УДК 81’322. 2’33

**О. В. Лазаренко**

ХГУ «НУА», г. Харков, Украина, lazolvlad@gmail. com

# МОДЕЛИРОВАНИЕ ПРОЦЕССА ПОНИМАНИЯ ТЕКСТА С ИСПОЛЬЗОВАНИЕМ ИНВАРИАНТНОЙ РЕПРЕЗЕНТАЦИИ СИТУАЦИЙ В СИСТЕМЕ АВТОРЕФЕРИРОВАНИЯ

Предложена процедура построения инвариантной репрезентации ситуации с использованием текстовых баз, ситуационных моделей и модели заголовка для смыслового анализа текста в системе автоматического реферирования.

АВТОМАТИЧЕСКОЕ РЕФЕРИРОВАНИЕ, ТЕКСТОВАЯ БАЗА, СИТУАЦИОННАЯ МОДЕЛЬ, ИНВАРИАНТНАЯ РЕПРЕЗЕНТАЦИЯ

## Введение

Работа над созданием искусственного интеллекта (ИИ) длится не первое десятилетие. Однако ключевая задача – научить компьютеры работать подобно человеческому мозгу – так же далека от своего решения, как и раньше. Одна из причин такого положения состоит в том, что разработчики ИИ (по мнению известного разработчика компьютеров в Силиконовой долине Джеффа Хокинса) «хотят достичь поставленной цели, обойдя вниманием вопрос о сути разума, о том, что означает слово “понимать”... Этим самым они “выплеснули с водой ребенка”— создавая мыслящие механизмы, забыли о разуме! Но все попытки создания искусственного интеллекта без учета особенностей естественного обречены на провал» [1].

Сравнивая работу человеческого мозга с работой компьютера, Дж. Хокинс задался вопросом – *какая составляющая разума отсутствует в компьютере?*

В поисках ответа на этот вопрос в августе 2002 года он открыл научно-исследовательский центр по изучению мозга, в котором первостепенное значение уделили изучению неокортекса — части головного мозга человека, ответственной за интеллект [1].

В ходе изучения неокортекса были обнаружены определенные особенности его работы, среди которых для нас представляет особый интерес способность мозга использовать инвариантные репрезентации объекта с сохранением его наиболее важных признаков на основе относительных измерений, пропорций и других характеристик, в которых возможны существенные упущения в сравнении с конкретным образом. Оказалось, что мозг запоминает важные взаимосвязи внешнего мира, а не привязывается к отдельным его элементам.

Аналогичным образом действует человеческий мозг и при чтении текстовой информации. Согласно гипотезе, выдвинутой голландским ученым А. ван Дейком, при чтении текста люди часто обрабатывают информацию не полностью или неточно и, тем не менее, понимают текст. «Языковому пользователю нет необходимости дожидаться конца абзаца, главы или целого текста, чтобы понять, о чем идет речь в тексте или в его фрагменте, ... пользователь языка может догадаться о теме текста уже после минимума текстовой информации из первых пропозиций. Догадку может подтвердить самая различная информация: заглавие, тематические слова, тематические первые предложения...» [2], то есть наиболее важные аспекты и их взаимосвязи.

В процессе работы над созданием интеллектуальной системы автоматического реферирования [3] нами были исследованы различные аспекты процесса понимания текста [4, 5]. В результате мы пришли к выводу о необходимости разработки не только ситуационных моделей, позволяющих уйти от использования таких сложных хранилищ знаний как онтологии, но и *инвариантных репрезентаций ситуаций*, представляющих собой наборы наиболее важных признаков, выделенных на основе относительных характеристик ситуаций, в которых возможны существенные упущения в сравнении с конкретной ситуацией, описываемой в конкретном тексте. Все это позволит, на наш взгляд, не только упростить автоматизацию процесса понимания текстов, но и глубже разобраться в механизмах понимания текста человеком.

