УДК 004.8 DOI: 10.30929/1995-0519.2018.1.29-37

МОДЕЛЬ И МЕТОД ТЕМАТИЧЕСКОЙ СТРУКТУРИЗАЦИИ ТЕКСТА НА ОСНОВЕ СТОХАСТИЧЕСКИХ АВТОМАТОВ

И. В. Шевченко, А. С. Лебединец, Д. О. Васильев
Кременчугский национальный университет имени Михаила Остроградского
ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: ius.shevchenko@gmail.com

Предложена модель анализа структуры текста, отличающаяся тем, что для трассировки текста используются стохастические матрицы и автоматы. Это позволяет изучать изменение значимости отдельных аспектов и сюжетных линий по длине текста, выявлять программу репрезентации словесных образов. Приведены этапы метода контент-анализа текста. Метод отличается тем, что в содержании выделяются тематические аспекты, прослеживается изменение их значимостей, формируются стохастические матрицы ассоциативных связей сущностей, и формулируются аннотации по каждому значимому аспекту содержания. Это позволяет тематически структурировать анализируемый текст. Проведена экспериментальная проверка метода и модели. Получены графики изменения значимостей аспектов и сформированы аннотации по каждому значимому аспекту. Результаты эксперимента подтверждают работоспособность метода.

Ключевые слова: анализ текста, тематическая структуризация, стохастические матрицы, аспектные аннотации.

МОДЕЛЬ І МЕТОД ТЕМАТИЧНОЇ СТРУКТУРИЗАЦІЇ ТЕКСТУ НА ОСНОВІ СТОХАСТИЧНИХ АВТОМАТІВ

І. В. Шевченко, О. С. Лебединець, Д. О. Васильев
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: ius.shevchenko@gmail.com

Запропоновано модель для аналізу структури тексту, яка характеризується використанням стохастичних матриць та автоматів для трасування тексту. Це дозволяє вивчити зміну значення деяких аспектів і сюжетних ліній за довжиною тексту, виявляти програму репрезентації словесних образів. Наведені етапи методу контент-аналізу тексту. Метод відрізняється тим, що в змісті виділяються тематичні аспекти, простежується зміна їх значимості, формуються стохастичні матриці асоціативних зв'язків сутностей і формулюються анотації по кожному значимому аспекту змісту. Це дозволяє тематично структурувати аналізований текст. Проведено експериментальну перевірку методу та моделі. Для експериментів з перевірки працездатності запропонованого методу та моделей було обрано два тексти – технічний текст обсягом 2400 слів (без заголовка, висновків і переліку посилань) і художнього тексту з обсягом 1500 слів. Ключові слова вибрані за допомогою програми TextAnalyst. У сюжетних лініях побудовані графіки зміни значущості аспектів. Висвітлені важливі аспекти та формування анотацій. Отримано графіки зміни значущості аспектів і сформовано анотації для кожного важливого аспекту. Коефіцієнт стиснення становив 10%. Результати експерименту підтверджують працездатність методу. Як перевагу запропонованого підходу можна відзначити відносну простоту виконання та повноту і навіть деяку надмірність відображення відповідної інформації в збірці анотацій. Надмірність легко усунути шляхом порогової обробки знайденого матеріалу для кожного аспекту. Оригінальна структура документа практично не має значення. Це можуть бути довільні тексти або набори значень полів гетерогенної бази даних. Недоліком запропонованого підходу на даній стадії є відсутність механізму автоматичного виявлення кореференції, тобто множинність варіантів позначення однієї й тієї ж сутності. Цей недолік частково компенсується можливістю навчання системи шляхом поповнення тезаурусів предметної області. У подальших дослідженнях планується проаналізувати тенденції у значенні аспектів, вивчити співвідношення аспектів, їх узгодженість та побудувати фазову структуру тексту. Використовуючи бази знань, можна виявити протиріччя, зіткнення та побудувати значущі інтерпретації.

Ключові слова: аналіз тексту, тематична структуризація, стохастичні матриці, аспектні анотації.

АКТУАЛЬНОСТЬ РАБОТЫ. Существующие методы анализа и моделирования текстов, используются в информационно-поисковых и информационно-аналитических системах различной направленности. При этом решаются задачи классификации документов по тематическим категориям, определения авторства, выявление плагиата, извлечения данных и знаний, автоматического реферирования и аннотирования и др. Любой связный текст является сложным системным образованием. Однако при решении конкретных задач требуется выявлять только отдельные, значимые для решения данной задачи признаки. Для выявления значимых признаков нужно рассматривать текстовые структуры как совокупность устойчивых связей элементов текста. Это особенно важно при решении задачи автоматического выявления «смысла» текста в сжатом изложении его содержания, способном донести до читателя не просто набор ключевых слов, но и нечто большее – адекватное понимание сути прочитанного. При этом нужно помнить, что содержание объективно, и поэтому его можно моделировать, смысл же всегда субъективен. Текст сам по себе не имеет смысловой структуры. Смысловая структура является принадлежностью не текста, а смысловой сферы личности, воспринимающей и осмысливающей текст [1].

Существующие подходы к анализу связных текстов можно разделить на лексико-аналитические,

использующие поэтапное применение методов морфологического, синтаксического и семантического анализа, и вероятностно-статистические – использующие традиционные методы анализа данных – описательную статистику, кластеризацию, корреляционный анализ и т.п.

