УДК 808.2: 159.937

Н. Г. ШВЕЦ

ИНФОРМАЦИОННАЯ СИСТЕМА СЕМАНТИЧЕСКОГО ПОИСКА ИЗОБРАЖЕНИЯ К ТЕКСТУ РЕКЛАМНОГО ОБЪЯВЛЕНИЯ

Минский государственный лингвистический университет

В связи со стремительным развитием информационных технологий современное общество испытывает потребность в решении проблем накопления, хранения и автоматизированной обработки семиотически неоднородных массивов текстов. Важной задачей разрабатываемых технологий является поиск оперативных способов выдачи самой разнообразной информации в соответствии с запросом пользователя. Целью предлагаемой работы является разработка компьютерной информационной системы, позволяющей к заданному тексту рекламного объявления подобрать наиболее подходящую по содержанию иллюстрацию. При этом основное содержание текста предлагается представлять в виде определенного набора главных и второстепенных ключевых слов, а содержание иллюстрации – в виде многоуровневого комплекса дескрипторов. Близость этих содержаний (и соответственно корреляцию между изображением и вербальным компонентом рекламного объявления) можно определить через максимальное число совпадений опорных слов и дескрипторов с учетом их степени важности. Важной особенностью данной системы является использование тезаурусного способа представления содержания изображения, представляющее собой описание иллюстрации в виде некоторого набора слов и отношений между этими словами в рамках некоторой предметной области. Преимуществом данного метода является то, что он позволяет учесть как доминирующие, так и второстепенные черты изображения. Анализ результатов работы компьютерной системы показал ее достаточно высокую эффективность. Таким образом, было доказано, что формальная взаимосвязь вербального текста и иллюстрации рекламного объявления вполне осуществима на лексико-семантическом уровне.

Ключевые слова: *информационная система, семантический поиск, рекламное объявление, компьютерная модель, ключевое слово, основное содержание, тезаурусное описание.*

Введение

Исследование механизмов порождения и восприятия смысла креолизованных текстов (в том числе и печатных рекламных объявлений), в которых в рамках единого сообщения сплетены вербальные и невербальные компоненты, вызывает большой интерес. Вследствие постоянного увеличения объема семиотически неоднородных массивов текстов современное общество испытывает острую потребность в их автоматизированной обработке. Способность автоматической системы осуществлять семантический поиск визуальной информации может быть использована в приложениях, способных принести конкретную пользу. Предлагаемая работа посвящена разработке компьютерной информационной системы, позволяющей к заданному тексту рекламного объявления подобрать наиболее подходящую по содержанию иллюстрацию.

Общая схема построения компьютерной модели

Конкретизируя в целях нашего исследования понятие рекламное объявление (РО), будем называть им семиотически неоднородный текст, содержащий вербальный (словесный) компонент (непосредственно рекламный текст) и визуальный (невербальный) компонент (изображение), представленный в письменной форме, заранее подготовленный, обладающий автономностью, направленный на донесение до адресата определенной информации с целью привлечения внимания к тому или иному виду товара.

Гипотеза исследования 1000 печатных РО по теме «Косметика и парфюмерия» состоит в том, что, представляя содержание текста РО в виде определенного набора главных и второстепенных опорных (или ключевых) слов, а содержание иллюстрации РО – в виде многоуровневого комплекса дескрипторов, можно

определить близость этих содержаний через максимальное число совпадений опорных слов и дескрипторов с учетом их степени важности.

Общая схема построения компьютерной модели включает следующие основные этапы [1, с. 9]:

1) постановка задачи;

2) разработка модели;

3) проведение компьютерного эксперимента.

Под моделью в компьютерной лингвистике понимается формализованное описание ряда существенных лингвистических свойств объекта, системы нескольких объектов, процесса или явления, обладающее структурным или функциональным подобием [2, с. 94; 3]. Такое описание может быть выражено конечным набором предложений какого-либо языка, математическими формулами, таблицами, графиками, специальными знаками или какими-нибудь схемами.

Рассмотрим подробнее перечисленные выше этапы разработки формальной модели.

Начнем с постановки задачи. Из 1000 проанализированных РО выбраны 200 РО, которые относятся к трем предметным областям: «Шампунь», «Крем для лица», «Краска для волос». Каждое РО состоит из двух основных частей: вербальной, содержащей текст рекламы, и изображения.

