Автоматический анализ дискурсивной структуры научного текста

Е.И. Большакова, Н.В. Баева
МГУ им. М.В.Ломоносова, Факультет ВМиК
bolsh@cs.msu.su; baeva@vog.ru

Рассматривается задача автоматического распознавания дискурсивной структуры научно-технического текста. В этой связи обсуждаются стилевые особенности дискурсивно-логической организации научно-технических произведений, отмечаемые в ряде лингвистических работ и обнаруженные авторами при исследовании научных текстов из разных предметных областей. Научный дискурс строится как взаимосвязанная последовательность дискурсивных приемов описания и аргументирования, реализуемых в соответствующих сегментах текста и помечаемых обычно дискурсивными маркерами. С учетом рассмотренных особенностей намечается процедура распознавания в заданном тексте дискурсивных приемов и соответствующих текстовых сегментов. Процедура базируется на поверхностном синтаксическом анализе фраз текста и словаре общенаучной лексики, который охватывает слова и выражения общенаучной речи, используемые обычно как дискурсивные маркеры. Кратко характеризуется состав словаря, а также высказываются предложения по использованию результатов распознавания для прикладных целей – литературно-научного редактирования, аннотирования и реферирования.

Введение

Для решения многих прикладных задач автоматизированной обработки научно-технических текстов – литературно-научного редактирования, извлечения текстовых знаний, реферирования и аннотирования – необходимы не только алгоритмы терминологического анализа, но и процедуры выявления и учета композиционно-логической, дискурсивной организации текстов. Действительно, выделенное в результате терминологического анализа множество ключевых слов текста позволяет обозначить тематику текста, но никак не основное его содержание. Однако, если проблема автоматического выделения терминов в текстах широко исследуется уже более 30 лет, то вопросам автоматического выявления дискурсивной структуры текстов посвящено необоснованно мало работ.
В статье рассматривается задача автоматического распознавания композиционно-речевой структуры научного текста, для решения которой были изучены и учтены стилевые особенности дискурсивной организации научных произведений. Эти особенности изучались нами текстах разных жанров (в статьях, монографиях, аннотациях), взятых из разных предметных областей (в том числе – из области информатики). Преимущественно рассматривались научные статьи (как относящиеся к «ядру» функционального стиля) из области естественных и технических наук.

Согласно исследованиям, посвященным научному дискурсу и указывающим основные его особенности [4, 7, 8, 9], научный дискурс представляет собой рассуждение, которое организуется как взаимосвязанная последовательность дискурсивных приемов описания, классификации, сравнения, оценивания и др. Каждому приему соответствует свой сегмент текста, состоящий из одного или нескольких предложений и содержащий обычно дискурсивные маркеры – слова и словосочетания, эксплицитно помечающие примененные операции.

Дискурсивные маркеры (*по этой причине, кроме того, в действительности, однако* и т.п.) называются также словами-организаторами (реже – словами-скрепами [3] или опорными словами [11]), поскольку их основной функцией является структурно-смысловая организация научного текста – оформление и упорядочение рассуждений, связывание отдельных текстовых фрагментов. Дискурсивные слова и выражения встречаются в текстах разных стилей и текстах на разных естественных языках, они относятся к метатекстовому компоненту дискурса. Замечено также, что стили текстов отличаются количественным и качественным «вкладом» метатекста в текст [2]. Можно с уверенностью сказать, что для научного стиля речи этот вклад наибольший.

Изученные особенности научного дискурса легли в основу рабочей гипотезы, согласно которой задача автоматического распознавания дискурсивной структуры научного текста может быть решена на основе поверхностного синтаксического анализа текста и лексикона дискурсивных слов. В работе [11] аналогичная идея выдвигалась и обсуждалась применительно к текстам любого стиля, но не была воплощена в работающей системе. Идея гласила, что при беглом, сквозном прочтении текста его общая логико-композиционная структура может быть понята по опорным словам, поскольку «семантика текстоорганизующих конструкций отделяется от содержательной информации текста». Алгоритмы дискурсивного анализа были реализованы для текстов на японском языке [14, 15], но в них использовался достаточно глубокий синтаксический анализ и не учитывались все особенности научного дискурса (хотя работоспособность алгоритмов была проверена именно на научно-технических текстах).

