УДК 811.161.1’42:004.89

А. В. Дыбина

АНАЛИЗ ПРЕДЛОЖЕНИЙ СО СЛОВАМИ-УКАЗАТЕЛЯМИ НА ОБЪЕКТ ОПИСАНИЯ ИЗ ТЕКСТОВОЙ БАЗЫ В СИСТЕМЕ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ

Резюме

У статті наводяться результати дослідження речень зі словами-покажчиками на об’єкт дослідження з текстової бази в системі автоматичного реферування. Дається перелік слів-покажчиків на об’єкт, їх семантико-синтаксичний аналіз та трансформація в реферативне речення для побудови автоматичного реферату. Всі положення ілюструються прикладами.

Summary

The article presents results of the analysis of sentences including words-indexes of the of research from the text database in the systems of automatic summarization. The list of words-indexes is given, their semantic and syntactic analysis and transformation in the summary sentence has been performed to construct an automatic summary. All statements are illustrated by examples.

Ключевые слова: автоматическое реферирование, текстовая база, семантико-синтаксический анализ, когнитология.

В 70-х годах прошлого века внимание ученых гуманитарных и социальных дисциплин было приковано к изучению связного текста, который в западной лингвистике получил название дискурса. Изначально исследователи дискурса разрабатывали методы его анализа, необходимые для полевых исследований бесписьменных языков, т. е. преобладал интерес к исследованию свободных предложений, изолированных от контекста. Вскоре в европейской и американской лингвистической науке обнаружилась тенденция в сторону контекстно-зависимого анализа текста, поскольку только контекст позволяет однозначно понять то или иное лексическое или грамматическое явление.

В социальных науках актуальным стало изучение использования языка в социальном контексте, социальной вариативности языкового употребления, речи в естественных условиях.

Результаты исследований дискурса в гуманитарных дисциплинах позволили ученым в области искусственного интеллекта обратиться к изучению проблем семантической памяти и к созданию когнитивных моделей понимания связного текста для разработки и внедрения программ автоматической обработки текстов (АОТ) в целом и систем автоматического реферирования (АР) в частности. Это позволяет говорить об *актуальности* проводимых нами исследований, поскольку системы АР, основанные на глубинном понимании смысла текста, остаются предметом дискуссий о возможности создания подобного рода систем и попыток их построения. Анализ систем АР, представленных в Интернете, свидетельствует о том, что современные системы основаны не на понимании текста, а на статистических методах, в основе которых лежит предположение о значимости слова, зависящей от частоты употребления в тексте: «значимыми предложениями считаются те, которые содержат «совокупность» значимых слов для данного документа» [1, с. 10].

В целом проблемы исследования текста охватывают широкий круг задач. Например, деятельность ученых Украинской Лингвистической Лаборатории, созданной в Киевском государственном университете им. Т. Г. Шевченко на основе кафедры математической информатики и отдела интеллектуализации информационных технологий, направлена на разработку теоретических вопросов по проблемам анализа текстов, создание систем автоматического понимания естественноязыковых объектов и др. В Институте кибернетики им. В. М. Глушкова НАН Украины сфера научных исследований представлена достаточно широко: от разработки общей теории системного анализа и математического моделирования искусственного интеллекта до решения проблем информатизации общества. Научно-исследовательский институт НАН Украины Украинский языково-информационный фонд занимается исследованиями системного строения естественного языка, разработкой и созданием информационно-лингвистических систем и многими другими вопросами.

Исходной теоретической посылкой исследования можно считать концепцию понимания текста, предложенную голландским

лингвистом, одним из основателей теории текста и анализа дискурса, Адрианусом ван Дейком (Adrianus Van Dijk), и описанную в работе [2]. Согласно концепции понимания связного текста ван Дейка, рефератом или резюме научного текста является совокупность макроструктур, которые адресат строит в процессе понимания текста. Построение макроструктур происходит в процессе применения макроправил, что позволяет переходить от текста к реферату. В нашем исследовании текстовая база (ТБ) представляет собой основу макроструктуры текста, т. е. семантическое представление научного текста.

Анализ последних публикаций в области АОТ в целом и АР в частности показал недостаточную изученность вопросов моделирования понимания текста с целью построения интеллектуальной системы АР, основанной на процессах понимания.

