УДК 004.912:81’37

О. В. Лазаренко

ЗНАНИЕОРИЕНТИРОВАННЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ. СЕМАНТИЧЕСКИЙ АСПЕКТ: ОПЫТ НАУЧНОЙ ШКОЛЫ

Статья посвящена рассмотрению основных направлений работы научной школы «Знаниеориентированные информационные системы. Семантический аспект». Рассмотрены основные достигнутые результаты и современное состояние начных исследований.

Благодаря современным информационным технологиям человек получил практически неограниченный доступ к информации, накопленной человечеством. Работать с такими информационными массивами стало чрезвычайно сложно. Резко возросла потребность в системах, способных существенным образом снизить напряжение при работе с большими массивами информации. К числу таких систем относятся системы, способные анализировать и обрабатывать информацию, в частности, системы автоматического реферирования. Несмотря на проведение активных исследований в области автоматической обработки информации задача создания высококачественного автоматического реферата остается до сих пор нерешенной. Все попытки автоматизации реферирования, которые предпринимались до последнего времени, привели к однозначному выводу о том, что понимание текста без опоры на знания невозможно. Поэтому особую актуальность приобретает задача изучения процессов понимания текстов и разработка интеллектуальных систем реферирования с опорой на знания. Чтобы приблизится к решению этих задач в сфере информационных технологий наметился переход от технологий, ориентированных на исследование количественной стороны информационных процессов, к технологиям, основанным на знаниях как качественных составляющих этих процессов [1].

В рамках нашей школы ведутся исследования, целью которых является моделирование процесса реферирования, его формализация и создание на основе разработанных формализмов системы автоматического реферирования (АР) с опорой на знания.

Одним из перспективных направлений в данной области исследований является, на наш взгляд, поиск путей и методов автоматического сжатия (обобщения) текста.

Приступив к моделированию процесса реферирования, представляющего собой совокупность сложнейших процессов понимания, обобщения и сжатия смысла, мы сформулировали гипотезу, согласно которой процесс реферирования следует начинать с изучения не самих процессов понимания, а с изучения их результата – с реферата. Причем не развернутого, информативного, а сжатого, индикативного и лишь научных текстов. И не только потому, что он наиболее востребован сегодня в системах поиска в Интернете, но и, в первую очередь, потому, что рассматриваем его как отправную точку в исследовании процесса реферирования, как объект наиболее простой по форме, но отражающий особенности любого типа реферативного текста. Иными словами, мы начали с исследования наиболее общих закономерностей для всех видов сжатия текста, т. е. сознательно пошли на теоретическую и эмпирическую неполноту на начальном этапе. С тем, однако, чтобы в дальнейшем, оттолкнувшись от понимания изученных закономерностей, расширять область исследования.

Моделирование процесса реферирования, таким образом, может быть сведено к нескольким самостоятельным, но взаимообусловленным задачам:

1) исследованию синтаксической и семантической структуры реферативных конструкций и первичных текстов и построению модели реферата в виде набора синтаксических схем;

2) моделированию процедуры компрессии (обобщения) текста при реферировании;

3) построению правил наполнения синтаксических конструкций реферата понятиями, обобщающими семантику исходного текста;

4) классификации понятий, заполняющих актантные структуры реферативных конструкций;

5) построению модели представления знаний в системе АР в виде антологий, содержащих независимую от ситуации информацию, и текстовой базы, являющейся «информационным ядром», содержащим информацию, зависящую от ситуации.

Для решения этих задач необходимо было:

• разработать методику формирования признаков, характеризующих реферат как результат компрессии информации на разных уровнях процесса реферирования;

• провести исследование синтаксической структуры реферативных конструкций и первичных текстов и построить модель процесса сжатия на синтаксическом уровне;

• проанализировать механизмы сжатия на семантико-синтаксическом, собственно семантическом и лексико-семантическом уровнях и построить модель процесса сжатия на семантическом уровне;

• построить модель реферата в виде типовых для индикативных рефератов семантико-синтаксических структур;

• выявить и классифицировать лексику, используемую в формировании индикативных рефератов;

• разработать на базе созданных словарей антологию верхнего уровня, антологию общенаучной лексики и антологию предметных областей;

• разработать механизм построения текстовой базы для заполнения модели реферата необходимой информацией.