В настоящей работе в общих чертах намечается процедура автоматического распознавания в заданном тексте дискурсивных приемов и соответствующих текстовых сегментов, а также их структурно-смысловых связей. Результатом распознавания является дискурсивно-композиционная схема текста. Особенность нашего подхода заключается во всестороннем учете специфики научного дискурса и в опоре на компьютерный словарь русской общенаучной лексики [1], отражающем эту специфику. В словаре представлены общенаучные слова и выражения (*предположим, что; в заключение* и т.п.), используемые в научных текстах из различных естественных и технических областей.

Ниже обсуждаются особенности научного дискурса, кратко характеризуется состав и структура словаря общенаучной лексики, и в общих чертах описываются основные шаги процедуры дискурсивного анализа. Высказываются также предложения по использованию распознанной дискурсивно-композиционной схемы текста для получения его реферата или аннотации.

Особенности научного дискурса

Дискурс представляет собой взаимосвязанную и взаимообусловленную последовательность отдельных речевых актов (дискурсивных приемов), детерминируемую коммуникативной целеустановкой [5, 8]. Глобальная коммуникативная цель научного произведения – сообщение о результатах проведенного исследования и объяснение способа их получения, формулировка новых идей и их обоснование.

Как следствие, типичное научное изложение состоит главным образом из рассуждений, и отличительной особенностью функционального стиля научной прозы является не ее насыщенность специальными терминами (это характерно и для научной фантастики), но особый формально-логический способ изложения материала.

Научный дискурс организуется как логическая последовательность шагов рассуждения – речемыслительных действий, соответствующих интеллектуальным операциям над мыслями об объекте исследования [8, 9, 10]. К типичным действиям и операциям относятся обоснование вывода, выдвижение гипотезы, введение термина и понятия, приведение фактов и доказательств, подведение итогов и т.п. В норме эти операции вводятся автором научного текста и эксплицитно помечаются при помощи разнообразных дискурсивных слов и выражений (слов-организаторов научной мысли).

Наиболее явными маркерами мыслительных операций служат ментальные перфомативные высказывания (такие как *особо подчеркнем, далее мы докажем*), которые включают широкий круг ментальных перфомативных глаголов (*выразим, учтем, рассмотрим* и т.п.) [9]. Перформативные высказывания не только помечают, но и квалифицируют соответствующий шаг рассуждения и выстраивают содержание текста в форме рассуждения.

В работе [9] описаны эквивалентные виды ментальных перформативных высказываний:

- канонические, с глаголом в 1 лице множественного числа (*мы покажем*);
- «установочные», с модальным или оценочным словом (*необходимо/нетрудно заметить*);
- в форме деепричастия или деепричастного оборота (*резюмируя вышесказанное*);
- в безличной форме (*представляется, что…*).

Описаны также дескриптивные (косвенные) варианты ментальных перформативов (*как мы уже отмечали*), представляющие собой замаскированные перформативы, используемые либо для перифразирования (*эти данные приводятся в таблице 3* вместо канонического *мы приводим эти данные в таблице 3*) либо для установления связей между высказываниями текста (*далее кратко изложен, выше мы уже указали, что…*).

Кроме ментальных перформативов к дискурсивным словам относятся «сигналы очередности и логической последовательности» [2] (*во-первых*, *прежде всего*), коннекторы (в основном союзы и союзные слова: *тем не менее*, *следовательно* и т.п.) и другие выражения, организующие структуру текста-рассуждения. Среди них могут быть метатекстовые операторы (*подчеркивается, что…*, *по мнению автора*), предполагающие в своем составе сентенциальный аргумент.

Переход от одной мысли к другой в научном тексте осуществляется не только при помощи дискурсивных слов, но с использованием так называемых общенаучных переменных [11] – это абстрактные существительные, называющие аппарат научно-познавательной деятельности (*анализ*, *гипотеза, проблема, аргумент, следствие, идея, понятие, процедура, модель* и др.). Эти существительные играют важную роль в структурно-семантическом упорядочении научной информации и часто употребляются в научных текстах с перформативными глаголами, образуя с ними устойчивые глагольно-именные словосочетания (*подвергнуть анализу, проводить аналогию*) [9, 12].