Первый подход предполагает создание корпусов текстов предметных областей, множества специальных словарей и баз грамматических знаний. Несмотря на значительные успехи в этом направлении, лексико-аналитические методы остаются очень трудоемкими и сложными, что ограничивает их применение. Кроме того, применение глубокого лингвистического анализа делает результат обработки менее устойчивым к ошибкам при анализе имен собственных и технических терминов – лексических единиц, представляющих наибольший интерес при извлечении семантических отношений.

Большинство систем анализа и обработки текста в той или иной степени связаны с вычислением его вероятностно-статистических характеристик. Лидирующее место в этой категории занимает инструментарий так называемого контент-анализа – метода качественно-количественного анализа содержания документов с целью выявления различных фактов и тенденций, отраженных в этих документах.

Возможность применения количественных методов основана на вероятностном характере языка. Это подтверждается следующими фактами: дискретность единиц; массовость языковых единиц; повторяемость их в высказывании; возможность выбора определенного элемента из ряда однородных. Основной задачей статистической лингвистики является применение точных методов и использование математического аппарата для раскрытия закономерностей функционирования единиц языка в речи, а также установление закономерностей построения текста [2].

Сказанное выше говорит о том, что разработка инструментов контент-анализа, способных выявить содержательную структуру и «суть» связного текста остается актуальной задачей.

Анализ литературных данных и постановка проблемы. Все исследования последнего времени в области интеллектуального анализа текстов, компьютерной лингвистики в той или иной степени опираются на системный подход к естественному языку. Все более распространяющаяся тенденция – рассматривать связный текст или корпус текстов как некую системную целостность. Применение системного подхода оправдано, поскольку язык обладает всеми свойствами и характеристиками, присущими сложным системам. Однако он обладает и своими особенностями. Системность языка можно определить как некоторое основополагающее свойство, обусловленное его сложным составом и сложными функциональными задачами. Можно выделить следующие свойства и фундаментальные качества естественного языка:

– принципиальная нечеткость значения языковых выражений;

– динамичность языковой системы;

– образность, основанная на метафоричности;

– семантическая мощь словаря, позволяющая выражать любую информацию с помощью конечного набора элементов;

– гибкость в передаче информации.

Понятие языка связано с массовостью лингвистических элементов, описывающих ситуации внешнего мира. Именно в массовости явлений проявляются некие закономерности, для выявления которых необходим количественный анализ. Употребительность языковых элементов является проявлением их функциональной значимости в речи. Для оценки этой значимости необходимо использовать некоторую количественную меру. Из всех количественных методов наибольшие возможности для решения конкретных задач и охвата основных фактов языка имеются у вероятностно-статистического анализа. В основе использования вероятностно-статистического метода анализа лежит представление о тексте как о последовательности случайных событий, которыми являются конкретные употребления лингвистических единиц [3].

Под пониманием текста будем подразумевать процесс перевода содержания исходного текста в любую другую форму его закрепления. Это могут быть процессы: парафразы, пересказа той же мысли другими словами; перевода на другой язык; содержательной компрессии, в результате которого может образовываться минитекст, воплощающий в себе основное содержание исходного текста – реферат, аннотация, резюме, набор ключевых слов.

Считается понятным то, что может быть иначе выражено [4]. Понимание включает: репрезентацию текста по его компонентам; выведение общего содержания текста на основе непосредственно данных в нем языковых единиц и установление отношений между ними. Что касается интерпретации, то она обозначает выделение некоторого смысла, переход в восприятии текста на более глубинный уровень понимания, связанный с процедурами логического вывода и получением выводных знаний и с соотнесением языковых знаний с неязыковыми [5].

Рассмотрим теперь роль ключевых слов в восприятии и передаче «сути» текста. Исследователи обращают внимание на то, что при знании ключевых слов восприятие знаменательной лексики текста улучшается при наличии ассоциативных связей. Важным является также тот факт, что восприятию помогают не только отдельные ключевые слова, но и их последовательность. Некая последовательность ключевых слов сама по себе является семантическим предсказанием. В связи с этим исследователи обращают внимание на механизм вероятностного прогнозирования. Прогнозирование – общая особенность человеческого поведения, связанного не только с речевой деятельностью. Восприятие любого сообщения сопряжено с предугадыванием того, что последует дальше, и это, в конечном счете, способствует адекватному пониманию всего контента сообщения [6].

На уровне смыслового прогноза общеупотребительной лексики важно не угадывание конкретных слов, а выдвижение гипотез по релевантным для данного контекста понятиям.

Последовательность ключевых слов составляет смысловой стержень произведения, его структуру и отражает прямое и последовательное, накапливающееся воздействие на читателя через слова, из которых автор создает свое произведение. Отношения ключевых слов в последовательности своего тематического ряда с ключевыми словами других тематических рядов составляет смысловую композицию произведения, его смысловой образный ритм, ход мысли в процессе отыскивания свойств вещи, смысловой образ которой конструируется. Значение структуры смысловой композиции в том, что она показывает смысловую связь, поскольку слова можно извлечь из текста, но саму связь – нет. Она является системным свойством данного текста.

В работе [7] отмечается, что для выявления смысловой структуры произведения словесности необходимо выделить строевые, ключевые слова произведения, определяющие словесное ядро образа в каждом фрагменте текста, составляющем относительно отдельную образную или тематическую структуру.

Таким образом, ключевые слова или словообразы образуют динамическую и разветвленную систему. Для читателя они служат вектором текстового и подтекстового развития авторской мысли, играя важнейшую роль в формировании всех выделяемых видов подтекста [8]. «Если считается, что смысл не «конструируется» в процессе понимания, а лишь приписывается, то следует признать, что в памяти должен хранится полный набор готовых смыслов и задача заключается лишь в том, чтобы актуализировать соответствующий данному тексту смысл» [9].

