Визуализация семантической структуры и реферирование текстов на естественном языке

М.Ю. Попов
Волгоградский Государственный Технический Университет
p_michael@vistcom.ru

А.В. Заболеева-Зотова
Волгоградский Государственный Технический Университет
zabzot@vstu.ru

С.А. Фоменков
Волгоградский Государственный Технический Университет

Ключевые слова: естественный язык, текст, семантика, визуализация, реферирование, структура текста, межфразовое единство, анализ текста, представление структуры, цельность, связность, высказывание, семантический класс, «Смысл↔Текст», теория риторической структуры.

В данной работе предлагается подход к анализу и представлению семантической структуры текста. Рассматриваются вопросы визуализация структуры текста, графическое представление используется как база для восприятия общей структуры текста и его содержательной части. Предлагаются эффективные нестатистические способы реферирования текстов, использующие семантическую структуру текста на разных уровнях. Показана возможность объединения различных лингвистических теорий с целью получение единой формальной модели текста.

Вступление

Работа с текстом на естественном языке (ЕЯ) до сих пор остается сложной задачей для вычислительной лингвистики. Несмотря на решенность задачи морфологического анализа, работа с текстом идет либо в рамках предложения ЕЯ, либо сразу со всем текстом, как набором линейно упорядоченных слов, словосочетаний и предложений. При этом большинство систем использует вариации статистических методов анализа [5], игнорируя при этом лингвистическую взаимосвязанность и нелинейность ЕЯ. Незадействованность промежуточных уровней представления текста, особенно в виде семантических структур, объясняется, прежде всего, отсутствием эффективных формализмов описания структуры текста.

В данной работе рассматривается подход к представлению семантической структуры текста. Поскольку семантический анализ не является самоцелью, а лишь предварительным этапом интеллектуальной обработки текста, то в качестве выходной информации было выбрано графическое представление структуры текста.

Текстовый документ, в абстракции от своего графического оформления (далее просто текст), содержит значительный объем ЕЯ информации, при этом часть информации оказывается отвлеченной. Следовательно, необходим инструмент для выявления и наглядного представления значимой информации.

Данная работа ориентирована, прежде всего, на нехудожественные тексты, которые принципиально отличаются от художественных следующими основными признаками:

- однозначность восприятия;

- отсутствие непосредственной связи между коммуникацией и жизнедеятельностью человека;
- отсутствие эстетической функции;
- эксплицитность содержания.

Текст как объект лингвистического анализа

Текст определяется как динамическая единица высшего порядка, как письменное произведение, обладающие признаками связности и цельности – в информационном, структурном и коммуникативном плане. Текст представляет собой объединенную по смыслу последовательность знаковых единиц, основными свойствами которой являются связность и цельность [2]. Как функционально–семантико–структурное единство он обладает определенными правилами построения, выявляет закономерности смыслового и формального соединения составляющих его единиц.

Любой текст представляет собой знаковую модель некоторого мыслительного содержания и воспроизводит фрагмент концептуальной системы автора текста, поэтому при анализе текста важно соотношение «автор↔текст↔читатель». Иными словами текст есть «линейная развертка» некоторой системы знаний. Однако текст по своей природе не линеен. Высокая популярность технологии гипертекстов подтверждает нелинейный характер ЕЯ текста. Таким образом, линейность печатных текстов есть веками устоявшийся принцип. Линеаризация происходит на этапе «автор→текст», при этом на текст дополнительно накладывается авторский стиль, а на этапе «текст→читатель» происходит обратное преобразование.

Текст как объект лингвистического исследования представляется прежде всего как информационное и структурное единство, как функционально завершенное речевое целое. Именно это качество текста в настоящее время дает возможность определить достаточно четкие закономерности текстообразования. Целостность и связность – основные конструктивные признаки текста – отражают содержательную и структурную сущность текста. Различают следующие связанности текста:

- функциональная – заключается в функциональном взаимодействии смыслов слов и словосочетаний текста;
- структурная – обеспечивается наличием тема-рематических связей в тексте;
- семантическая – обеспечивается единством темы и микротемы.