Очень важным является выделение типичных приемов научного дискурса, соответствующих операциям научного мышления. Сравнив типологии приемов, предлагаемые в работах [4, 7], с результатами нашего анализа текстов, мы остановились на следующем наборе приемов:

- Описание, констатация и характеризация;
- Конкретизация (уточнение) и добавление информации;
- Логические операции и причинно-следственные связи;
- Выделение (подчеркивание) информации, актуализация внимания;
- Определения и допущения;
- Цитирование, иллюстрование и приведение примеров;

- Обобщение и резюмирование (подведение итогов);
- Классификация, аналогия и сравнение;
- Выражение мнения и оценивание;
- Пожелания и рекомендации.

Каждому дискурсивному приему может соответствовать в тексте не одно, а несколько последовательных предложений (в общем случае – сверхфразовое единство), соответствующий сегмент текста будем называть дискурсивным. Некоторые дискурсивные приемы могут использоваться как средство реализации другого приема: например, аргументирование производится при помощи примеров и аналогий. В результате включения одного приема в другой происходит вкладывание соответствующих дискурсивных сегментов, и формируется иерархическая структура текста, для упрощения восприятия которой и служат многочисленные дискурсивные слова.

Однако в научных текстах обычно эксплицитно помечаются и разграничиваются не все шаги и операции. Это нормально, если соответствующие ментальные операции могут быть легко восстановлены читателем по контексту. В ином случае недостаток дискурсивных маркеров повышает неопределенность текста и затрудняет его понимание.

К средствам организации научного дискурса относятся также такие способы структуризации текста, как рубрикация, нумерация, абзацное членение, разбиение на разделы и подразделы. Наблюдается системность всех организующих средств: они функционируют в тексте, замещая и дополняя друг друга. Например, рубрикация используется в составе ментального перформатива: *Перечислим основные положения: А)….В)…*; а заголовки разделов по сути замещают перформатив *Перейдем к*.

Дискурсивно-композиционной структурой научного текста естественно считать взаимосвязанную последовательность всех употребленных в тексте дискурсивных приемов и средств структуризации.

Компьютерный словарь общенаучной лексики

Словарь общенаучной лексики [1, 13] разрабатывается уже более двух лет и охватывает широкий круг семантически и грамматически разнородных слов и выражений общенаучной речи, используемых как дискурсивные маркеры. К числу таких выражений относятся именные и глагольно-именные словосочетания (*сравнительное исследование*, *опровергнуть гипотезу, предположим, что*), предложно-именные сочетания (*в общих чертах*), причастные и деепричастные обороты (*упомянутый выше*, *суммируя все это*), составные предлоги и союзы (*в случае*, *благодаря тому, что*). Подчеркнем, что словарь содержит только общеупотребительные слова естественного языка и не зависит от конкретной научной области (тем самым представляемая им лексика инвариантна в научном языке).

При построении словаря была проведена классификация слов и выражений, классифицирующим принципом был функционально-семантический. Согласно нему, выражения были разбиты на классы исключительно по их смыслу и коммуникативной роли в тексте, без учета их грамматической формы и синтаксических характеристик. В итоге получилось 32 смысловые группы. Каждая группа либо является классом синонимичных (эквивалентных по смыслу) выражений, либо включает несколько близких по семантике подгрупп эквивалентности. Все группы выражений были объединены по общности их функции в научном дискурсе в 5 основных разрядов:

1. Структуризация текста, актуализация внимания (*далее, наконец, перейдем к*);
1. Логические и причинно-следственные связи (*по этой причине, в силу доказанного, однако*);
2. Другие приемы научного рассуждения: конкретизация, выделение, цитирование, иллюстрирование, сравнение и др. (*как пишет автор, к примеру, с одной стороны*);

3. Авторская оценка, в широком ее понимании (*по-видимому, к сожалению, успехом представляется*);
4. Общенаучные переменные (*следствие, проблема*, *теория* и др.) и устойчивые словосочетания с ними.

В общем случае каждая группа эквивалентности содержит несколько синонимичных выражений разной грамматической природы, к примеру, группа следственной связи включает слова и словосочетания *значит, итак, таким образом, тем самым, как видим, следовательно* и др. В группах могут встречаться ментальные перформативные высказывания (*как видим*). В целом словарь охватывает все виды перформативных выражений, а также их дескриптивные варианты.