Из изложенного выше следует:

– автор текста решает прямую задачу, структурируя информацию в некотором пространстве словесных образов. Читатель решает обратную, некорректную задачу, интерпретируя текст автора. Естественно, что смысл, вкладываемый автором в текст, может быть интерпретирован бесконечным числом вариаций, отличающихся как объемом, так и извлечённым смыслом. Таким образом, текст является носителем множества интерпретируемых образов, а читатель – системой распознавания словесных образов. От того, насколько система распознавания научена (настроена) на определённую комбинацию образов, насколько велика разделительная способность системы и каким образом распознанные образы будут интегрированы в единую структуру, зависит, в конечном счете, успешность передачи исходного смысла текста читателю.

– системный подход к тексту должен предполагать его декомпозицию – в первую очередь, на тематические аспекты, символами которых выступают ключевые слова (КС).

– смысловое содержание скрыто в архитектуре текста – его структуре и динамических свойствах. В совокупности КС должны образовывать логическую структуру текста.

Под динамикой текста следует понимать чередование, частоту переходов и близость появления в нём ключевых слов, отражающих в своей совокупности основное содержание текста.

Проблема состоит в том, что в настоящее время практически нет инструментальных средств, прямо направленных на выявление указанных выше динамических свойств текста и использования этих свойств для тематической структуризации текста.

Цель работы – усовершенствовать метод автоматического реферирования текста путем тематической структуризации с применением вероятностных автоматов как инструмента трассировки текста и выявления динамики чередования и частоты ключевых слов по аспектам.

МАТЕРИАЛ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ. На концептуальном уровне становится ясно, что от читателя – человека или программы – ожидается деятельность по развертыванию неких смыслопорождающих моделей текста. Понимание всегда происходит на основе имеющихся знаний. Читатель должен соотнести содержание прочитанного с собственным опытом, то есть проявить рефлексию. Это позволяет высказать предположение о том, что при обработке текста необходимо генерировать «промежуточные репрезентации» и использовать их для тематической структуризации текста. При этом, после прочтения все новых фрагментов текста, в сознании читателя происходит интерактивный процесс перестройки промежуточной информационной структуры в некоторую конечную структуру, отражающую полное «осмысление» информации в целом.

Модельное представление процесса обработки текста. Для решения задачи тематической структуризации текста требуется разработать комплекс моделей, при помощи которых решается задача построения наиболее адекватной структуры системы словесных образов, отражающей смысл, заложенный в текст автором. Очевидно, что любое предложение и любой абзац могут быть по разному интерпретированы с учётом онтологии предметной области (ПрО). В рамках онтологии ПрО могут проявляться разные аспекты, влияющие на проявление определенных словесных образов. С учётом этого представим онтологию в виде [10]:

$$O = <E(AT), ER, EA, F, AS, AR> , \qquad (1)$$

где E – набор сущностей, отождествленных с множеством образов и выражаемых ключевыми словами, AT – множество атрибутов сущностей; $ER \subseteq E \times E$ – множество отношений сущностей, $EA \subseteq E \times AS$ – проекция сущностей на аспекты, $F: E \times ER$– функции интерпретации отношений и сущностей: AS – множество аспектов, определяющих локальные смыслы фрагментов текста, $AR \subseteq AS \times AS$ – пересечение аспектов. Тогда многоаспектное семантическое пространство можно выразить набором

$$SS = <O, M>, \quad (2)$$

где M – набор метрик для вычисления степени близости сущностей и релевантности результатов поиска. Пространство SS условно разделено на пересекающиеся подпространства, каждое из которых соответствует одному аспекту. Соответственно, за каждым аспектом закрепим определенный смысл, выражающийся в комплекте ключевых слов и фраз и их последовательности.

Пусть T – входной текст, в котором представлены словоформы $X = \{x_1, \dots, x_m\}$ имеющие между собой скрытые отношения $R=\{r_1, r_2, \dots, r_v\}$. Задача тематической структуризации может быть решена при использовании некой функции структуризации $\varphi: X \to TI$, где TI – суперпозиция отношений R, выраженная на ЕЯ.

На системном уровне построим структурную матрицу взаимосвязей исходного текста, набора моделей преобразования и полученных словесных образов для множества аспектов (табл. 1). Зафиксированные текстовые последовательности – X_i , $i = \overline{1,m}$ характеризуют содержание фрагментов связного текста. Модель преобразования должна обрабатывать фрагмент текста, фиксировать динамические изменения содержания по выбранному аспекту и прогнозировать вероятности сочетания аспектных сюжетных линий в виде дискретных распределений.

Таблица 1 – Структурная матрица взаимосвязей фрагментов текста, моделей преобразования и сжатых словесных образов для одного аспекта

Входы	Модели	Выходы					
		F_1	F_2	..	F_j	..	F_n
X_1	M_1	Y_{11}	Y_{12}	..	Y_{1j}	..	Y_{1n}
X_2	M_2	Y_{21}	Y_{22}	..	Y_{2j}	..	Y_{2n}
..	..	..	..	..	..	..	..
X_i	M_i	Y_{i1}	Y_{i2}	..	Y_{ij}	..	Y_{in}
..	..	..	..	..	..	..	..
X_m	M_m	Y_{m1}	Y_{m2}	..	Y_{mj}	..	Y_{mn}

Выходы системы F_j , $j = \overline{1,n}$ – значимости j-х аспектов, полученные как суперпозиции дискретных распределений вероятностей появления определенных КС при определенном содержании фрагмента X. Связи между данными X_i и выходами F_j осуществляют вероятностные модели. Каждая модель M_i , описывающая связь текста X_i и выходной величины Y_j , представима отображением:

$$M_i : X_i \to Y_i , \quad (3)$$

где Y_i – дискретное распределение вероятностей переходов между КС. В общем случае каждая модель M_i может быть представлена матрицей, показанной в табл. 2. Здесь x_l^i – элемент текста X_i ; Y_k^i – k-й элемент дискретного распределения вероятности; I_{lk} – информационное (смысловое) содержание, отражающее связь элемента x_l^i с вероятностью перехода к k-му КС.