Необходимо для текста любого из упомянутых РО подобрать наиболее подходящее по содержанию изображение.

Каждая компьютерная модель опирается на некоторую базу данных (БД). В нашем исследовании она состоит из:

1) таблиц основного содержания (ТОС) исследуемых рекламных текстов;

2) формальных представлений изображений исследуемых РО, заданных в виде их тезаурусных описаний;

3) текстов РО;

4) изображений РО.

Списки опорных слов, тезаурусные описания, а также тексты и иллюстрации рекламных объявлений в БД делятся на 3 группы, соответствующие следующим:

1) шампунь;

2) крем для лица;

3) краска для волос.

В нашем исследовании при выделении ключевых слов текста учитывается абсолютная частота употребления знаменательных слов (с учетом всех их возможных синонимов и замен) и количество абзацев, в которых они встретились. В современных системах автоматической обработки текста статистические методы, как правило, дополняются другими методами [4, с. 42]. Поэтому в целях получения более качественного результата мы использовали комплексный подход: статистический метод в сочетании с позиционным методом извлечения ключевых слов из текста. Преимущество выбранной нами методики состоит в возможности классифицировать слова конкретного текста в зависимости от степени их важности для семантической структуры текста по нескольким группам, а также в гибкой применимости данной методики к текстам РО с разным количеством абзацев.

Были получены ТОС для каждого исследуемого текста, которые были дополнены ключевыми словами из заголовков (КСЗ) соответствующих РО. В ТОС опорные слова в соответствии с предметными свойствами своих референтов в общем случае могут быть разделены на следующие группы [5]:

1) слова-объекты;

2) слова-признаки;

3) слова-действия;

5) прочие слова.

Например, в ТОС (табл. 1) текста РО № 1 (рис. 1) ГОС и КСЗ разделены на:

1) слова-объекты: *шампунь*, *волосы*, *лаборатория*, *фрукты*, *концентрат*;

2) слова-признаки: *активный*, *укрепляющий*;

3) прочие слова: *сила*, *блеск*.

Таблица 1. **Основного содержания текста РО № 1**

Тип опорных слов	Опорные слова текста			
	объекты	признаки	действия	прочие
ГОС1	шампунь			
ГОС2	волосы			
ГОС3	лаборатория			
ГОС4	фрукты			
ГОС5	концентрат			
ГОС:		активный		
ГОС7				сила
ГОС8				блеск
КСЗ		укрепляющий		

В настоящем исследовании вызывает большой интерес тезаурусное представление содержания изображения, представляющее собой описание иллюстрации в виде некоторого

Рис. 1. Рекламное объявление № 1

набора слов и отношений между этими словами в рамках некоторой предметной области. Преимуществом данного метода является то, что он позволяет учесть как доминирующие, так и второстепенные черты изображения.

Методика тезаурусного описания изображений РО включала следующие этапы [6]:

1) выявление языкового, денотативного и коннотативного содержаний изображений исследуемых РО;

2) определение дескрипторов, образующих первый уровень тезауруса;

3) выделение следующих типов отношений между дескрипторами: ЦЕЛОЕ–ЧАСТЬ, СОМАТЕМА–ПРИЗНАК, СОМАТЕМА–ДЕЙСТВИЕ, ОБЪЕКТ–ПРИЗНАК, ОБЪЕКТ–ДЕЙСТВИЕ, АССОЦИАЦИЯ;

4) систематизация отношений типа СОМАТЕМА–ПРИЗНАК по ряду параметров (цвету форме и т. п). Такие параметры были выделены для описания следующих соматем: «Женщина», «Волосы», «Глаза», «Губы», «Зубы», «Кожа», «Лицо», «Нос», «Ресницы», «Тело (Фигура)»;

5) систематизация отношений типа ОБЪЕКТ – ПРИЗНАК по следующим параметрам: «форма», «цвет», «содержимое», «название», «фирма-изготовитель».

Тезаурусное представление изображения РО № 1 приведено в табл. 2.