Постановка задачи. Цель данного исследования состоит в обобщении результатов семантико-синтаксического анализа предложений из текстовой базы, содержащей предложения со словами-указателями на один из выделенных смысловых аспектов – объект.

Объект исследования – статьи научно-технического характера и научно-гуманитарная литература. *Предмет* исследования – текстовая база предложений со словами-указателями на смысловой аспект – объект.

Новизна исследования заключается в разработке алгоритма формирования текстовой базы на основе четырех смысловых компонентов – объекта, результата, метода и цели для формирования автоматического реферата в системе АР.

Исходя из идей, изложенных в публикациях [3; 4], была сформирована текстовая база из научно-технических и научно-гуманитарных статей, содержащая предложения непосредственно из текста со словами-указателями на смысловой аспект – объект. Установлено, что стиль изложения научно-технических текстов значительно отличается от стиля научно-гуманитарных статей четкостью, точностью изложения, логичностью, объективностью, тогда как статьи научно-гуманитарного характера носят более описательный характер, предложения нередко содержат оценочные суждения. Научно-технические статьи обнаруживают более четкую

структуру, тогда как в научно-гуманитарной литературе не прослеживается строгое членение текста на смысловые блоки: введение, основная часть, заключение.

В ходе семантико-синтаксического анализа предложений из текстовой базы были изучены предикаты, указывающие на объект исследования, и классифицированы на три группы в зависимости от значения глаголов.

К **первой** группе относятся слова-указатели, которые выражаются глаголом в 3-м лице единственного или множественного числа несовершенного вида с частичкой -ся в настоящем времени, семантику которых можно определить как выражение идеи рассуждения, получения знания, или существительным с глаголом: *анализируется*, *описывается*, *изучается*, *могут изучать*, *рассматривается, обсуждаются, определяется (как), высказывается, упоминается, дается обзор, речь идет о*, *имеются в виду*, *трактуется как*. Особую разновидность такого типа предложений составляет схема, в которой предикат выражается причастием *рассматриваемый* или деепричастием *изученный*, что в реферате может быть заменено на глагол в 3-м лице: *рассматривается*, *изучается*. Семантическую схему такого типа предложений можно определить следующим образом: «отношение между субъектом и его предикативным признаком – состоянием как результатом действия» [5, § 1923, § 2367, § 2273]. В данной синтаксической структуре предложений наряду с предикатом присутствует грамматический субъект – компонент сигнификативного значения предложения, или другими словами, элемент структуры синтаксического содержания предложения. Грамматический субъект в предложениях такого типа представлен существительным в единственном или множественном числе. Однако в них отсутствует реальный субъект – денотат (референт), который находится за пределами структуры предложения. Реальным субъектом в предложениях, используемых в рефератах научных текстов, является исследователь (исследователи), ученый (ученые), теоретик (теоретики), практик (практики) и т. п. – человек или группа лиц, непосредственно проведшие исследования, описанные как в оригинальном тексте, так и в реферате. В любом случае реальный

субъект предложений в реферате – автор оригинального текста (статьи), значит, его можно не называть для краткости, что и происходит в рефератах.

После предикатов в данном случае всегда следует объект в именительном падеже, выраженный лексикой из предметной области. Объект может выражаться с помощью существительного, существительного с прилагательным, несколькими существительными. Прилагательное носит оценочный характер и является факультативным в актанте-объекте.

Употребление несовершенного вида глаголов обусловлено категориальным грамматическим значением данного вида – отсутствия указания на признак ограниченности действия пределом [6, § 1438, с. 605]. Именно потому, что действие не ограничено пределом, оно может быть представлено как протекающий процесс. Это значение несовершенного вида реализуется в конкретно-процессных типах ситуаций, в которых совершенный вид не может употребляться, поскольку не может обозначать развернутый, развивающийся процесс.

Конкретно-процессный тип употребления несовершенного вида глаголов в научном тексте также обусловлен планом настоящего времени. Предикаты в тексте имеют форму настоящего комментирующего [6, § 1504, с. 631] времени и обозначают действие, одновременное не с моментом речи, а со временем существования и восприятия той ситуации, которая описана в исходном тексте, в то время как в реферате вместо несовершенного вида предпочтение отдается употреблению глаголов совершенного вида, что обусловлено необходимостью подведения результатов исследования объекта, описанного в статье.