Таблица 2 – Структурная матрица модели M_i

Входы	Выход Y					
	Y_1^i	Y_2^i	..	Y_k^i	..	Y_p^i
x_1^i	I_{11}	I_{12}	..	I_{1k}	..	I_{1p}
x_2^i	I_{21}	I_{22}	..	I_{2k}	..	I_{2p}
..	..	..	..	..	..	..
x_l^i	I_{l1}	I_{l2}	..	I_{lk}	..	I_{lp}
..	..	..	..	..	..	..
x_r^i	I_{r1}	I_{r2}	..	I_{rk}	..	I_{rp}

Предложенная обобщенная структура взаимосвязей комплекса моделей обработки текста является концептуальной основой для построения системы динамического анализа текста, на основе которого строится тематическая структура. Этот подход позволяет дополнять и уточнять модельное описание процесса обработки текста без качественного изменения структуры системы, применять различные математические и феноменологические модели для описания различных языковых явлений, использовать различные уровни детализации для исследования внутритекстовых процессов и свойств с гарантированным сохранением общей целостности, масштабируемости и прозрачности.

От статического описания перейдем к описанию динамических свойств модели преобразования M. Как было сказано выше, для анализа динамики смысловых оттенков и сюжетных линий должна производиться трассировка текста на предмет обнаружения ключевых слов и интенсивности их чередования. Роль трассировочной функции выполняет вероятностный автомат, работающий в скользящем окне с заданной шириной. При этом создается последовательность стохастических матриц, содержащая локальные репрезентации содержания «прочитанного» текста. Именно вероятностный автомат является динамической частью преобразовательной модели, показанной в выражении (3).

Структура вероятностного автомата показана в выражении (4) [11]:

$$A = < X, Y, Q, P >, \quad (4)$$

где X – множество входных сигналов, Y – множество выходных сигналов, Q – множество состояний, P – функция вероятностных переходов:

$P = X \times Q \to Q' \to R^+$, такая, что для любой тройки $<x, q, q'>$ выполняются свойства: $p(x, q, q') \geq 0$,

$$\sum_{(x,q) \in X \times Q} p(x,q,q') = 1 .$$

Алгоритм работы вероятностного автомата стандартный и приводится в [11]. Входными сигналами автомата являются внутренние события, а именно – достижение первого и последнего КС в окне заданного фрагмента текста. Счетчик событий инициализируется при определении объема текста и количества окон анализа. Выходными сигналами автомата являются промежуточные и финальные матрицы переходных вероятностей. Именно результаты расчета вероятностных распределений используются для подготовки процесса понимания «прочитанного»

текста. Здесь учитываются интенсивности переходов между КС, которые отражают интенсивность ассоциаций как между КС, так и между сюжетными линями в тексте. Этим наш подход отличается от подхода, в котором учитываются лишь относительные частоты вхождения КС в текст.

Помимо вероятностного автомата модель M включает в себя и функцию тематической структуризации текста φ. Отождествим аспекты с понятием «локального смысла» как некоторой интерпретации текста, выраженной последовательностью сущностей-КС. Концептуально основу модели составляют наборы КС, распределенные по «аспектам-смыслам» (матрица $EA \subseteq E \times AS$). Список аспектов, вообще говоря, может охватывать любые мыслимые темы. Каждому аспекту сопоставлено отдельное ключевое слово, обозначающее «смысл аспекта», определенный экспертным путём. Имеется «локальный смысл» абзаца (или нескольких абзацев в пределах окна анализа) и «глобальный смысл» всего текста. Суперпозиция «локальных смыслов» абзацев должна дать приближение к «глобальному смыслу» текста. Задача тематической структуризации – найти некое оптимальное выражение «локальных смыслов», учесть их чередование и локальную значимость в отдельных фрагментах текста.

Каждому фрагменту текста (окну) X_w, $w=1..W$, в котором происходит локальная обработка, сопоставляются подмножества КС (сущностей E_w) и соответствующих аспектов AS_w. С помощью стохастических матриц, содержащих относительные частоты (вероятности) переходов от E_{wk} к E_{wt}, и матрицы связи EA, вычисляются степени значимости аспектов для данного окна текста X_w. С учетом сказанного раскроем содержание функции φ:

$$\varphi: X_w \times E_w \times AS_w \rightarrow SV_w, \qquad (5)$$

где SV_w – вектор значимости аспектов (significance vector) для текущего окна X_w. Вычисление степени значимости аспектов имеет важное значение для определения смысла. «Именно значимость аспекта выводит на уровень понимания текста, что, собственно, и является целью создания текста» [12].

Метод анализа текста. Общая последовательность действий по обработке текста с описанными выше моделями состоит из нескольких этапов.

Этап 1. Подготовка тезаурусов по аспектам.

От количества аспектов и качества тезаурусов зависит и качество результатов анализа. Текст должен проецироваться на знания системы, а знания сосредоточены в словарях. Среди аспектов можно выделить предметные, содержащие возможные ключевые слова заданной предметной области (ПрО) или нескольких областей, и утилитарные, содержащие образы действия или оценки в виде лексем типа «поставить задачу», «цель работы», «хороший».

