РАЗРАБОТКА ТЕКСТОВОЙ БАЗЫ НА ОСНОВЕ АНАЛИЗА СТРУКТУРЫ НАУЧНОГО ТЕКСТА

Анастасия Дыбина

***Аннотация**: В статье описан алгоритм построения текстовой базы, основанный на анализе структуры научного текста, понимаемого как связного, в терминах когезии и когерентности в рамках лингвистики текста. Исследования проводятся в области автоматического реферирования с целью построения интеллектуальной системы, основанной на глубинной семантике текста.*

***Ключевые слова**: автоматическое реферирование, текстовая база, структура текста, связный текст, когезия, когерентность.*

***ACM Classification Keywords**: E. Data, E.2 Data storage representation.*

Введение

С точки зрения необходимости развития современного общества, повышения эффективности обработки постоянно увеличивающегося объема научно-технической информации в сети Интернет, возрастает и признается наиболее актуальным внедрение в практику систем и технологий, основанных на знаниях. “Сегодня интеллектуальные системы олицетворяют одно из наиболее впечатляющих и приносящих огромную практическую отдачу приложений современных компьютерных технологий к решению многих задач” [Бакаев, 1993]. Тем не менее, несмотря на возрастание сложности поставленной задачи, практика создания интеллектуальных автоматизированных систем, в частности, системы автоматического реферирования, на разработку которой направлены исследования автора, сталкивается со многими проблемами, наиболее значимыми из которых, по мнению ряда ученых, является вопрос о методах извлечения знаний из текстов на естественном языке, их формализации, структурирования и использования в компьютерных системах для решения сложных задач, требующих значительного опыта человека. Предложенная в данной статье концепция создания системы автоматического реферирования основывается на понимании текста как связного, структура которого исследуется в терминах когезии и когерентности в рамках лингвистики текста, и включает разработку текстовой базы, являющейся семантическим представлением текста, на основе анализа структуры текста, и онтологии: метаонтологии для описания категорий реферативных конструкций и онтологии предметных областей [Лазаренко, 2009].

Предпосылки исследования

Исследованием связного текста занимается научная дисциплина – лингвистика текста, которая оформилась в 60–70-х годах XX в. Возникновение и развитие новой отрасли лингвистики связано с комплексом изменений, которые произошли в лингвистических ценностях, и отразило весь период развития лингвистической науки в целом, поскольку достаточная изученность предложения вне связи с контекстом показала, что сами по себе предложения не являются законченными и независимыми

единицами языкового произведения и не представляют возможности теоретического описания языковых явлений, и связано, прежде всего, с именем русского литературоведа и исследователя фольклора Владимира Яковлевича Проппа, книгу которого “Морфология сказки” научный мир увидел в 1928 г. Переводы книги на английский и французский языки способствовали формированию лингвистики текста многих западных исследователей, таких как Клода Леви-Стросса (Claude Lévi-Strauss), Ролана Барта (Roland Barthes), Цветана Тодорова (Tzvetan Todorov) и Альгирдаса-Жюльена Греймаса (Alguirdas Julien Greimas). Под влиянием чешской лингвистической школы складывается теория текста Роланда Харвега (Roland Harveg), Адрианус ван Дейка (Adrianus Van Dijk), Вольфганга Дресслера (Wolfgang Dressler), Карела Гаузенблаза (Karel Gauzenblaz), Петра Сгалла (Petr Sgall), Ирэны Беллерт (Irena Bellert) и других ученых.

В отечественной лингвистике формирование лингвистики текста связано с именами А. М. Пешковского, И. А. Фигуровского, Н. С. Поспелова, Н. Ю. Шведовой, Т. Г. Винокур, Е. В. Падучевой, Т. М. Николаевой, С. И. Гиндина, И. Р. Гальперина, В. И. Перебейнос и других исследователей.