Исследователи также выделяют локальную и глобальную связности.

Локальная связность – это связность линейных последовательностей (высказываний, межфразовых единств). Данный вид связности определяется межфразовыми синтаксическими связями (вводно-модальными и местоименными словами, видовременными формами глаголов, лексическими повторами, порядком слов, союзами и т.д.).

Глобальная связность – это то, что обеспечивает единство текста как смыслового целого, его внутреннюю цельность. Глобальная связность (а она приводит к содержательной целостности текста) проявляется через ключевые слова, тематически и концептуально объединяющие текст и его фрагменты.

Связность текста проявляется через внешние структурные показатели, через формальную зависимость компонентов текста. Целостность же текста усматривается в связи тематической, концептуальной, модальной. Таким образом, понятие цельности текста ведет к его содержательной и коммуникативной организации, а понятие связности – к форме, структурной организации.

Структура текста

Текст реализует структурировано представленную деятельность, а структура деятельности предполагает субъект и объект, сам процесс, цель, средства и результат. Эти компоненты структуры деятельности отражаются в разных показателях текста – содержательно-структурных, функциональных, коммуникативных.

Текст имеет свою микро- и макро семантику, микро- и макроструктуру. Семантика текста обусловлена коммуникативной задачей передачи информации; структура текста определяется особенностями внутренней организации единиц текста и закономерностями взаимосвязи этих единиц в рамках текста как цельного сообщения.

Следует различать внешнюю (композиционную) и внутреннюю структуры текста. На уровне композиционном выделяются: предложения, абзацы, параграфы, разделы, главы, подглавки, страницы и др. Все композиционные элементы, кроме предложения, лишь косвенно связаны с внутренней структурой, далее они рассматриваться не будут. Предложения задают границы действия знаков препинания, анафорических и катафорических ссылок. Под структурой текста будет пониматься его внутренняя структура.

Единицами внутренней структуры текста являются:

- высказывание – реализованное предложение, любое высказывание есть предложение ЕЯ, однако обратное не верно;
- межфразовое единство (МФЕ или сверхфразовое единство) – ряд высказываний, объединенных семантически и синтаксически в единый фрагмент. Ядром МФЕ является высказывание-зачин, которое является «автосемантичным» в том смысле, что оно не подчинено напрямую другим высказываниям и сохраняет смысл, даже будучи выделенным из контекста;
- фрагменты-блоки – совокупность межфразовых единств, обеспечивающих тексту целостность благодаря реализации дистантных и контактных смысловых и тематических связей.

Единицы семантико-грамматического (синтаксического) и композиционного уровня находятся во взаимосвязи и взаимообусловленности, в частном случае они даже в «пространственном» отношении могут совпадать, накладываясь друг на друга, например, межфразовое единство и абзац, хотя при этом они сохраняют свои собственные отличительные признаки.

С семантической, грамматической и композиционной структурой текста тесно связаны его стилевые и стилистические характеристики. Каждый текст обнаруживает определенную более или менее ярко выраженную функционально-стилевую ориентацию (научный текст, художественный и др.) и обладает стилистическими качествами, диктуемыми данной ориентацией и, к тому же, индивидуальностью автора.

Стилистические качества текста подчинены тематической и общей стилевой доминанте, проявляющейся на протяжении всего текстового пространства.

Построение текста определяется темой, выражаемой информацией, условиями общения, задачей конкретного сообщения и избранным стилем изложения.

Анализ текста

Одним из важных моментов анализа является многоуровневость представления структуры анализируемого текста. Используется следующая иерархия уровней:

- исходный текст как линейная последовательность символов;
- линейная последовательность морфологических структур;
- линейная последовательность высказываний;

- сеть взаимосвязанных МФЕ.

Соседние уровни явно связаны друг с другом, и на различных этапах анализа все уровни сохраняются, что позволяет воспользоваться информацией с любого уровня представления.