Этап 2. Предварительная обработка анализируемого текста.

2.1. Преобразование всех символов к верхнему или нижнему регистру.

2.2. Удаление «стоп-слов». Под «стоп-словами» понимаем слова, не оказывающие влияния на тематику документа, например, артикли, союзы и предлоги.

2.3. Выделение предложений и абзацев (графематический анализ).

2.4. Стемминг. Данная процедура заключается в выделении значимой части слова с помощью отсечения суффиксов и окончаний.

Этап 3. Определение тематики текста. Это делается путем выявления слов с высокой значимостью по Зипфу и проверки этих слов в тезаурусах ПрО. Происходит выбор ПрО по критерию максимального числа совпадений в тезаурусах.

Этап 4. Выделение и группировка аспектных терминов. Каждое предполагаемое КС употребляется в тексте с определенной частотой. Соберем КС в матрицу $E \times AS$, в которой строки – это КС, а столбцы – аспекты. Группируем КС по аспектам, используя пороговое преобразование

$$H = \begin{cases} 1 \; if \; F \geq T \\ 0 \; \text{otherwise} \end{cases}, \qquad (6)$$

где F – экспертная оценка значимости КС в аспекте; T – заданный порог.

Этап 5. Формирование стохастической матрицы SMT для всего текста. Каждый элемент k-й строки SMT рассчитывается по формуле:

$$f_{kt} = \sum_{t=1}^{p} \alpha_{kt}, \qquad (7)$$

где $\alpha_{kt} = 1/(d_{kt}+1)$, d – расстояние между предложениями, в которых обнаружены ключевые слова-сущности E_k и E_t.

После заполнения строк матрицы, значения в каждой строке нормируются по условию $\sum f_t = 1$.

Этап 6. Определение величины значимости аспектов в тексте. С помощью стохастического автомата производится расчет финальных вероятностей перехода $E_k \rightarrow E_t$. Затем определяется значимость аспектов. При этом следует учесть, что встречаемые в тексте КС в разной степени относятся к разным аспектам (аспекты пересекаются). То есть, несколько КС «работают» на один и тот же аспект, но с разной силой (весом). Поэтому, в качестве меры значимости j-го аспекта необходимо вычислить для каждого аспекта скалярное произведение:

$$S_j = \sum_{k=1}^{p} a_{kj} f_{fin}^{k}, \qquad (8)$$

где f^{k}_{fin} – финальная вероятность переходов от других КС к k-му КС, a_{kj} – весовой коэффициент, учитывающий важность k-го КС в j-м аспекте: $a_{kj} = 1 - \frac{n_k}{n}$, где n_k – число аспектов, в которые входит k-е КС, n – общее число аспектов. Результат – вектор SV значимости аспектов по всему тексту.

Совокупность выделенных по значимости аспек-

тов и соответствующих компонентов текста позволяет установить основные опорные вехи текстового содержания и выстроить исходный замысел текста, а также определить, как соотносятся выделяемые компоненты по степени важности.

Этап 7. Оконная обработка текста.

7.1. Текст размечается границами окон – фрагментов, объемом 80...200 слов. Границы фрагментов совпадают с границами абзацев. Окно анализа скользит по абзацам с перекрытием в один-два абзаца.

7.2. Формируется стохастическая матрица окна и производится расчет финальных вероятностей для найденных в окне КС. По методике этапа 6 определяется значимость каждого аспекта в данном окне. Для текущего окна формируется вектор SV_w.

Результат: таблица *ST* (significance table), в которой каждая строка – вектор SV_w окна, каждый столбец – значимость одного аспекта.

Этап 8. Построение графиков изменения значимости каждого аспекта вдоль текста. Абзацы могут иметь разную доминирующую аспектную тематику. При смене аспекта изменится и состав КС, соответственно, на графике будут наблюдаться скачки значимостей.

При ранжировании оценок значимости можно зафиксировать картину смены доминантных аспектов. Если аспекты привязаны к сюжетным линями текста, то по графикам можно отслеживать, как переплетаются сюжетные линии. В работе мы придерживаемся гипотезы А. И. Новикова о том, что признаком смысла может служить явление, похожее на явление доминантности, т.е. выявленные доминантные аспекты несут определенный смысл, связанный с содержанием текста.

Этап 9. Сравнение с матрицами глобальной значимости. Если в локальных матрицах и глобальной матрице доминируют одни и те же аспекты, эти аспекты признаются глобально доминирующими.

Этап 10. Формирование аннотаций по аспектам.

10.1. Для каждого отобранного аспекта AS_g, g=1..G, выбираются окна X^g, в которых значимость аспекта превышает его среднюю значимость. Образуется подмножество $AS_g \times X^g$.

10.2. КС данного аспекта попарно анализируются на степень ассоциации в выбранном окне X^g. Для выбора используются стохастические матрицы, сформированные на этапах 6 и 7. Отбираются КС, для которых выполняется условие: $\alpha_{kt} \geq CT$., где *CT* – пороговое значение, влияющее на степень сжатия текста (compression threshold).

Данный алгоритм выполняется для всех, выбранных по аспекту, окон. Предложения текста, в которых оказались выбранные КС, помечаются для выборки. Результат – массив номеров предложений – *APN*.

10.3. В аннотацию включаются предложения, попавшие в массив *APN*.