С исследованием категорий связного текста связаны работы отечественных лингвистов И. Р. Гальперина, С. И. Гиндина, Н. Н. Леонтьевой, Н. С. Валгиной, З. Я. Тураевой, Л. В. Орловой, И. А. Белокопытовой и некоторых других, а также практические исследования западных лингвистов Роберта-Алена де Богранда (Robert de Beaugrande) и В. Дресслера. Среди множества категорий, выделяемых данными учеными, для нашего исследования наибольшую практическую значимость имеют категория смысловой связности – когерентность и категория композиционной связности – когезия, изученные на научных статьях различных предметных областей в рамках двух подходов к исследованию структуры текста, описанных Г. Г. Москальчук, А. А. Стриженко и Л. И. Кручининой, – прагмастилистического и композиционно-тематического, предполагающих изучение того, как отдельные предложения связываются в единое содержательное и структурное целое, что позволяет воспринимать текст как единое коммуникативное целое [Стриженко, 1985].

В области автоматической обработки текста и разработки технологий автоматического реферирования, включающего процедуры понимания, обобщения и сжатия смысла исходного текста [Лазаренко, 2007], представляют интерес практические исследования Э. Ф. Скороходько, В. Е. Берзона, И. П. Севбо, Р. Г. Пиотровского, О. В. Лазаренко, Д. И. Панченко, Г. П. Луна, Д. Г. Лахути и других ученых.

Особо следует отметить практические исследования американского основоположника крупнейших современных подходов к моделированию процесса понимания связного текста Роджера Шенка (Roger Schenk) и нидерландского ученого Адрианус ван Дейка.

В настоящее время разработкой интеллектуальных систем обработки языка занимаются Украинский языково-информационный фонд НАН Украины, факультет кибернетики Киевского национального университета им. Т. Шевченко, Украинская Лингвистическая Лаборатория КНУ им. Т. Г. Шевченко, Институт прикладной и математической лингвистики МГЛУ, Санкт-Петербургский государственный университет и другие.

Постановка проблемы

Актуальность наших исследований заключается в том, что они направлены на разработку когнитивной или семантико-контекстной модели автоматического реферирования, основанной на глубинном понимании смысла текста на естественном языке.

Обзор теоретических и практических исследований отечественных и зарубежных ученых показал достаточную изученность смысловой и композиционной структуры художественного, публицистического, и, в некоторой степени, научного текстов. Тем не менее, отмечается отсутствие комплексных исследований, направленных на изучение структуры текста с целью выявления глубинной семантики для построения индикативного реферата автоматическим способом. **Цель** статьи – описать структуру научного текста с целью выделения предложений, содержащих ключевые слова и слова-указатели на смысловые аспекты: объект, результат, метод, цель, и построения текстовой базы.

Изложение основного материала

Обозначив основные категории связного текста – когезию и когерентность, проанализировав структурную и смысловую связность научных текстов на базе описанных категориальных признаков в работе [Дибіна, 2010], мы подошли к проблеме построения текстовой базы.Результаты, полученные на сегодня, можно сформулировать следующим образом:

1. Мы разделяем точку зрения [ван Дейк, 1988], согласно которой “каждый тип текста имеет свои языковые и когнитивные расхождения”, поэтому при анализе определенного текста следует учитывать его функциональную принадлежность.

2. Анализ текста мы начали с изучения заголовка научных статей, поскольку в нашем подходе он имеет решающее значение для выделения в тексте среди ключевых слов тех, которые понимаются как “функционально нагруженные лексемы, которые … образуют семантический стержень текста” [Дыбина, 2011].

Анализ семантико-синтаксической структуры заголовка проведен в статье [Лазаренко, 2004]. На основе описанной сравнительной характеристики понятий, которые принимают участие в заполнении семантической структуры заголовка, мы провели исследование заголовков научных статей, и выявили, что в его построении участвуют два актанта: первый – отглагольное существительное (иногда с прилагательным), выражающее действие, направленное на достижение результата, второй – существительное, представляющее объект исследования. Например, “Лингвистический анализ (действие, направленное на достижение результата) текста: *речевая манипуляция* (объект)”, “О математическом моделировании (действие, направленное на достижение результата) *отсубстантивных имен существительных* (объект) русского языка” и подобные.

Еще один тип заголовков, выраженный только одним актантом в виде существительного и представляющий объект исследования. Приведем примеры: “Слово в вычислительной лингвистике”, “Внешнее низкочастотное электрическое поле человека”, “О психодинамическом эксперименте (теория

среднего)". Следует отметить, что первые два типа заголовков являются наиболее распространенными в научной литературе.