Предварительным этапом анализа является нормализация текста (графематический анализ), приводящий исходный текст к каноническому виду. Являясь определенного рода текстовым препроцессором, графематический анализатор решает следующие задачи: удаление нетекстовых символов, разделение цепочки символов на слова, выделение цифр, чисел, дат, неизменяемых оборотов и сокращений, деление на предложения и абзацы. Результатом анализа является линейная последовательность слов, включая служебные (знаки препинания, метки конца предложения).

Морфологический анализ решает частную задачу приведения всех слов к каноническому виду. Это первый этап анализа, в котором появляется явная многозначность, которая обусловлена совпадением морфологических структур у различных словоформ. Цель морфологического анализа состоит в получении основ, т.е. словоформ с отсечёнными окончаниями. Причём каждой словоформе ставится в соответствие значения грамматических категорий, т.е. совокупности грамматических значений (род, падеж, склонение и т.д.) Результатом является линейная последовательность морфологических структур, каждая из которых может иметь несколько вариантов.

Семантический анализ предполагает наличие естественно-семантического словаря. Входами такого словаря являются. Точность семантического анализа целиком определяется полнотой и корректностью семантического словаря. Следует отметить, что здесь под семантическим анализом понимается лингвистический семантический анализ, т.е. слова ЕЯ соотносятся с некоторыми «семантическими классами», которые никак не соотносятся с реальным миром.

На этапе семантического анализа происходит отбор нужных для данного предложения морфосемантических альтернатив и связывание слов в единую структуру. Результатом семантического анализа предложений является упорядоченное множество записей суперпозиций из базисных функций (лексических функций в терминологии модели «Смысл↔Текст») и семантических классов (базовых понятий) [1, 9]. Часть семантических классов в лингвистической формуле может оказаться незаполненными, что может объясняться неполнотой исходного предложения или наличием референций. Исходный порядок морфем не сохраняется, однако сохраняется линейность текста, который на этом этапе представляется как последовательность предложений на семантическом языке.

В структуре текста предложение на семантическом языке соответствует одному высказыванию. Связанные межфразовыми отношениями высказывания отвечают за локальную связность текста.

Референциальный анализ использует вводно-модальные и местоименные слова, формы глаголов, порядок слов, союзы для установления связей между предложениями семантического языка.

Для выделения в тексте МФЕ используется понятие микротемы (частной темы) и текущего контекста. МФЕ монотематично, поэтому переход к другой микротеме есть граница сверхфразовых единств. Текущий контекст есть набор семантических классов, встречавшихся в текущем МФЕ. С его помощью реализуется разрешение локальных референций («этот», «который», «его») и выделение начального высказывания – ядра МФЕ.

Формальное разделение высказываний на тему и рему позволяет выделять в структуре текста тема-рематические структуры, которые могут использоваться при построении реферата.

Единство темы проявляется в следующих свойствах текста:

- регулярная повторяемость ключевых слов;
- синонимизация ключевых слов;
- повторную номинацию ключевых слов;
- тождество референции, т.е. соотношением слов (имен и их заместителей) с одним и тем же предметом изображения;
- наличие импликации, основанное на ситуативных связях. Наличие одних отображаемых предметов предполагает наличие и других, ситуативно связанных с ними.

Анализ структуры связей использует элементы Теории Риторических Структур [7, 10]. В частности, используется базовый набор риторических отношений как связей между МФЕ и построение нелинейной сети МФЕ. Открытость набора связей предполагает его расширение и адаптацию для анализа структуры текстов.

Предпосылками использования данной теории являются высокая степень совпадения терминов «межфразовое единство» и «дискурсивная единица», а также «предложение семантического языка», «высказывание» и «элементарная дискурсивная единица». Несмотря на отсутствие единой устоявшей терминологии, есть основания полагать, что сходные термины различных теорий описывают одни и те же сущности ЕЯ текста.

Визуализация и реферирование

Как было отмечено, визуализация используется как наглядное представление работы системы в графическом виде. Наличие направленных связей предполагает использование ориентированных графов, при этом часть информации (например, тип связи) может кодироваться цветом. Важной представляется решение задачи о минимальном пересечении связей путем перемещения вершин графа, поскольку оно повышает наглядность визуализации.