Этап 11. Формируется вторичный текст, состоящий из названий аспектов и аннотаций по данному аспекту. Формирование совокупности аннотаций по аспектам производится следующим образом:

11.1. Пользователь выбирает интересующие его аспекты. Если выбор аспектов происходит автоматически, и для *j*-го аспекта выполняется условие

$$CV(ASj) \leq 0{,}25, \qquad (9)$$

где *CV* – коэффициент вариации, то по данному аспекту аннотация не формируется. Условие означает, что тематика аспекта распределена равномерно по тексту, и информация по данному аспекту будет представлена в аннотациях по другим аспектам.

11.2. Для каждого выбранного аспекта определяется временная точка w_j (номер окна) пересечения уровня значимости с линией среднего. Точки ранжируются по возрастанию значения w_j. Соответственно, очередность появления значений номера аспекта *j* будет определять очередность размещения в сжатом тексте аннотаций по *j*-м аспектам.

Содержание всех аннотаций в совокупности может оказаться избыточным. Но в данном случает избыточность (полнота) предпочтительнее потери информации при рассмотрении реальной проблемы в разных аспектах. Степень сжатия содержания текста легко регулировать, изменяя порог *CT*.

Эксперименты и обсуждение результатов. Для экспериментов по проверке работоспособности предлагаемых метода и моделей были выбраны два текста – технический, объёмом 2400 слов (без заголовка, выводов и списка литературы), и художественный, объёмом 1500 слов.

На подготовительном этапе, при помощи программы TextAnalyst из текстов выбирались тематические аспекты и наборы ключевых слов.

Для статьи «Имитационная модель функционирования карьерного автотранспорта», аспекты и ключевые слова представлены следующим образом:

Аспект 1. «Автосамосвалы»: Порожний / Загруженный / Движение / Скорость / Погрузка / Разгрузка / Время / Очередь / Дороги / Маневры / Простои.

Аспект 2. «Экскаваторы»: Автомобили / Закрепление / Задержки / Простои / Погрузка / Очередь / Карьер / Технология / Поломка / Емкость ковша.

Аспект 3. «Маршрутизация»: Скорость движения / Диспетчер / Управление / Загруженный / Порожний / Карьеры / Транспорт / Время / Участки пути / Адресация.

Аспект 4. «Имитационная модель»: Структура / Алгоритм / Управление / Процесс / Выборка / Моделирование / Статистика / Реализация / Время / Результаты.

На рис. 1 показано изменение значимости аспектов А1, А2, А3, А4.

Видно, что аспект «Автосамосвалы» хотя и доминирует по всему тексту, но распределение его значимости практически равномерно. Значимость других аспектов изменяется довольно резко.

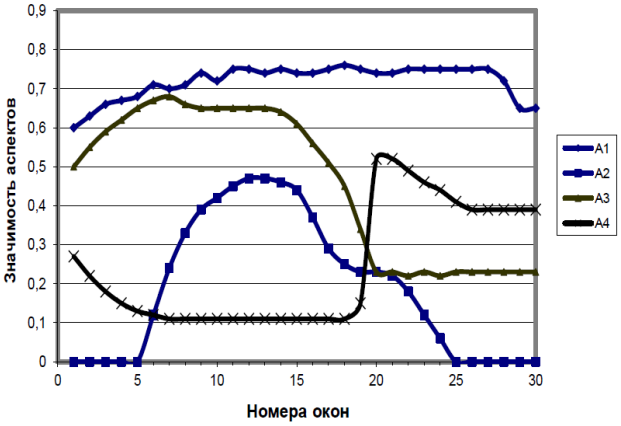

Рисунок 1 – Графики изменения значимости аспектов по тексту статьи

Соответственно, были сформированы аннотации по трем аспектам:

<*Маршрутизация*> <Множество технологически возможных маршрутов движения автосамосвалов в карьере y(k,j) от k-го пункта погрузки к j-му пункту разгрузки можно выразить через совокупность последовательностей вершин графа

$$Bm = \{Zm_1, Zm_2, \ldots Zm_{b-1}, Zm_b\},$$

где b – число вершин на маршруте m.

Таким образом

$$Zm_1 \in P,\ Zm_b \in R \text{ и } \{Zm_2..Zm_{b-1},\} \in \{Q \cup D\},$$

где P – множество пунктов погрузки, R – множество пунктов разгрузки; Q – множество пунктов промежуточных транспортных задержек и D – пункт управления.

Каждая дуга K_{kj} графа соответствует элементарному участку транспортной сети карьера и характеризуется протяженностью L_{ki} и уклоном U_{ij}.>

<*Экскаваторы*> <При сравнительно небольшом числе автосамосвалов, работающих в карьере (15–20 шт.), они закрепляются на смену за экскаваторами. При такой системе организации работ, носящей название закрытого цикла, могут возникать неизбежные простои экскаваторов или автосамосвалов вследствие неисправности одного из этих механизмов или задержки автосамосвалов. При большем числе автосамосвалов и экскаваторов возникают простои экскаваторов из-за недостатка автосамосвалов, вызванного поломками, задержками в пути и при разгрузке. Для более полного использования экскаваторов и автомобилей на многих предприятиях применяется распределение автомобилей под экскаваторы в процессе работы.>

<*Имитационная модель*> <Описываемая в данной статье имитационная модель интегрируется с системой контроля и маршрутизации роботизированных и традиционных транспортных средств, выполняющих маневры на разгрузочных и перегрузочных площадках. Разработанная имитационная модель базируется на принципах: агрегативности, т.е. горно-транспортный комплекс рассматривается как совокупность относительно обособленных агрегатов (самосвалов, экскаваторов, пунктов контроля и управления) с четко определенными функциями, событийности – состояние системы рассматривается моделью только в «критические» моменты времени, когда в системе происходят те или иные изменения. Для получения достоверных данных при использовании предложенной имитационной модели необходимо осуществлять многократный прогон программы на ЭВМ>.

