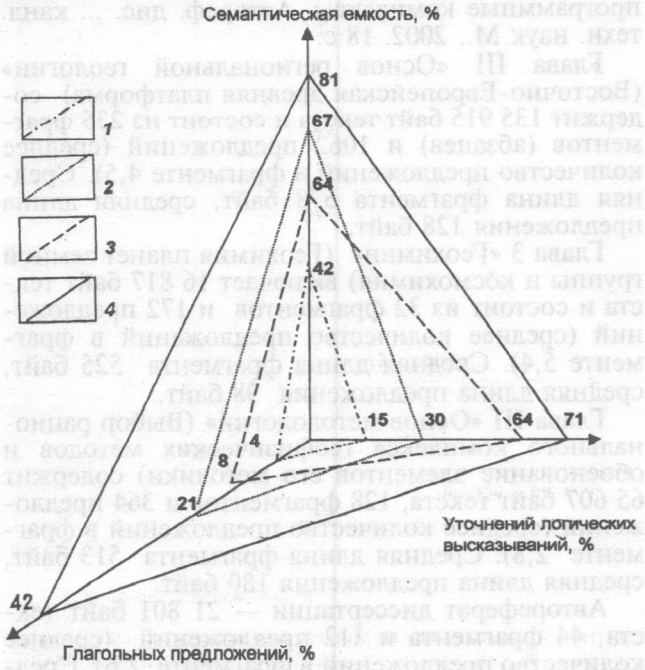

Рис. 2. Результаты анализа геологических текстов: *1* — региональная геология; *2* — геохимия; *3* — геофизика; *4* — геоинформатика

меньше глагольных предложений, тем сильнее повествовательный характер текста. И наоборот, в текстах, посвященных логическим построениям, доля глагольных предложений велика.

Рассмотрим результаты анализа четырех текстов контрольного примера, представляющих четыре различных направления геологической науки: 1) региональной геологии, 2) геохимии, 3) геофизики, 4) геоинформатики (рис. 2). В первом тексте доля логических высказываний составляет всего лишь 7%. Третий текст контрольного примера наиболее насыщен логическими построениями и выводами: доля простых логических высказываний составляет 12%, доля сложных — 18%.

Высокая семантическая емкость учебника по геохимии определяется большим объемом фактографического материала, а текста по информатике достигнута за счет детального описания алгоритмов импорта геолого-геофизических данных.

Заключение

Анализ геологических текстов, проведенный с помощью системы понимания текста на естественном языке, показал способность системы к построению терминологических систем в такой трудноформализуемой области естественного языка как геологический. В то же время проведенный анализ указывает на направления развития средств терминологики, основным из которых является разработка интерфейса между системой понимания текста и системами распознавания речи.

Основная проблема включения разработанных систем распознавания речи в средства терминологики в том, что система понимания текста настроена на морфологически, синтаксически и семантически правильные тексты. Учитывая высокий процент ошибок распознавателей речи, потребуется существенное расширение базы правил, а также, возможно, функций лингвистического процессора, а также увеличения функций процессора логических выводов и прагматического процессора.

ЛИТЕРАТУРА

1. Дудецкий В.Н. Компьютерная модель обыденной интуиции // Тез. докл. III международной конференции «Новые идеи в науках о Земле». Т. 4. М.: Полимаг, 1997. С. 246.
2. Дудецкий В.Н. Система автоматизации проектирования трудноформализуемых задач в области технологии разведки // Геоинформатика. 1999. № 4. С. 14—22.
3. Дудецкий В.Н. Организация данных в системе понимания текста на естественном языке // Геоинформатика. 2000. № 2. С. 14—21.
4. Дудецкий В.Н. Система понимания текста на естественном языке. М., 2001. 107 с.
5. Смирнова А.С. Построение автоматизированных фактографических информационно-поисковых систем в геологии. М.: Недра, 1976. 346 с.
6. Смирнова А.С. Информационный анализ в геологии. М.: Недра, 1985. 274 с.

Московский государственный
геологоразведочный университет
Рецензент — А.В. Веселовский