В других заголовках, наоборот, на первом месте стоит объект исследования, за которым следует действие, направленное на достижение результата: "*Этнические стереотипы* (объект) в медийном дискурсе: механизмы интерпретации и попытка классификации (действие, направленное на достижение результата)", что является не очень распространенным типом заголовков в научных статьях.

Ключевые слова в заголовке позволили найти в тексте фрагменты, описывающие эти понятия в соответствии с определенной структурно-композиционной последовательностью. А через эти слова выделить слова-указатели, указывающие на смысловые аспекты, встречающиеся в реферате – объект, результат, метод и цель.

3. Согласно концепции понимания, предложенной в работе [ван Дейк, 1988], для описания глобального смысла текста необходимо построить схемы, которые бы обеспечивали "быстрый анализ структур и построение относительно простой и неизменной семантической конфигурации", а для построения этой схемы "необходимо дать электронной вычислительной машине параметры построения текста" [Бук, 2010]. Поэтому другой, не менее важной составляющей процесса понимания текста, является описание его связности, т.е. установление важных связей между предложениями текста.

4. С этой целью мы провели исследование смысловой (когерентной) и композиционной (когезиальной) связности научных текстов [Дибіна, 2010; Дыбина, 2011] и пришли к выводу, что композиционную связность научного текста можно представить следующей схемой: введение, где мы выделяем преамбулу (обоснование актуальности, постановка задачи, история вопроса) и постановочную часть (определение объекта исследования, метода), основная часть (изложение материала) и заключение (формулирование результатов исследования, перспективы), при этом данный тип связности не всегда проявляется в последовательном расположении обозначенных частей текста, что свидетельствует о том, что когезиальная связность может быть только вспомогательным инструментом смыслового анализа, в то время как когерентность должна стать основным объектом исследования в выявлении "метатекста".

5. Описанная и построенная интенсиональная модель индикативного реферата в работе [Лазаренко, 2007], позволила определить и четко обозначить те способы подачи знания в системе автоматического реферирования,которые необходимы для заполнения модели реферата экстенсиональной семантикой, а именно: текстовая база и онтология.

6. Текстовая база в нашем понимании – это семантическое представление текста, в нашем случае – научного, содержащее информацию, выраженную самим текстом [Лазаренко, 2011]. Т.е. объектом исследования должны стать референциальные значения текста, поскольку именно в них содержится вся основная информация научного текста.

Построенная модель индикативного реферата, классификация лексем, участвующих в заполнении данной модели, классификация лексем заголовка позволили подойти к построению текстовой базы, ограничив ее

содержание на данном этапе выделением нескольких смысловых аспектов текста – объекта, результата, цели и метода. С этой целью в тексте были выделены слова-указатели на эти аспекты. В качестве примеров рассмотрим некоторые из них

Объект: рассмотрим, рассматриваются; называется; понимается; описывается; представляет собой; (целесообразно, можно, будем) рассматривать и др.

Результат: в заключение следует отметить, отметим, что; (был) описан, использован, получен, предложен, рассмотрен; полученные в результате; эксперименты, результаты показали (свидетельствуют); разработана и др.

Объект + результат: целью (статьи, данной работы), (нашей) целью (является, было); предпринята попытка, представляет попытку и др.

Метод: методами … установлено, что; ведется с использованием; описывается метод; (широко) используется метод; математическим аппаратом для построения является; одним из распространенных методов является и др.

С помощью этих указателей в каждом тексте были выделены предложения, которые и вошли в текстовую базу соответствующего текста как главные (но не все) смысловые аспекты.

Статья 1. *О моделировании фонетического отношения.*

В настоящей статье математически описываются фонетические признаки звука, а также связь между признаками и фонемой.

Статья 2. *Слово в вычислительной лингвистике.*

Нами дано определение слова для систем вычислительной лингвистики, работающих с письменной формой русского языка, и на основании этого определения предложен алгоритм выделения слова из текста.