Тезаурус является иерархической структурой, состоящей из дескрипторов различных уровней [7, с. 219; 8; 9]. Первый уровень тезау-

Таблица 2. **Тезаурусное представление изображения РО № 1**

Дескриптор (его уровень тезаурусной иерархии)	Мероним (его уровень тезаурусной иерархии)	Признак (его уровень тезаурусной иерархии)	Действие (его уровень тезаурусной иерархии) (ассоциация)
женщина (1)	голова (2) руки (2)		
флакон (1)	шампунь (2)	зеленый (2) прямоугольный (2)	
фон (1)	фрукты (2)		
схема-пояснение (1)	волос (2) вещество (2)		
голова (2)	волосы (3)		
руки (2)			завязывать волосы в узел (3) (сила волос)
шампунь (2)		fructis (фруктис) (3) garnier paris (3)	
волос (2)	корень (3)		
вещество (2)	молекулы (3)	активное (3) зеленое (3)	
фрукты (2)	дольки (3)		
волосы (3)		блестящие (4) сильные (4) густые (4) шелковистые (4) длинные (4) гладкие (4)	
корень (3)		здоровый (4) сильный (4)	
молекулы (3)		зеленые (4) красные (4)	проникать в корень (4) (сила и блеск волос)
дольки (3)		зеленые (4) желтые (4)	

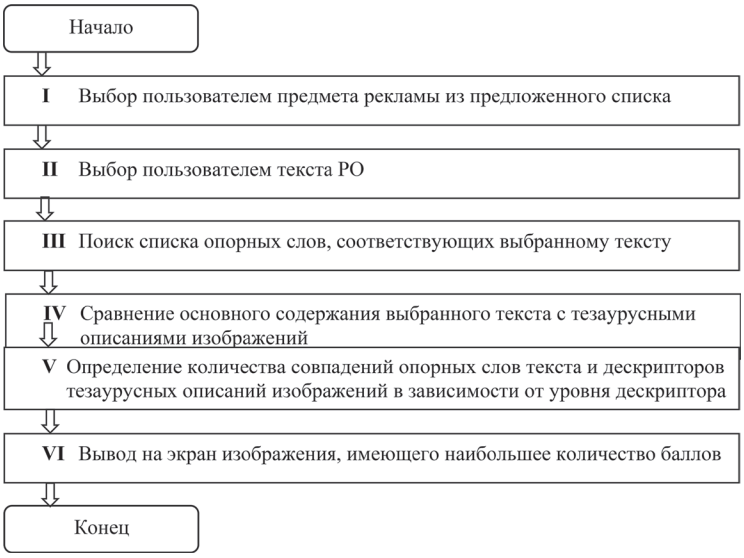


Рис. 2. Основные блоки принципиального алгоритма выбора иллюстрации к заданному тексту РО

руса иллюстрации анализируемого РО (уровень тезауруса в табл. 2 обозначается цифрой в скобках) образуют следующие дескрипторы: *женщина*, *флакон*, *фон* и *схема-пояснение*.

Между дескриптором первого уровня *женщина* и дескрипторами второго уровня *голова* и *руки* устанавливаются системные отношения типа ЦЕЛОЕ (холоним) – ЧАСТЬ (мероним).

Дескриптор первого уровня *флакон* связан с дескриптором второго уровня *шампунь* отношением ЦЕЛОЕ–ЧАСТЬ, а с дескрипторами второго уровня *зеленый* и *прямоугольный* – отношением ОБЪЕКТ–ПРИЗНАК.

Между дескриптором первого уровня *схема-пояснение* и дескрипторами второго уровня *волос* и *вещество* устанавливаются отношения типа ЦЕЛОЕ–ЧАСТЬ.

Третий уровень тезауруса образуют следующие дескрипторы: *волосы*, *завязывать волосы в узел*, *сила волос*, *Fructis (Фруктис), Garnier Paris*, *корень*, *молекулы*, *активное*, *зеленое*, *дольки*, *волосы*.

Дескрипторами четвертого уровня тезауруса являются: *блестящие*, *сильные*, *густые*, *шелковистые*, *длинные*, *гладкие*, *здоровый*, *сильный*, *зеленые*, *красные*, *желтые*, *проникать в корень*, *сила и блеск волос*.