Основной ***целью*** наших исследований на этом этапе стала разработка процедуры построения ситуационных моделей текстов и инвариантных репрезентаций ситуаций на базе этих моделей. Для достижения поставленной цели, опираясь на построенные ранее семантико-контекстную модель текста, модель заголовка, модель реферата и разработанную процедуру построения текстовых баз с учетом влияния каждой их них на анализ смысла текста, нами была разработана процедура выделения наиболее важных смысловых составляющих определенной ситуации, образующих инвариантную репрезентацию ситуации.

## 1. Моделирование процесса реферирования

Цель наших исследований состоит в изучении процедуры реферирования, моделировании процесса реферирования, его формализации и создании на основе разработанных формализмов системы автоматического реферирования (АР) с опорой на знания.

Работа по созданию системы автоматического реферирования предполагала исследование наиболее общих закономерностей реферирования, выраженных в конечном продукте – реферате, т. е. на начальном этапе была сознательно допущена теоретическая и эмпирическая неполнота. С тем, однако, чтобы в дальнейшем, оттолкнувшись от понимания изученных закономерностей, расширять область исследования.

В результате моделирование процесса реферирования было сведено к нескольким самостоятельным, но взаимообусловленным задачам, решение которых осуществлялось поэтапно.

1 этап. Изучение особенностей синтаксической и семантической структур реферата. Моделирование компрессии на всех уровнях. Построение модели унифицированного реферата. Разработка алгоритма наполнения модели реферата соответствующей семантикой.

2 этап. Разработка процедуры семантического анализа текста с целью сжатия его смысла в процессе реферирования. Построение семантико-контекстной модели реферирования, включающей модель заголовка и текстовую базу. Выбор средств представления знаний в системе автоматического реферирования.

3 этап. Совершенствование технологии построения текстовой базы с использованием методов когнитивной лингвистики. Разработка процедуры построения ситуационных моделей и инвариантной репрезентации ситуации, обеспечивающих глубинный анализ смысла текста.

Проведенные исследования синтаксической и семантической структуры реферативных предложений позволили построить модель реферата в виде типовых для индикативных рефератов семантико-синтаксическиих конструкций и сформулировать правила порождения реферативных предложений.

Разработанная модель была положена в основу первой версии компьютерной программы АвтоРеферат, в которой осуществлялось порождение реферативного предложения, но пока без глубинного анализа смысла первичного текста.

Анализ результатов работы программы показал правильность порождения реферативных предложений в соответствии с разработанной моделью реферирования, и, как и ожидалось, указал на смысловую неполноту этих предложений и необходимость более глубокого смыслового анализа первичного текста, что составило задачу второго этапа исследований.

Для решения этой задачи необходимо было перейти к изучению семантических отношений исходного текста и реферата и построению модели представления знаний в системе автоматического реферирования.

Нами была разработана семантико-контекстная модель, включившая в себя заголовок, являющийся концентрированным выражением смысла исходного текста; текстовую базу, являющаяся «информационным ядром» текста, содержащим информацию, зависящую от ситуации, описанной в тексте, и онтологии, содержащие независимую от ситуации информацию: онтологию верхнего уровня, онтологию общенаучной лексики и онтологии предметных областей.

По нашему мнению, анализ содержания текста должен включать анализ заголовка. Заголовки пишутся максимально сжато, лаконично, в них опущены все семантически второстепенные элементы. Следовательно, идет речь о компрессии содержания текста в максимальной степени. Поэтому мы рассматриваем заголовок как реферат минимального объема или как текст с максимальным уровнем обобщения содержания. Сравнительный анализ компрессии в заголовке и реферате позволил отследить четкое продолжение процедуры свертывания (компрессии) в заголовке в сравнении с процедурой компрессии в реферате как на семантическом, так и на синтаксическом, уровнях. При сохранении тех же смысловых составляющих в реферате и в заголовке они имеют свои синтаксические и грамматические особенности их выражения, повышающие уровень компрессии текста заголовка.