Проведенное нами исследование смысловой и синтаксической структур реферата позволило сделать выводы о том, что такое обобщение в реферировании и к каким особенностям в структуре реферативных предложений оно приводит, и на основании выявленных особенностей синтактико-семантической структуры этих предложений построить модель индикативного реферата [2].

Не углубляясь в детали этих исследований, остановимся на некоторых выводах, имеющих принципиальное значение для понимания специфики проведенного нами исследования.

Анализ синтаксической структуры реферативных предложений показал, что набор характерных для них синтаксических конструкций очень ограничен. Это преимущественно простые предложения с неопределенно-личными и страдательно-возвратными конструкциями. Он также целиком подтвердил распространенную в современной лингвистике вербоцентрическую концепцию, согласно которой главной составляющей простого предложения является элементарная

глагольная конструкция, состоящая из основного глагола (предикатного ядра) и зависимых от него обязательных элементов (актантов). Общей оказалась семантика синтаксических структур реферативных предложений: «отношение между субъектом и его предикативным признаком – состоянием как результатом действия».

Все это в полной мере отвечает основной функции реферата – краткому изложению того, что исследовано, открыто, описано. То есть сжатие в процессе реферирования имеет свою ярко выраженную специфику: уход от развернутых описаний,– и осуществляется на синтаксическом уровне за счет использования в тексте простых предложений с фиксированной семантикой синтаксических конструкций: «результат действия» или «направленность действия на достижение результата».

Анализ большого корпуса индикативных рефератов показал, что сжатие на смысловом уровне происходит не только в рамках отдельного предложения, но и всей содержательной структуры реферата. Оказалось, что индикативный реферат, как правило, состоит из двух предложений со значениями – *объект исследования* и *результат исследования*. При этом первым предложением в реферате является предложение со значением *объекта*, а второе – со значением *результата*. Это разрешило представить модель реферата в виде набора синтаксических конструкций с этими значениями.

В отличие от реферата исходный текст содержит большой массив общеупотребительной и общенаучной лексики. Поэтому основная задача реферирования заключается в переходе от развернутой смысловой структуры текста к обобщенной смысловой структуре реферата.

Справиться с этой задачей может лишь система, которая способна анализировать содержание текстовых документов не только по формальным, но и, в первую очередь, по смысловым признакам. Наш подход к решению этой задачи является одной из попыток создания такого рода системы.

Анализ содержания текста мы начали с анализа его заголовка, который рассматриваем как реферат минимального объема или как текст с максимальным уровнем обобщения смысла. Поскольку

главной составляющей процедуры реферирования является компрессия (сжатие), нас также интересует, в чем заключается отличие компрессии при переходе от текста к заголовку, от текста к реферату и от реферата к заголовку, и в чем она выражается. Компрессия при переходе от текста к реферату (*компрессия 2*) описана в работе [2]. Изучая смысловую структуру заголовка в сравнении с реферативной структурой, мы попробовали разобраться, как выражается компрессия при переходе от реферата к заголовку (*компрессия 3*).

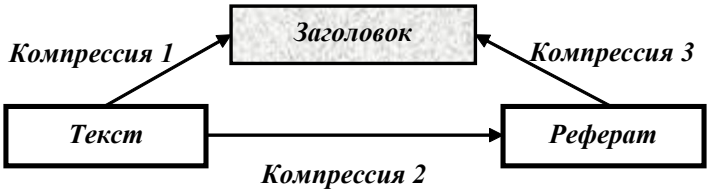

Рис. 1. Виды компрессии в системе реферирования

В результате исследования смысловой и синтаксической структуры заголовка было обнаружено его сходство со структурой реферата. Также как и в индикативном реферате, смысловая структура заголовка состоит из двух метазначений – *объект* и *результат*. Но в отличие от реферата они являются элементами смысловой структуры одного предложения и употребляются в обратной последовательности: сначала – *результат*, потом *объект*. Такое сходство смысловых структур реферата и заголовка послужило основанием для изучения взаимосвязи текстов и их заголовков для того, чтобы с помощью информации, содержащейся в заголовке, выявить в тексте те лексические единицы, которые необходимы для семантического наполнения модели реферата данного текста.