Эти три аспекта выбраны программой для формирования сжатого реферативного текста. Как видно, аспектные аннотации в совокупности дают практически полное представление о содержании статьи. Информация структурирована по тематике аспектов. Степень сжатия составила 10 %, что вполне допустимо для реферата.

Приведем и аннотацию по аспекту <Автосамосвалы>, предложения в которую отбирались при значения порога *CT*=0,72:

<*Автосамосвалы*> <Транспортная сеть в карьере является сложной системой, состоящей из значительного числа активных элементов – автосамосвалов и экскаваторов. В транспортном цикле наибольший удельный вес составляет время движения автосамосвала в груженом и порожнем состоянии. Продолжительность погрузки автосамосвала зависит от модели и вместимости ковша экскаватора, характера разрабатываемого грунта и параметров забоя, схемы подъезда под погрузку, грузоподъемности автосамосвала, угла поворота экскаватора при погрузке и других факторов.>

Анализ показывает, что данная аннотация привносит сравнительно мало новой информации в реферат.

Анализ художественного текста (рассказа Рея Бредбери «Всё лето в один день») в программе TextAnalyst показал, что выбранные программой сущности (КС) и их связи не дают представления о смысле текста, а отражают (неполно) лишь информационную картину происходящего. Попытка ручного подбора КС привела к тому, что выбирающий ориентировался на те сущности и связи, которые в тексте поданы имплицитно, в частности, на слова, несущие эмоциональную окраску, т.е. использовал собственное представление о смысле текста. Понятно, для выявления эмоциональной составляющей следует формировать соответствующую базу знаний и тезаурусы. Сделан вывод, повторяющий утверждения в работах [1–5, 12], что сам текст является кодом, а смысл текста, вложенный автором, открывается лишь в той степени, в какой к его восприятию (репрезентации) подготовлен читатель.

В целом можно сказать, что смысловой ритм текста – динамика значимостей тематических аспектов (сюжетных линий) – отражает программу репрезентации образов. В дальнейших исследованиях планируется провести анализ трендов значимостей аспектов, изучить корреляцию аспектов, их согласованность, формализовать и построить фазовую структуру текста. При использовании соответствующих баз знаний и тезаурусов можно выявлять противоречия, коллизии, строить смысловые интерпретации.

В качестве достоинства предлагаемого подхода можно отметить относительную простоту реализации и стопроцентную полноту и даже некоторую избыточность отображения релевантной информации в сборнике аннотаций. Избыточность легко устраняется путем пороговой обработки найденного материала по каждому аспекту. Структура документа при этом не имеет практически никакого значения. Это могут быть произвольные тексты или совокупности значений полей гетерогенной базы данных.

Недостатком предлагаемого подхода на данном этапе является отсутствие механизма автоматического выявления кореференции, т.е. множественности вариантов обозначения одной и той же сущности. Этот недостаток частично компенсируется возможностью непрерывного обучения системы путем пополнения тезаурусов предметной области.

ВЫВОДЫ. Предложена математическая модель анализа динамики текста, отличающаяся тем, что для трассировки текста используются стохастические матрицы и автоматы, что позволяет изучать изменение значимости отдельных аспектов и сюжетных линий по длине текста, выявлять программу репрезентации словесных образов.

Сформулированы этапы метода контент-анализа текста, отличающего тем, что в содержании выделяются тематические аспекты, прослеживается их динамика, формируются стохастические матрицы ассоциативных связей сущностей, и формулируются аннотации по каждому значимому аспекту содержания, что позволяет тематически структурировать анализируемый текст. При наличии соответствующих тематических тезаурусов это позволяет максимально приблизиться к сути текста, а в перспективе – и к смысловой интерпретации.

Проведена экспериментальная проверка метода и модели. Получены графики изменения значимостей аспектов и сформированы аннотации по каждому выявленному аспекту. Результаты эксперимента подтверждают работоспособность метода.

В дальнейших исследованиях планируется провести анализ трендов значимостей аспектов, изучить корреляцию аспектов, построить фазовую структуру текста. При использовании соответствующих баз знаний можно выявлять противоречия, коллизии, строить смысловые интерпретации.

ЛИТЕРАТУРА

1. Новиков А.И. Текст и его смысловые доминанты / под ред. Н.В. Васильевой, Н.М. Нестеровой, Н.П. Пешковой. М.: Институт языкознания РАН, 2007. 224 с.

2. Суркова А.С. Идентификация текстов на основе информационных портретов. *Вестник Нижегородского университета им. Н. И. Лобачевского*. 2014. № 3 (1). С. 145–149.

3. Bolshakov I.A., Gelbukh A. Computational linguistics : models, resources, applications. Mexico: First edition, Universidad Nacional Autónoma de Mèxico, 2004. 186 p.

4. Леонтьев А.А. Основы психолингвистики. М.: Смысл, 2005. 288 с.

5. Антипов А.Г., Аносова К.М. Акты номинативной деривации в аспекте познавательной деятельности индивида (на материале диалектного словообразования). *Филологические науки. Вопросы теории и практики*. Тамбов: Грамота, 2008. № 1 (1). Часть I. С. 8–11.

6. Матвеева Л.Ю. Вероятностное прогнозирование звучащей речи: к постановке проблемы (обзор). *Саратовский научно-медицинский журнал*. 2015. № 11 (2). С. 216–220.