На базе выявленных аналогий была предложена модель заголовка, отражающая сходство его с рефератом, с целью более эффективного использования этой модели в процессе автоматического реферирования.

Наличие одних и тех же семантических компонентов в заголовке, реферате и тексте, являющихся разными формами выражения понятия, позволили описать смысловые структуры словосочетаний на разных уровнях компрессии информации. Однако для этого необходимо было использование онтологий предметных областей и общенаучной лексики. Это послужило основанием для включения в разрабатываемую нами систему автоматического реферирования онтологий как средства представления знаний.

В процессе анализа текста с целью выбора необходимой для построения реферата информации мы оказались перед необходимостью отбора тех предложений, которые содержат указания на объект, результат, цель и метод, т. е. смысловые аспекты, образующие смысловую структуру реферата. Выбор этих предложений представляет определенные трудности в силу разнообразия описания этих аспектов в исходном тексте.

Проанализировав ситуацию, мы пришли к выводу о необходимости поэтапного приближения к выбору необходимых предложений из текста. Поскольку предложений, указывающих на определенный смысловой аспект, может быть несколько, целесообразно выделить их из текста в текстовую базу, которая представляет

собой расширенное информационное ядро текста. Анализ заголовка, слов-указателей на смысловые аспекты и предложений из текстовой базы позволяет выбрать необходимую информацию для реферата.

В соответствии с концепцией понимания, предложенной в работах голландского лингвиста ван Дейка, для описания глобального содержания текста необходимо построение схемы, обеспечивающей «быстрый анализ поверхностных структур и выстраивание относительно простой и жесткой семантической конфигурации». Такие структуры текста (называемые макроструктурами) представляют собой обобщенное описание основного содержания дискурса, которое читатель строит в процессе понимания, и являются фактически рефератом или резюме. Предполагается, что последовательно применяя макроправила, можно построить формальный переход от исходного текста к реферату, состоящему из нескольких предложений. К числу таких правил относятся правила сокращения (несущественной информации), обобщения (двух или более однотипных пропозиций) и построения (т. е. комбинации нескольких пропозиций в одну), что, по сути, оказалось эквивалентно построенным нами моделям компрессии текста и модели реферата.

Построение макроструктур читающими – это одна из разновидностей, так называемых, стратегий понимания дискурса.

Таким образом, двигаясь от изучения реферата к вопросам анализа первичного текста и обосновав необходимость построения текстовой базы для полноценного смыслового наполнения модели реферата, мы получили важный результат в виде одного из возможных объяснений процесса понимания дискурса читателем.

Исходя из этих рассуждений и разработанных нами моделей различных этапов реферирования, мы смогли подойти к решению задачи построения семантико-контекстной модели реферирования, которая способствует глубинному проникновению в содержание текстов. Эта модель базируется на использовании текстовой базы, заголовка и онтологий нескольких видов. Подробное описание этих исследований можно найти в работах [3, 4, 5].

Для реализации глубинного анализа текста мы остановились на трех видах онтологий: онтологии предметной области, промежуточной онтологии общенаучной лексики, позволяющей связать понятия из разных предметных областей, и онтологии верхнего уровня, содержащей категории, описывающие смысловые аспекты реферата. Построение онтологий предполагало дальнейшие исследования в области понимания, что связано с изучением процессов концептуализации и категоризации в языке и что стало объектом наших дальнейших исследований.

## 2. Формирование инвариантных репрезентаций ситуаций на базе ситуационных моделей

В своих дальнейших исследованиях процесса понимания текста [6] мы пришли к выводу о том, что представление знаний в системе автоматического реферирования в виде онтологий предметных областей является не самым удачным решением для задачи реферирования текстов. В ходе разработки процедуры смыслового анализа текста с использованием текстовых баз, точнее, с построением текстовых баз при выборе предложений, описывающих главные смысловые аспекты текста, мы вышли на понятие ситуационной модели. В разрабатываемой нами системе автоматического реферирования ситуационная модель формируется в виде накопителя текстовых баз определенной тематики, автоматически извлекаемых из текста в процессе его смыслового анализа в соответствии с разработанным алгоритмом извлечения основных смысловых аспектов текста.