Основным критерием максимальной близости по содержанию любого текста РО и иллюстрации, взятых из БД, является максимальное количественное совпадение ключевых слов текста с дескрипторами тезаурусного описания изображения, выбранными определенным способом.

Формальная модель процесса выбора иллюстрации к заданному тексту РО может быть представлена в виде принципиальной схемы алгоритма, основные блоки которого приведены на рис. 2.

Основной принцип работы алгоритма заключается в следующем.

Пользователь по своему желанию может выбрать на экране в меню предмет рекламы: «Крем для лица», «Шампунь» или «Краска для волос» (блок I), а затем – текст РО (блок II).

Далее компьютер в автоматическом режиме подбирает иллюстрацию к данному тексту и высвечивает ее на экране. Для этого сначала осуществляется поиск в БД списка опорных слов, соответствующих выбранному тексту (блок III), и его сравнение с тезаурусным описанием каждой иллюстрации, относящейся к выбранному предмету рекламы (блок IV). Затем определяется количество совпадений опорных слов текста и дескрипторов всех иллюстраций (блок V). В результате на экран выводится иллюстрация, которая имеет наибольшее количество баллов, которое зависит не только от количества совпадений, но и от уровня дескриптора (блок VI).

Реализация разработанной модели в виде компьютерной программы

Для проведения компьютерного эксперимента была написана программа на языке C#.

Программа работает следующим образом. Сначала из меню пользователь выбирает один из предметов рекламы (рис. 3), затем – один из

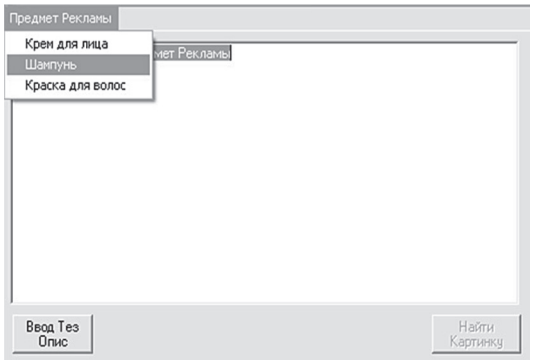


Рис. 3. Выбор из меню предмета рекламы

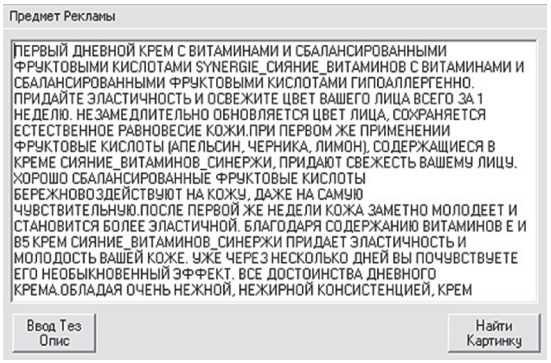
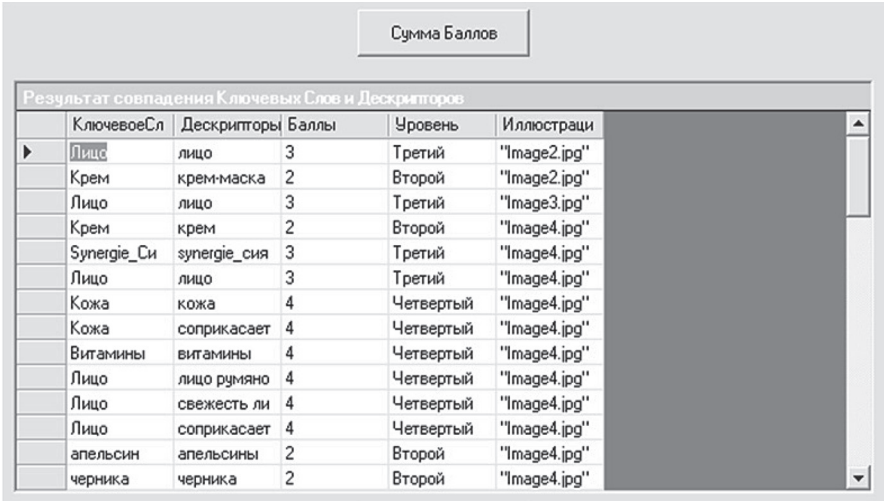