В исследовании [2] была проведена классификация лексем, принимающих участие в заполнении реферативных конструкций. В последующем была построена классификация лексем заголовка [3] и проведено ее сравнение с первой классификацией. Эти исследования позволили нам увидеть четкое продолжение процедуры компрессии в заголовке в сравнении с процедурой компрессии в реферате как на

семантическом, так и на синтаксическом уровнях. При сохранении тех же смысловых составляющих и в реферате и в заголовке они имеют вместе с тем синтаксические и грамматические особенности их выражения, которые повышают уровень обобщения смысла текста в заголовке.

Наличие одних и тех же семантических компонентов в заголовке, реферате и тексте, являющихся разными формами выражения одного и того же понятия, позволяет описать смысловые структуры словосочетаний на разных уровнях обобщения информации. А это, в свою очередь, позволило построить модель реферирования в виде процедуры перехода от заголовка к тексту и дальше к реферату в ходе его содержательного конструирования.

Для оптимизации процедуры автоматического извлечения именных групп из текста при заполнении модели реферата в нашей системе необходимо использование текстовой базы, антологий верхнего уровня и предметных областей. Антология верхнего уровня представляет собой вырожденную антологию в виде словаря категорий реферативных конструкций – *объект, результат, цель, инструмент*. Антология предметной области представляется в виде таксономии понятий конкретной области знаний. Для построения текстовой базы мы отталкивались от понятий, содержащихся в заголовке документа, по которым отыскиваются соответствующие им именные группы в тексте.

Анализ смысловой и синтаксической структур заголовка показал, что заголовок является аналогом индикативного реферата в максимально сжатом виде, что позволило нам рассматривать заголовок в качестве отправной точки в разработке системы автоматического реферирования.

Процесс сжатия является, на наш взгляд, самым трудным, поскольку предполагает свертывание смысла путем поиска наиболее емких средств и форм представления информации. Под сжатием подразумевается совокупность операций аналитико-синтетической переработки информации, преследующих цель создания вторичных документов, или выражение содержания исходного текста в более экономичной форме. При этом содержание реферата должно оставаться семантически адекватным и эквивалентным первоначальному документу.

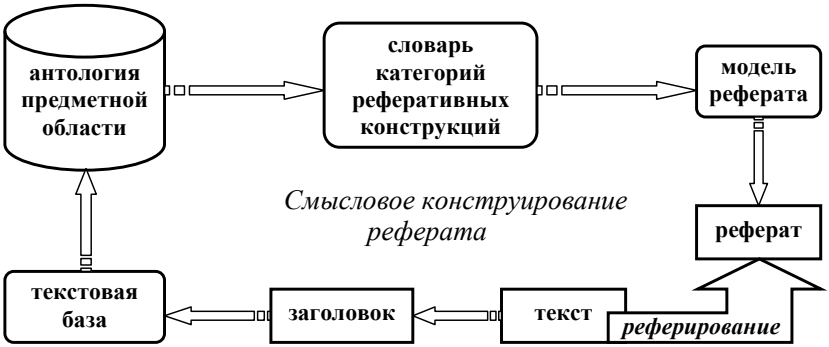

Рис. 2. Общая схема смыслового конструирования реферата

В идеале модель понимания для систем АР должна учитывать сложную иерархию всех (качественно разных) видов понимания – от языкового до психологического, и, более того, включать понимание текста каждым человеком соответственно его внутреннему миру, индивидуальному смысловому контексту. Такой подход, охватывающий все стороны изучаемого явления, пока далек от практической реализации. Вместе с тем все эти вопросы остаются в поле зрения современных исследователей. Организация и функционирование человеческих знаний является предметом когнитивной науки. Изучение того, как структуры языкового знания представляются и участвуют в переработке информации, относится к области интересов когнитивной лингвистики.