Полученная таким образом ситуационная модель, в свою очередь, стала основой для создания *инвариантной репрезентации ситуации*, представляющей собой набор наиболее важных признаков, выделенных на основе относительных характеристик ситуации, в которых возможны существенные упущения в сравнении с конкретной ситуацией, описываемой в конкретном тексте.

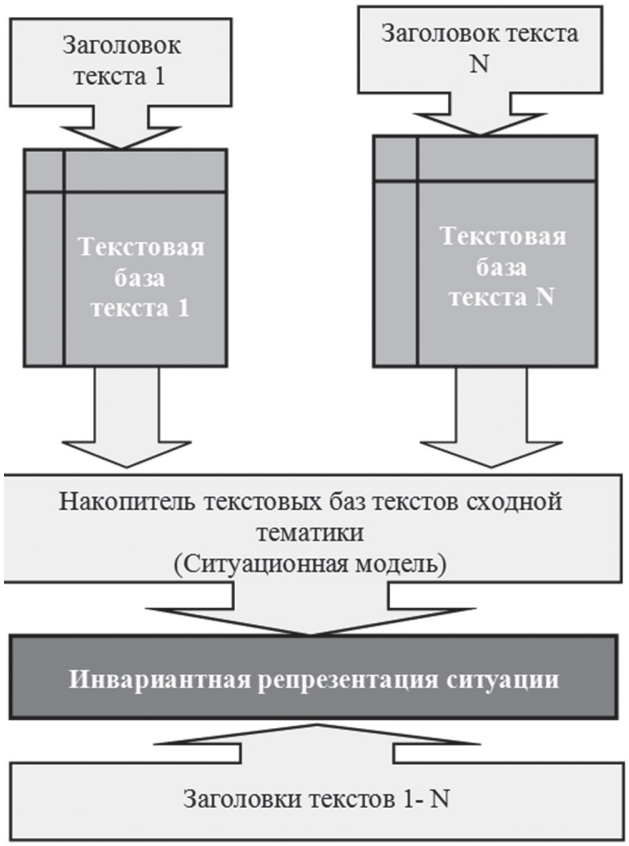


Рис. 1. Процедура разработки инвариантной репрезентации ситуации в системе автоматического реферирования

Возникает вопрос: какие признаки считать наиболее важными для описания данной ситуации, чтобы включить их в инвариантную репрезентацию? Существенную помощь в выборе таких признаков могут оказать заголовки статей как концептуальные инварианты текстов, на базе которых создавалась данная ситуационная модель.

Таким образом, с использованием текстовых баз, задающих контекстную семантику, и формируемых на их основе ситуационных моделей, содержащих информацию, актуализируемую в процессе понимания текста, а также заголовков всех текстов, описывающих схожие ситуации, можно выделить наиболее важные смысловые составляющие определенной ситуации, которые и образуют их инвариантную репрезентацию.

Может показаться, что все это уводит нас от основной задачи исследований. Однако это не так. Чем глубже и яснее мы представляем процесс понимания, осуществляемый человеком, тем точнее будет настройка системы автоматического реферирования на анализ смысла исходного текста.

Помимо этого исследование и моделирование некоторых аспектов процесса понимания приоткрывают нам тайны работы мозга. Мы видим, как работы в разных областях, таких как исследование механизмов работы мозга (Дж. Хокинс), разработка стратегий понимания дискурса (А. ван Дейк) и моделирование процесса реферирования (Лазаренко О. В.) сошлись в одной точке – инвариантных представлениях, лежащих в основе указанных процессов.