7. Субботина М.В. Дистрибутивная метафора как метод исследования произведения словесности. *Научный вестник Воронежского государственного архитектурно-строительного университета. Современные лингвистические и методико-дидактические исследования*. 2016. Вып. 1 (29). С. 56–67.

8. Лелис Е.И. Подтекст как лингвоэстетическая категория в прозе А.П. Чехова. Ижевск: «Удмуртский университет», 2013. 424 с.

9. Новиков А.И. Смысл: семь дихотомических признаков. *Теория и практика речевых исследований*. М., 1999. С. 68–82.

10. Тертышный В.А., Шевченко И.В. Модель и метод многоаспектного поиска фактографических данных для поддержки принятия решений. *Вісник Кременчуцького національного університету імені Михайла Остроградського*. 2016. Вип. 5/2016 (100). С. 19–25.

11. Лупал А.М. Теория автоматов: учебное пособие. СПб: СПбГУАП, 2000. 119 с.

12. Валгина Н.С. Теория текста. М.: Логос, 2003. 173 с. Режим доступа: http://evartist.narod.ru/text14/01.htm (Дата обращения 17.01.2018).

THE MODEL AND METHOD OF TEXT THEMATIC STRUCTURIZATION ON THE BASIS OF STOCHASTIC AUTOMATICS

I. Shevchenko, A. Lebedinets, D. Vasilyev

Kremenchuk Mykhailo Ostrohradskyi National University

vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E-mail: ius.shevchenko@gmail.com

Purpose. To propose a model for analyzing the text structure, which is characterized by the use of stochastic matrices and automata for text tracing. This allows you to study the change in the significance of certain aspects and story lines along the length of the text, to reveal a program for the representation of verbal images. **Results.** Stages of the text content analysis method have been given. The method difference is that the thematic aspects are distinguished in the content, the change in their significance is traced, stochastic matrixes of associative connections of entities are formed, and annotations are formulated for each significant aspect of the content. This allows thematically to structure the

analyzed text. An experimental verification of the method and model is carried out. The graphs of the change in the significance of the aspects are obtained and annotations are formed for each significant aspect. For experiments to test the working capacity of the proposed method and models, two texts were selected – a technical text with a volume of 2400 words (without a title, conclusions and a list of references), and an artistic text, with a volume of 1500 words. The keywords were selected using the TextAnalyst program. The plot lines were constructed graphs of the significance change of aspects. Significant aspects were highlighted and annotations were formed. The compression ratio was 10%. **Practical value.** The experiment results confirm the method operability. As a merit of the proposed approach, one can note the relative simplicity of implementation and 100% completeness and even some redundancy of the display of relevant information in the collection of annotations. Redundancy is easily eliminated by threshold processing of the found material for each aspect. The original structure of the document has practically no value. It can be arbitrary texts or sets of values of fields of a heterogeneous database. In future studies, it is planned to analyze trends in the significance of aspects, to study the correlation of aspects, their consistency, and to construct the phase structure of the text. When using knowledge bases, it is possible to identify contradictions, collisions, and build meaningful interpretations. References 12, figure 1, table 2.

Key words: text analysis, thematic structuring, stochastic matrices, aspect annotations.

REFERENCES

1. Novikov, A.I. (2007), *Tekst i ego smyislovyie dominantyi; pod red. N.V. Vasilevoy, N.M. Nesterovoy, N.P. Peshkovoy* [Text and his semantic dominants], Institut yazyikoznaniya RAN, Moscow, Russia.

2. Surkova, A.S. (2014), "Authentication of texts on the basis of informative portraits", *Transactions of the Nizhny Novgorod. N.I. Lobachevsky University*, no. 3 (1), pp. 145–149.

3. Bolshakov, I.A., Gelbukh, A. (2004) Computational linguistics: models, resources, applications, First edition, Universidad Nacional Autónoma de Mèxico, Mexico.

4. Leontev, A.A. (2005), *Osnovyi psiholingvistiki* [The basics of psycholinguistics], Smyisl, Moscow, Russia.

5. Antipov, A.G., Anosova, K.M. (2008), "Acts of nominative derivation in the aspect of cognitive activity of individual (on material of dialectal word-formation)", *Philological sciences. Questions of theory and practice*, iss. 1, no. 1, part I, pp. 8–11.

6. Matveeva, L.Yu. 2015, "Probabilistic prognostication of sounding speech: to raisingof problem(review)", *Saratov scientific medical journal*, iss.11, no. 2, pp. 216–220.

7. Subbotina, M.V. (2016), "A distributive metaphor as method of research of work of literature"**,** *Scientific herald of the Voronezh State Architectural and Construction University. Modern linguistic and methodological and didactic studies*, iss. 1, no. 29, pp. 56–67.

8. Lelis, E.I. (2013), *Podtekst kak lingvoesteticheskaya kategoriya v proze A.P. Chehova* [Subtext as a linguistic esthetic category in prose. Chekhov], «Udmurtskiy universitet», Izhevsk, Russia.

9. Novikov, A.I. (1999), "Sense: seven dichotomy signs", *Theory and practice of speech researches*, pp. 68–82.

10. Tertyishnyiy, V.A., Shevchenko, I.V. (2016), "Model and method for multidimensional search of factual data for decision support", *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, iss. 5, no. 100, pp. 19–25.

11. Lupal, A.M. (2000), *Teoriya avtomatov: uchebnoe posobie* [Theory of automata: a tutorial], SPbGUAP, SPb, Russia.

12. Valgina, N.S. (2003), *Teoriya teksta: uchebnoe posobie* [Text theory: a tutorial], Logos, Moscow, URL: http://evartist.narod.ru/text14/01.htm.

Стаття надійшла 22.01.2018.