Морфологическая категория лица, тесно связанная с категорией времени и выражающая соотнесенность или несоотнесенность действия с участниками речевого акта, также имеет немалое значение для обобщения в процессе реферирования. Формы 3-го лица выражают отнесенность действия к неодушевленным предметам и соответственно несоотнесенность к участникам речевого акта, выступают в неопределенно-личном употреблении, соотносящем

действие с субъектом – неопределенным количеством лиц или с одним лицом. Очевидно, что здесь употребление данной категории лица обусловлено синтаксической структурой предложений.

Категория рода в настоящем времени не отражается. Приведем примеры:

«Анализируются вероятностные, а не достоверные веса множества индивидуальных систем синхронных и асинхронных связей головного мозга».

«В этой статье *описывается* метод формирования грамматических категорий, в том числе словообразовательных».

«В статье *изучаются* экспериментально проверяемые условия, при выполнении которых произвольная динамическая система, обследуемая методом нуль-органа, может быть математически описана интегралом свертки».

«Рассматриваются работы по автоматическому анализу и синтезу письменного языка в США».

«Определяется идейный потенциал, в первую очередь, состоянием общественных наук, литературы и искусства, их идейным уровнем, их силой воздействия на умы и сердца людей».

«По коренным вопросам о сущности языка и задачах языкознания у нас *высказываются* весьма различные взгляды».

«В статье «Социолингвистика» *упоминаются* методы полевого исследования языка: анкетирование, интервьюирование, наблюдение (в частности, включенное)».

«Ниже предполагается лишь *дать обзор* господствующих в настоящее время теорий и показать, на каких принципах они основаны и каких результатов можно достичь с их помощью».

«Имеются в виду следующие вопросы: 1) стратификация семантики, соотношение понятий «план содержания текста» и «смысл текста»…».

«Текст, *рассматриваемый* с точки зрения его языкового содержания и языкового выражения, выступает как билатеральная система».

Предложения в тексте могут быть распространены причастным оборотом, вводимым причастием действительного или страдатель-

ного залога настоящего времени (*обследуемая*), или придаточным предложением, которое вводится местоименным прилагательным «*который*», выступающим в роли союзного слова (*при выполнении которых*), при этом связь, устанавливаемая при помощи союзного слова, имеет характер подчинения [6, § 1684, с. 719] и всегда отсылает к референту, уже упоминавшемуся в главном предложении, т. е. ситуация, представленная в главной части предложения, предшествует ситуации, представленной в придаточной части предложения [5, § 2970, с. 549].

Союзное слово располагается непосредственно в начале придаточного предложения, сразу за существительным, которое оно замещает [5, § 2881–2882, с. 514–515], или входит в качестве зависимого компонента в состав субстантивного словосочетания, и располагается после существительного [5, § 2881, с. 514]. Однако такой порядок слов, отодвигающий местоименное слово в конец или середину придаточного предложения, не имеет общенормативного характера и стилистически не соответствует нейтральному, неэкспрессивному стилю научного изложения [5, § 2881, с. 515].

Причастие, входящее в состав обособленного оборота, формально может зависеть от актанта-объекта или той части предложения, которая является объяснением объекта исследования и находится после предиката, а семантически причастие зависит от сказуемого, называя состояние субъекта, сопутствующего совершению действия, названного глаголом-сказуемым (*понимается совокупность семантических и структурных механизмов, содействующих целенаправленному изменению плана содержания лексической единицы*; *будем понимать две построенные морфологические функции, характеризующие зависимость текстов от смысла*).

Причастные обороты и придаточные предложения, вводимые причастиями и союзными словами, базируются на анафорическом (отсылочно-заместительном) употреблении [5, § 1685, с. 720]. В реферате они могут быть эквивалентны семантике самостоятельных простых предложений. В этом случае семантика таких простых предложений будет выражать результат действия.

Фактически, функция причастного оборота или придаточного

предложения сводится к расширению семантики объекта исследования и представлению его значительно шире.

В автоматическом реферате предложения первого типа имеют незначительные изменения:

«Анализируются / описываются / рассматриваются / изучаются вероятностные, а не достоверные веса множества индивидуальных систем синхронных и асинхронных связей головного мозга».