Результаты этих исследований подтверждают прямо и косвенно тот факт, что мозг в процессе распознавания объекта, фактов, ситуаций вспоминает важные взаимосвязи внешнего мира, а не привязывается к отдельным его элементам [1]. Более того, по словам того же Дж Хокинса «единственный способ, которым человек может познать этот изменчивый мир, – найти инвариантную структуру для переменного потока информации. ... В каждом конкретном случае мозг сопоставляет инвариантную структуру с текущими данными. ... Именно инвариантная форма хранится в вашем мозге, и именно с ней впоследствии сравниваются новые входные сигналы. Запоминание, припоминание и распознавание — все это происходит на уровне инвариантных форм» [1].

Таким образом, попытка разработать процедуру формирования инвариантной репрезентации ситуаций из текстов схожей тематики имеет под собой вполне реальную основу – именно в таком виде информация хранится в мозге, что позволяет ему распознавать новую информацию, содержащую в себе схожие смысловые составляющие.

**Выводы**

В статье предложена процедура смыслового анализа текста, позволяющая обеспечить более качественный результат автоматического реферирования за счет использования доступа к информации, необходимой для наполнения смысловой структуры реферата, путем:

1) выделения макроструктуры текста и формирования на ее основе текстовой базы;

2) формирования в автоматическом режиме ситуационных моделей в виде накопителей текстовых баз, используемых для более точного понимания смысла конкретного текста;

3) извлечения необходимой информации из текста через актуализацию ее с помощью инвариантных репрезентаций ситуаций.

**Список литературы: 1.** *Хокинс Дж., Блейксли С.* Об интеллекте / Дж., Хокинс, Блейксли С. – М. : Издательский дом "Вильямс", 2007. – 240 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.** *Лазаренко О. В.* Моделювання семантичних зв'язків «Текст-Реферат» в системах автоматичного реферування / О. В. Лазаренко, Д. І. Панченко. – Х. : Изд-во НУА, 2014. – 176 с. **4.** *Лазаренко О. В.* Семантіко-контекстна модель реферування / О. В. Лазаренко, Д. І. Панченко// Бионика интеллекта: науч.-техн. журнал. 2014. № 1(80). – С. 19-24. **5.** *Лазаренко О. В.* Разработка интеллектуальной системы автоматического реферирования с использованием текстовых баз и ситуационных моделей / О. В. Лазаренко // MegaLing'2013. Горизонти прикладної лінгвістики та лінгвистичних технологій : доп. міжнар. наук. конф., Україна, Київ, 20-23 листопаду 2013 г. **6.** *Лазаренко О. В.* Процедура формирования инвариантной репрезентации ситуации для автоматизации процесса понимания текста в системе автоматического реферирования О. В. Лазаренко // Інтелектуальні системи та прикладна лінгвістика : III Всеукраїнська наук. -практ. конф. – Харків, НТУ «ХПІ», 17 квітня 2014 р.

***Поступила в редколлегию 29.08.20014***

УДК 81'322. 2'33

**Моделювання процесу розуміння тексту з використанням інваріантної репрезентації ситуацій в системі автореферування** / О. В. Лазаренко // Біоніка інтелекту: наук.-техн. журнал. – 2014. – № 2 (83). – С. 15–18.

У статті розглядаються питання побудови інваріантних репрезентацій ситуацій в системі автоматичного реферування. Запропонована процедура смислового аналізу тексту на базі інваріантних репрезентацій.

Іл. 1. Бібліогр.: 6 найм.

UDC 81'322. 2'33

**Modeling of the process of understanding the text by using the invariant representation of situations in a system of auto summarization** / O. V. Lazarenko // Bionics of Intelligence: Sci. Mag. – 2014. – № 2 (83). – P. 15–18.

In article the approach to creation of the invariant representation of situations in automatic summarization systems is considered. A invariant representation based procedure for semantic analysis of text is proposed.

Fig. 1. Ref.: 6 items.