В терминах когнитивной лингвистики можно сказать, что построенная нами интенсиональная модель реферата, описывающая его концептуальное значение, позволила более четко определить те средства представления знаний в системе автоматического реферирования, которые необходимы для наполнения модели реферата экстенсиональной семантикой.

В результате мы можем констатировать, что в число средств представления знаний, необходимых для смыслового конструирования модели реферата, входят как минимум:

• *текстовая база*, являющаяся семантическим представлением текста;

• *антологии*: метаантология для описания категорий реферативных конструкций (категорий, посредством которых естественный язык осуществляет концептуальное структурирование представлений о действительности) и антологии предметных областей, содержащие понятия определенной области знаний.

Текстовая база должна включать информацию, выраженную самим текстом, т. е. референциальное значение текста, так как в отличие от художественной литературы «для такой жанровой разновидности текста, как научная и техническая литература, характерна преобладающая роль референциальных значений» [4].

Поскольку «процесс понимания предполагает частичное планирование (или ожидание)» (в нашем случае, ожидание) «структур и значений предложений и целых текстов» [5], именно с создания таких структур и следует начинать разработку текстовой базы.

Выявленное и описанное на первом этапе наших исследований сходство смысловых структур реферата и заголовка послужило основанием для изучения взаимосвязи текстов и их заголовков для того, чтобы с помощью информации, которая содержится в заголовке, выявить в тексте те лексические единицы, которые необходимы для семантического наполнения модели реферата экстенсиональной семантикой данного текста.

Для осуществления этой процедуры необходимо определить те языковые единицы, входящие в текст, в которых и заключены референциальные значения текста. Что же это за единицы? Определяясь с ответом на этот вопрос, мы придерживались мнения о выборочном подходе к знаниям, необходимым для понимания текста. «Вместо более или менее сплошной активации всего имеющегося знания, нужного для понимания слова, предложения или конструкции с глобальной темой, имеет место … стратегическое использование знания, которое зависит от целей пользователя языка, объема знания, имеющегося в тексте и контексте, уровня переработки или степени связности, необходимых для понимания и являющихся критериями стратегического использования знания…» [5].

В соответствии с концепцией понимания, предложенной в работе [5], для описания глобального содержания текста необходимо построение схемы, обеспечивающей «быстрый анализ поверхностных

структур и выстраивание относительно простой и жесткой семантической конфигурации». Наряду с этим схема позволяет «обеспечивать некоторые общие ограничения на возможные локальные и глобальные значения текстовой базы».

Второй не менее важной составляющей процесса понимания текста является описание «локальной когерентности (связности)» или установление значимых связей между предложениями текста.

Ограничив решение этих задач рассмотрением в тексте смысловых аспектов, необходимых для смыслового конструирования реферата, мы подошли к задаче построения текстовой базы.

Одним из путей в решении этой задачи является построение когнитивной или семантико-контекстной модели, которая обеспечила бы глубинное проникновение в содержание текстов и его трансформацию с сохранением смысла. Целью нашего исследования на данном этапе является изучение и описание некоторых категориальных признаков текста, а именно когезии и когерентности, которые представляют для нас интерес с точки зрения анализа текста как единого семантического и структурного целого, а также анализ содержательной и структурной связности текста на основе описанных категорий. При этом основными средствами связности называют когерентность и когезию, которые имеют первоочередное значение в рамках исследований по структуре связного текста.

Когерентность или смысловая связность проявляется во взаимозависимости отдельных предложений на содержательном уровне, что обеспечивает цельность всего текста на смысловом уровне.