Проведенная классификация не является строгой, поскольку некоторые выражения из-за их неоднозначности (например, выражение *так что*) помещены сразу в несколько разных групп эквивалентности.

Отнесенные к пятому разряду общенаучные выражения включают как именные, так глагольно-именные словосочетания с общенаучными переменными (*сильный аргумент*, *привести аргумент*), причем в глагольно-именных сочетаниях обычно фигурируют ментальные перформативные глаголы.

Единицей словаря является отдельное слово или устойчивое словосочетание общенаучного лексикона. Словарная статья объединяет связанную со словом или словосочетанием морфологическую, синтаксическую (разрывность/неразрывность, описание валентностей) и семантическую (указание класификационной группы) информацию.

Ядром программной системы, поддерживающей словарь общенаучной лексики, является библиотека процедур, обеспечивающих разнообразные операции поиска словарной информации. Тем самым, словарь допускает использование в составе объемлющей системы автоматической обработки текста.

Распознавание дискурсивной структуры текста

Предлагаемая нами процедура осуществляет дискурсивный анализ только на уровне предложений (но не их составляющих). Две основные проблемы распознавания дискурсивной структуры текста – это отнесение каждого предложения текста к определенному дискурсивному приему, т.е. дискурсивная его характеризация, и разграничение дискурсивных сегментов текста, т.е. определение групп предложений, относящихся к одному и тому же сегменту (приему).

В идеале, отнесение предложения к одному из дискурсивных приемов должно основываться на нескольких факторах:

- Лексических (словарных) показателях (маркерах) приема;
- Анализе синтаксической структуры предложения;
- Соотнесении предложения с содержательным контекстом.

Сочетание всех этих факторов повышает успешность и надежность характеризации. В процедуре анализа последний фактор пока учтен упрощенно: при дискурсивной характеризации предложения учитываются (по возможности) уже установленные дискурсивные характеристики соседних предложений.

Процедура дискурсивного анализа последовательно обрабатывает предложения текста, выявляя вхождения в них словарных дискурсивных слов и выражений и проводя дискурсивную характеризацию предложений. Процедура использует поверхностный синтаксический анализ, который определяет лишь согласованность отдельных членов предложения (например, необходимое согласование составляющих в словарных выражениях) и общий вид предложения (в частности, наличие сентенциального аргумента у обнаруженного метатекстового оператора).

Заметим, что проблема дискурсивной характеризации отдельного предложения не всегда разрешима: например, если дискурсивные маркеры в предложении отсутствуют (что весьма нередко) – в таком случае для анализа привлекаются характеристики соседних предложений. Проблема характеризации может быть разрешима неоднозначно: например, если в предложении встречается несколько разнотипных маркеров, или, что чаще – если встречается один, но неоднозначный маркер. В таких случаях необходимо использовать эвристики, учитывающие, в частности, абзацную структуру текста и выделенные рубрики и выбирающие наиболее правдоподобный вариант характеризации данного предложения. К примеру, если дискурсивные маркеры в предложении отсутствуют, то предложение, не являющееся первым предложением абзаца, относится к предыдущему дискурсивному сегменту, в ином случае оно считается началом нового сегмента, соответствующего приему описания. Среди эвристических правил разграничения дискурсивных сегментов важную роль играют эвристики поиска конца сегмента, определенные для каждой словарной группы эквивалентных общенаучных выражений.

Результатом работы процедуры анализа является дискурсивно-композиционная схема текста, структурным ядром которого является дерево, фиксирующее иерархическую структуру дискурса. Листья этого дерева соответствуют предложениям обрабатываемого текста, нелистовые узлы – выделенным разделам/подразделам текста, рубрикам и дискурсивным сегментам, а ветви дерева – структурно-смысловым связям сегментов и предложений (логическим связям и связям подчинения/вхождения). Кроме указанных связей в дискурсивно-композиционной схеме возможны отсылки от листьев дерева к другим узлам, отражающие встреченные в тексте эксплицитные референтные ссылки. Таким образом, в общем случае дискурсивно-композиционная схема является графоподобной структурой, ребра и узлы которой помечены именами связей и приемов, обнаруженных при дискурсивном анализе.