Рис. 4. Окно с выбранным текстом



КлючевоеСл	Дескрипторы	Баллы	Уровень	Иллюстраци
Лицо	лицо	3	Третий	"Image2.jpg"
Крем	крем-маска	2	Второй	"Image2.jpg"
Лицо	лицо	3	Третий	"Image3.jpg"
Крем	крем	2	Второй	"Image4.jpg"
Synergie_Си	synergie_сия	3	Третий	"Image4.jpg"
Лицо	лицо	3	Третий	"Image4.jpg"
Кожа	кожа	4	Четвертый	"Image4.jpg"
Кожа	соприкасает	4	Четвертый	"Image4.jpg"
Витамины	витамины	4	Четвертый	"Image4.jpg"
Лицо	лицо румяно	4	Четвертый	"Image4.jpg"
Лицо	свежесть ли	4	Четвертый	"Image4.jpg"
Лицо	соприкасает	4	Четвертый	"Image4.jpg"
апельсин	апельсины	2	Второй	"Image4.jpg"
черника	черника	2	Второй	"Image4.jpg"

Рис. 5. Окно с результатами совпадения опорных слов текста и дескрипторов тезаурусных описаний иллюстраций

текстов, описывающих выбранный предмет рекламы.

После нажатия на кнопку «Найти картинку» (рис. 4) на экране высвечивается таблица с результатами совпадения опорных слов и дескрипторов всех иллюстраций БД. В этой таблице перечисляются опорные слова выбранного текста РО, а также совпавшие с ними дескрипторы определенного уровня из тезаурусных описаний иллюстраций. Каждому такому совпадению начисляется определенное количество баллов, которое зависит от уровня дескриптора.

Далее при нажатии на кнопку «Сумма баллов» (рис. 5) подсчитывается количество совпадений опорных слов текста и дескрипторов, описывающих изображения РО, и на экран выдается иллюстрация, имеющая наибольшее значение числа таких совпадений.

Заключение

Анализ результатов работы компьютерной программы показал ее достаточно высокую эффективность. Таким образом, было доказано, что формальная взаимосвязь вербального текста и иллюстрации РО вполне осуществима на лексико-семантическом уровне.

Предложенные идеи можно использовать при создании информационных систем семантического поиска визуальных материалов в криминалистике (при отождествлении предметов и их описаний) и музейном деле (при составлении документации на картины и другие экспонаты), в издательском деле (при компьютерном дизайне текстов), а также в вузовских учебных курсах по лингвистике текста, семиотике и прикладной лингвистике.

ЛИТЕРАТУРА

1. Зубова, И. И. Информационные технологии в лингвистике: учеб. пособие / И. И. Зубова. – Минск: МГЛУ, 2001. – 211 с.

2. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Большакова Е. И. [и др.]. – М.: МИЭМ, 2011. – 272 с.

3. Марчук, Ю. Н. Компьютерная лингвистика: учеб. пособие / Ю. Н. Марчук. – М.: АСТ: Восток – Запад, 2007. – 317 с.

4. Карпилович, Т. П. Моделирование процесса смысловой компрессии текста: когнитивно-дискурсивный подход / Т. П. Карпилович. – МГЛУ. – Минск: МГЛУ, 2003. – 226 с.

5. Швец, Н. Г. Определение основного содержания текстов рекламных объявлений / Н. Г. Швец // Материалы ежегодной науч. конф. преподавателей и аспирантов ун-та 16–17 апреля 2002 г. в 3 ч. / Минский гос. лингв. ун-т; отв. редактор Н. П. Баранова. – Минск, 2003. – Ч. 2. – С. 82–87.

6. **Швец, Н. Г.** Формальный способ представления изображений рекламных объявлений в виде тезауруса / Н. Г. Швец // Вестник МГЛУ. Сер. 1, Филология. – 2019. – № 2(99). – С. 161– 171.

7. Баранов, А. Н. Введение в прикладную лингвистику: учеб. пособие / А. Н. Баранов. – М.: Эдиториал УРРС, 2001. – 360 с.

8. Б**раславский, П. И.** Тезаурус как средство описания систем знаний / П. И. Браславский [и др.] // НТИ. Сер. 2. – 1997. – № 11. – С. 16–22.