В ходе исследования научных текстов мы изучили и описали два категориальных признака текста – когерентность и когезию, а также проанализировали смысловую и структурную связность научного текста. Первые результаты наших исследований показали, что содержательная связность заголовка, который выступает в качестве своего рода ключа к пониманию содержания текста, и непосредственно текста, осуществляется в основном с помощью средств лексемрекуррентности, имплицитной рекуррентности и референциальной рекуррентности. Использование того или иного типа связности определяется структурно-композиционной характеристикой текста.

В настоящее время ведется работа над построением схемы,

обеспечивающей быстрый анализ поверхностных структур текста и выстраивание относительно простой семантической конфигурации текстовой базы, ориентированной на задачи реферирования, в нашем случае, на заполнение модели реферата необходимой информацией.

Подводя промежуточный итог работы над созданием системы автоматического реферирования, можно отметить следующие полученные результаты:

1. для интеллектуализации процедуры реферирования научных текстов *разработана модель процесса компрессии* – наиболее сложного этапа реферирования, – на базе которой была *построена модель реферата*;

2. *построена модель заголовка* и проведены исследования связи модели заголовка с моделью реферата;

3. *разработана модель семантических связей текст – реферат* с использованием текстовой базы, антологий двух уровней и заголовка;

4. на базе разработанных моделей *создана первая версия системы автоматического реферирования* «АвтоРеферат», подтвердившая работоспособность предложенного подхода к созданию систем АР и наглядно продемонстрировавшая необходимость совершенствования работы системы за счет использования текстовой базы и антологий двух уровней.

Из выводов следует, что главная задача текущих исследований состоит в построении когнитивной или семантико-контекстной модели реферирования, которая обеспечила бы глубинное проникновение в содержание текстов и его трансформацию с сохранением смысла.

Список литературы

1. Ландэ Д. В. Поиск знаний в Internet. Профессиональная работа : пер. с англ. / Д. В. Ландэ. – М. : Изд. дом «Вильямс», 2005. – 272 с.

2. Лазаренко О. В. Моделювання узагальнення в системі автоматичного реферування / О. В. Лазаренко, А. А. Яковенко. – Харьков : Изд-во НУА, 2007. – 136 с.

3. Лазаренко О. В. Аналіз смислової структури заголовка як тексту з максимальним рівнем узагальнення / О. В. Лазаренко, Т. В. Попова // Проблеми семантики слова, речення та тексту : зб. наук. пр. Випуск 12 / відп. ред. Н. М. Корбозерова. – К. : КНЛУ, 2004. – С. 143–149.

4. Бархударов Л. С. Язык и перевод / Л. С. Бархударов // Вопросы общей и частной теории перевода. – М. : Междунар. отношения, 1975. – С. 71.

5. Дейк Т. А. ван Стратегии понимания связного текста / Т. А. ван Дейк, В. Кинч // Новое в зарубежной лингвистике. – 1988. – Вып. 23. Когнитивные аспекты языка.

Кирвас Виктор Андреевич – кандидат технических наук, профессор. Работает в рамках научной школы по знание-ориентированным информационным системам

В НУА работает с 2001 года. Читает дисциплины: «Компьютерные технологии и информатика», «Основы информатики».

Специализируется в области информационных систем в образовании. В настоящее время работает над теоретическими и методическими основами системы формирования информационно-коммуникационной компетентности студентов гуманитарных специальностей.

Имеет более 90 научных публикаций.

УДК 378.147.016:004

В. А. Кирвас

ПОНЯТИЙНО-ТЕРМИНОЛОГИЧЕСКИЙ АППАРАТ ПРОБЛЕМЫ ИНФОРМАЦИОННО-КОММУНИКАЦИОННОЙ ПОДГОТОВКИ ВЫПУСКНИКОВ СОВРЕМЕННОГО ГУМАНИТАРНОГО УНИВЕРСИТЕТА

В статье отражены авторские наработки по проблемам организации информационно-коммуникационной подготовки выпускников современного гуманитарного вуза. Исследованы подходы к определению категорий и понятий в области информационного образования.