Не исключен случай, когда распознавание дискурсивной структуры отдельных фрагментов текста или всего текста неуспешно. Это возможно, к примеру, если обрабатываемый текст содержит очень мало дискурсивных маркеров. В этом случае, поскольку обрабатывается научный текст, дискурсивную структуру текста следует считать дефектной. Возможность определения дефектности дискурсивной структуры текста полезна для системы автоматизированного литературно-научного редактирования, призванной выявлять различные типы дефектов научно-технического текста.

Полученная в результате распознавания дискурсивно-композиционная схема может быть использована для переработки исходного текста в аннотацию или реферат. При этом менее значимые дискурсивные сегменты (поддеревья схемы) могут быть просто отброшены, а более значимые сокращены или преобразованы с использованием специальных шаблонов. Одно из возможных направлений преобразования – это замена встреченного в тексте ментального перформативного высказывания вида: *Опишем теперь формальный язык, который…* на выражение *Был описан формальный язык, который…* при этом в реферат переносится дискурсивный сегмент, соответствующий приему описания упомянутого языка.

Очевидно, что распознавание дискурсивной структуры с опорой только на поверхностный синтаксис и словарь дискурсивных выражений представляет упрощенную модель дискурса, и поэтому необходимо, во-первых, исследование результатов работы процедуры на научных тестах из различных предметных областей, а во-вторых, дальнейшее развитие предложенной процедуры как за счет уточнения эвристик, так и учета других факторов внутритекстовой связности текста: анафорических ссылок, лексических повторов и др.

Литература

1. Большакова Е.И. О принципах построения компьютерного словаря общенаучной лексики //Труды Международного семинара Диалог ‘2002 по комп. лингвистике и интеллект. технологиям. М., 2002, Т. 1, с. 19-23.
2. Вежбицка А. Метатекст в тексте // Новое в зарубежной лингвистике. Вып. VIII. М.: Прогресс, 1978, с. 402-421.
3. Голубева А.И. Скрепы как особый вид связочных средств и их функционирование в научном тексте // Научная литература. Язык, стиль, жанры. М.: Наука, 1985. с.272-280.
4. Гордеева О.Н. О принципе дискурсивной организации научно-медицинской статьи // Вестник СпбГУ, Сер.2, 1992, Вып. 4, с.101-103.
5. Маккьюин К. Дискурсивные стратегии для синтеза текста на естественном языке // Новое в зарубежной лингвистике. Вып. XXIV. М.: Прогресс, 1989, с.311-356.
6. Митрофанова О.Д. Язык научно-технической литературы. М.: Изд-во МГУ, 1973. 147 с.
7. Николаев А.М. Языковые особенности и статистическая характеристика речевых актов, реализуемых в тексте рецензии на научно-техническую работу // НТИ. Сер. 2. 1998, № 6. с.28-34.
8. Николаев А.М. Описание семантики научного текста с позиций теории речевых актов (на материале рецензии на научно-техническую работу) // НТИ. Сер. 2. 1998, № 7, с.35-41.
9. Рябцева Н.К. Ментальные перформативы в научном дискурсе // Вопросы языкознания. 1992, № 4. с. 12-28.
10. Свинцов В.И. Логические основы редактирования текста. М.:Книга, 1972. 272 с.
11. Севбо И.П. Сквозной анализ как шаг к структурированию текста // НТИ. Сер. 2. 1989, № 2. с.2-9.
12. Словарь глагольно-именных словосочетаний общенаучной речи. М., Наука, 1973.
13. Bolshakova, E.I. Phraseological Database Extended by Educational Material for Learning Scientific Style. In: ACH/ALLC 2001: The 2001 Joint International Conference. New York University, New York, 2001, p. 147-149.
14. Kurohashi, S., Nagao M. Automatic Detection of Discourse Structure by Checking Surface Information in Sentences. In: COLING 94 Proceedings of the 15th Int. Conf. On Computational Linguistics. Vol. II, 1994, Kyoto, Japan, p. 1123-1127.
15. Ono, K., Sumita K, Miike S. Abstract Generation Based on Rhetorical Structure Extraction. In: COLING 94 Proceedings of the 15th Int. Conf. On Computational Linguistics. Vol. II, 1994, Kyoto, Japan, p. 344-348.