9. Шингарева, Е. А. О двух направлениях представления семантики текста (тезаурус и фрейм) / Е. А. Шингарева // НТИ. Сер. 2. – 1982. – № 8. – С. 1–8.

REFERENCES

1. Zubova, I. I. Informacionnye tehnologii v lingvistike: ucheb. posobie / I. I. Zubova. – Minsk: MGLU, 2001. – 211 s.

2. Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i komp'juternaja lingvistika / Bol'shakova E. I. [i dr.]. – M.: MIJeM, 2011. – 272 s.

3. Marchuk, Ju. N. Komp'juternaja lingvistika: ucheb. posobie / Ju. N. Marchuk. – M.: AST: Vostok – Zapad, 2007. – 317 s.

4. Karpilovich, T. P. Modelirovanie processa smyslovoj kompressii teksta: kognitivno-diskursivnyj podhod / T. P. Karpilovich. – MGLU. – Minsk: MGLU, 2003. – 226 s.

5. Shvec, N. G. Opredelenie osnovnogo soderzhanija tekstov reklamnyh ob#javlenij / N. G. Shvec // Materialy ezhegodnoj nauch. konf. prepodavatelej i aspirantov un-ta 16–17 aprelja 2002 g.: v 3 ch. / Minskij gos. lingv. un-t; otv. redaktor N. P. Baranova. – Minsk, 2003. – Ch. 2. – S. 82–87.

6. Shvec, N. G. Formal'nyj sposob predstavlenija izobrazhenij reklamnyh ob#javlenij v vide tezaurusa / N. G. Shvec // Vestnik MGLU. Ser. 1, Filologija. – 2019. – № 2(99). – S. 161– 171.

7. Baranov, A. N. Vvedenie v prikladnuju lingvistiku: ucheb. posobie / A. N. Baranov. – M.: Jeditorial URRS, 2001. – 360 s.

8. Braslavskij, P. I. Tezaurus kak sredstvo opisanija sistem znanij / P. I. Braslavskij [i dr.] // NTI. Ser. 2. – 1997. – № 11. – S. 16–22.

9. Shingareva, E. A. O dvuh napravlenijah predstavlenija semantiki teksta (tezaurus i frejm) / E. A. Shingareva // NTI. Ser. 2. – 1982. – № 8. – S. 1–8. E. I. Bol'shakova [et al.], Automatic natural language text processing and computer linguistics. Moscow: MIEM, 2011. – 272 p.

Поступила *28.10.2019*

После доработки *29.11.2019*

Принята к печати *01.12.2019*

SHVETS N. G.

INFORMATION SYSTEM OF SEMANTIC SEARCH OF THE IMAGE TO THE TEXT OF ADVERTISING ANNOUNCEMENT

Minsk State Linguistic University

In connection with the rapid development of information technology, modern society is in need of solving the problems of the accumulation, storage and automated processing of semiotically heterogeneous arrays of texts. An important task of the developed technologies is the search for operational methods for issuing a wide variety of information in accordance with the user's request. The aim of the proposed work is to develop a computer information system that allows you to select the most appropriate illustration for the given text of the advertisement. In this case, the main content of the text is proposed to be presented in the form of a certain set of main and secondary keywords, and the content of the illustration in the form of a multi-level complex of descriptors. The proximity of these contents (and, accordingly, the correlation between the image and the verbal component of the advertisement) can be determined through the maximum number of matches of support words and descriptors, taking into account their importance. An important feature of this system is the use of a thesaurus method of representing the image content, which is a description of the illustration in the form of a certain set of words and the relations between these words within a certain subject area.

Keywords: *information system, semantic search, advertisement, computer model, keyword, main content, thesaurus description.*

Швец Наталия Георгиевна, старший преподаватель кафедры информатики и прикладной лингвистики МГЛУ. Научные интересы: прикладная и математическая лингвистика. E-mail: shvetssnat@gmail.com.

Shvets Natalia Georgievna, senior Lecturer, Department of Informatics and Applied Linguistics, Minsk State Linguistic University. Research interests: applied and mathematical linguistics. E-mail: shvetssnat@gmail.com.