Во **вторую** группу входят слова-указатели, выраженные кратким страдательным причастием *посвящен, нацелен (на)* + отглагольное существительное, семантику которого можно определить как действие, направленное на достижение результата. Отглагольное существительное, образованное от слов-указателей первой группы, фактически является смысловым вариантом первой группы предикатов: *посвящен* (проблемам *описания*) *описанию / анализу / рассмотрению / обсуждению / определению*; *нацелен на описание / анализ / рассмотрение / обсуждение / определение*. Приведем примеры:

«Сообщение *посвящено описанию* транслятора с этого языка».

«Деятельность полевого лингвиста обычно бывает *нацелена на описание* языка, на непосредственное изучение языковых феноменов, в ходе которого могут решаться самые разнообразные конкретные исследовательские задачи».

В автореферате данные предложения также имеют незначительные изменения:

«Сообщение *посвящено* описанию деятельности полевого лингвиста».

Третью группу слов-указателей составили те, которые прямо не указывают на объект исследования, а связаны с его определением и являются дополнением к описанию объекта исследования: *называется*, *называем*, *называют*, *понимается (как)*, *будем понимать*, *представлен*, *представляя*:

«Отображение M в двухэлементное множество *называется* конечным предикатом».

В статье «Экспериментальные методы» сказано: «Экспериментальная работа с информантами (нередко в сочетании с наблюдением)

непосредственно в среде носителей языка *называется* обычно полевой лингвистикой».

«Идейным потенциалом народа мы *называем* его относительную способность быть активной стороной в идейном обмене с другими народами».

«Коль скоро наш язык *называют* natural language, «естественным», попытаемся оценить правильность такого наименования с точки зрения латинской корреляции natura || cultura».

«Документирование языка *понимается* как «долговременная многоцелевая фиксация языковых данных».

«Под конечной алгеброй *будем понимать* алгебру с конечным числом элементов».

«Отсубстантивные имена существительные *представлены* наименованием лиц и предметов по предмету, как-то с ними связанному».

«Язык, фольклор, литература, искусство, *представляя* разные формы объективации общественного сознания, образуют единый гуманитарный мир».

В **четвертую** группу вошли слова-указатели, которые выражаются глаголом совершенного вида в настоящем времени в повелительном наклонении: *рассмотрим, сформулируем, докажем*, что в реферате может быть употреблено в форме *рассматривается, формулируется, доказывается*. Употребление совершенного вида глаголов в данном случае обусловлено их способностью передавать концентрированный, целостный факт, ограниченный пределом [6, § 1438, с. 603]:

«Рассмотрим основные понятия, существенные для анализа семантики высказывания и целостного текста».

«Сформулируем и *докажем* теорему об условиях существования сверточного семейства предикатов».

В автореферате предикаты в повелительном наклонении предлагается заменить на глаголы в прошедшем времени:

«Рассмотрены / *сформулированы* / *доказаны* основные понятия, существенные для анализа семантики высказывания и целостного текста».

Выводы. Проанализирована текстовая база предложений, в которых встречаются слова-указатели на смысловой аспект «объект», что позволило классифицировать данные слова-указатели на четыре группы на основе семантического анализа выделенных предложений.

Дальнейшие исследования направлены на анализ всех смысловых аспектов индикативного реферата – результата, цели и метода, что позволит улучшить качество рефератов, получаемых в системе автоматического реферирования.

Список литературы

1. Лазаренко О. В. Моделювання узагальнення в системі автоматичного реферування : [монографія] / О. В. Лазаренко, А. А. Яковенко. – Х. : Вид-во НУА, 2007. – 136 с.

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

3. Побудова текстової бази в системі реферування з опорою на знання / Анастасія Дибіна, Ольга Лазаренко // Людина. Комп'ютер. Комунікація : зб. наук. пр. / Нац. ун-т «Львів. політехн.», Ін-т комп'ют. наук та інформ. технологій, каф. приклад. лінгвістики. – Л., 2010. – С. 75–76.

4. Дибіна А. В. Побудова текстової бази в системі автоматичного реферування на основі структурно-семантичного аналізу тексту / Дибіна А. В., Лазаренко О. В. // Проблема семантики слова, речення та тексту / Київ. нац. лінгвіст. ун-т. – К., 2011. – Вип. 26. – С. 105–111.

5. Русская грамматика. – М. : Наука, 1980. – Т. II. – 710 с.

6. Русская грамматика. – М. : Наука, 1980. – Т. I. – 788 с.