Исследователи, придерживающиеся машинной, формальной ориентации, предложили определение слова письменной формы языка в следующем виде: слово – набор букв алфавита, ограниченный с обеих сторон пробелами.

Статья 3. *Об изоморфизме алгебр конечных предикатов.*

Рассмотрим вопрос об изоморфизме одного класса конечных алгебр, а именно, алгебр конечных предикатов.

Под конечной алгеброй будем понимать алгебру с конечным числом элементов.

Рассмотрим произвольную алгебру с алгебраическими операциями и элементами, а также множество Ф.

Данные примеры наглядно показывают, что ограничение текстовой базы четырьмя смысловыми аспектами позволяет точнее выбирать информацию из текста для заполнения имеющейся модели реферата.

Выводы

Данное исследование представляет интерес с точки зрения разработки системы автоматического реферирования с опорой на процессы понимания текста на естественном языке. Поскольку “процесс понимания допускает частичное планирование (или ожидание) (в нашем случае – ожидание) структур и значений предложений и целых текстов” [ван Дейк, 1988], на разработку таких структур направлены наши изыскания. С этой целью мы проанализировали структурную и смысловую связность научного текста, выделили слова-указатели на смысловые аспекты – объект, результат, цель, метод, и с помощью этих слов-указателей выделили в каждом тексте предложения, указывающие на данные смысловые аспекты, которые и вошли в текстовую базу, необходимую для заполнения модели реферата.

Перспективы исследования

Проведенное исследование показало, что разработка метода выделения смысловых аспектов текста, необходимых для построения реферата, позволит создать интеллектуальную систему автореферирования, основанную не на статистических методах, а на глубинном понимании текста.

Литература

[1] Бакаев А.А., Гриценко В.И., Козлов Д.Н. Методы организации и обработки баз знаний. – К.: Наукова думка, 1993. – 150 с.

[2] Лазаренко О.В., Панченко Д.И. Роль онтологий при обработке знаний в семантическом “web” // Лінгвістичні студії. – Донецьк, ДонНУ, 2009. – Вип. 18. – С. 258-262.

[3] Стриженко А.А., Кручинина Л.И. Об особенностях организации текстов, относящихся к разным функциональным стилям: Монография. – Иркутск, Изд-во Иркут. ун-та, 1985. – 176 с.

[4] Лазаренко О.В., Яковенко А.А. Моделювання процессу узагальнення в системі автоматичного реферування: Монографія. – Х., Видавництво НУА, 2007. – 124 с.

[5] Дибіна А. Дослідження когезії та когерентності як основних категорій зв’язного тексту // Матеріали IV Міжнародної конференції молодих вчених CSE-2010 “Комп’ютерні науки та інженерія”. – Львів, Видавництво Національного університету “Львівська політехніка”, 2010. – С.128-129.

[6] Дейк ван Т.А., Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике. – Вып. 23: Когнитивные аспекты языка. – М., 1988. – С.153-211.

[7] Лазаренко О.В., Попова Т.В. Аналіз смислової структури заголовка як тексту з максимальним рівнем узагальнення // Проблеми семантики слова, речення та тексту: зб. наук. праць. – К. : КНЛУ, 2004. – С. 143-149.

[8] Бук С., Ровенчак А. Засади анотування внутрішніх елементів тексту у збалансованому текстовому банку даних української мови // Збірник наукових праць “Людина. Комп’ютер. Комунікація”. – Львів, Видавництво Національного університету “Львівська політехніка”, 2010. – С.66-69.

[9] Дыбина А.В. О структурно-композиционной связности научного текста // Программа и материалы конференции “Молодые ученые Харьковщины”, 2011. – Х., Изд-во НУА, 2011. – С.21-25.

[10] Лазаренко О.В., Дыбина А.В. От модели реферата к модели понимания текста // Доповіді міжнародної наукової конференції “Megaling’2011. Горизонти прикладної лінгвістики та лінгвістичних технологій”, 2011. – Крим, Партеніт. – С.87-88.

Информация об авторе

Анастасия Дыбина *– аспирантка, Харьковский гуманитарный университет «Народная украинская академия», ул. Лермонтовская, 27, Харьков, 61000, Украина, e-mail: generosite@mail.ru*