Многоуровневое структурирование текста позволяет очень гибко подходить к решению задачи реферирования. В частности, возможны следующие способы формирования реферата:

- удаление малозначащих МФЕ (по связям типа Elaboration, Background). Преимуществом метода является гарантированное сохранение значащей информации, недостатком – низкая степень сжатия. Возможен вариант статического ранжирования важности связей и отброс наименее значащих, при этом количество отброшенных МФЕ зависит от заданной степени сжатия информации;
- сокращение МФЕ – замена МФЕ его смысловым ядром, равным одному высказыванию. Преимущество метода – высокая степень сокращения, недостаток – возможная несвязность полученного реферата за счет удаления референций из МФЕ. Данный недостаток может быть скомпенсирован с помощью учета референциальных высказываний и оставления их в МФЕ;
- комбинированный способ является наиболее продвинутым, совмещая преимущества обоих методов. При этом возможно дополнительное уточнение реферата с помощью статистических методов, с использованием семантических классов, глобального контекста и синонимических связей в словаре.

Таким образом, учет и использование внутренней структуры текста позволяет составить высоко релевантный, компактный и связный реферат исходного ЕЯ текста. Также следует отметить, что использование семантического анализа автоматически означает снятие стилевой информации и, в некоторых случаях, оттенков смысла. Однако такие потери практически не влияют на качество реферата.

Заключение

Представляется, что дальнейшее развитие предлагаемого подхода приведет к созданию системы, объединяющей достоинства модели «Смысл↔Текст» и существующие теории о структуре и семантике ЕЯ текстов, а также восполнить пробел формального анализа промежуточных уровней строения текста. Выявление многоуровневой структуры текста позволяет по-новому и более качественно решить проблему построения рефератов ЕЯ текстов.

Natural language text abstracting and visualization of semantic structure

M. Y. Popov

A. V. Zaboleeva-Zotova

S. A. Fomenkov

Key words: natural language, text, semantic, visualization, abstraction, text structure, interphrase unity, text analysis, structure declaration, integrity, coherence, utterance, semantic class, «meaning↔text» theory, rhetorical structure theory.

In this work we offer an approach to natural language text semantic structure declaration and analysis. We review text structure visualization issues. Graphical representation is use as basis for text general structure and content perception. We offer effective non-statistic text abstracting methods. Those methods use different levels of text semantic structure. We have shown possibility of integration of different linguistic theories for elaboration of unified formal text model.

Литература

1. Апресян Ю.Д. Избранные труды, том I. Лексическая семантика: 2-е изд., испр. и доп. – М.: Школа «Языки русской культуры» РАН, 1995. – VIII с., 472 с.
2. Валгина Н.С. Теория текста: Учебное пособие. Москва: Изд-во МГУП «Мир книги», 1998. – 210 с.
3. Гальперин И.Р. Текст как объект лингвистического исследования. М.: Наука, 1981. – 140 с.
4. Демьянков В.З. «Событие» в семантике, прагматике и в координатах интерпретации текста // Изв. АН СССР. Серия литературы и языка. 1983. Т. 42. № 4. С.320-329.
5. Ермаков А.Е. Тематический анализ текста с выявлением сверхфразовой структуры Информационные технологии. – 2000. – N 11.
6. Заболеева-Зотова А.В. Естественный язык в автоматизированных системах. Семантический анализ текстов: Монография / ВолгГТУ. – Волгоград 2002. – 228 с.
7. Литвиненко А.О. Описание структуры дискурса в рамках Теории Риторической Структуры: применение на русском материале. // Труды Международного Семинара Диалог’2001 по компьютерной лингвистике и ее приложениям. Том 1. Теоретические проблемы. Аксаково 2001. С. 159-168.
8. Русский язык. Текст как целое и компоненты текста. – М., 1982.
9. Тузов В.А. Компьютерная семантика русского языка. // Труды Международного Семинара Диалог’2001 по компьютерной лингвистике и ее приложениям. Том 1. Теоретические проблемы. Аксаково 2001.
10. William Mann, Sandra A. Thompson. Rhetorical structure theory: A theory of text organization. University of Southern California, 1987.