DOI: 10.17117/na.2015.09.773

Поступила (Received): 29.09.2015

<http://ucom.ru/doc/na.2015.09.773.pdf>**Надеждин Е.Н.****Задача выявления цепочки ключевых слов и предложений при семантическом анализе текста****Nadezhdin E.N.****The task of identifying the chain of keywords and proposals in the semantic analysis of text**

Рассмотрены существующие подходы к построению и анализу семантической модели текста. Сформулирована задача формализованного поиска и выявления цепочек ключевых слов и предложений в процессе семантического анализа текста

Ключевые слова: семантический анализ текста, ключевые слова, цепочка ключевых слов и предложений

The existing approaches to the construction and analysis of the semantic model of the text. The problem of formalized research and identify chains of key words and phrases in the semantic analysis of the text

Key words: semantic analysis of the text, keywords, key chain of words and sentences

Надеждин Евгений Николаевич

Доктор технических наук, профессор, главный научный сотрудник

Государственный научно-исследовательский институт информационных технологий и телекоммуникаций

г. Москва, Брюсов пер, 21

Nadezhdin Evgeniy Nikolaevich

Doctor of Engineering Sciences, Professor, Chief Researcher

State institute of information technologies and telecommunications

Moscow, Bryusov Pereulok, 21

Одна из актуальных проблем компьютерной лингвистики заключается в создании естественного языкового интерфейса для общения человека с компьютером. В этой проблеме особое место занимают вопросы автоматического анализа текста и распознавания слитной человеческой речи [1, 6].

Предположим, что на вход системы автоматического распознавания (САР) интеллектуального робота поступает параметризованный речевой сигнал в виде последовательности векторов параметров в пространстве первичного описания. Параметризация сигнала традиционно производится пре-процессором, который моделирует восприятие речевого сигнала и фоновых помех.

На выходе САР результат распознавания формируется в виде одной или нескольких гипотез о произнесенной последовательности слов. Эти последовательности, дополненные оценкой сопутствующих речевых параметров, передаются в блок смысловой интерпретации, где и происходит принятие окончательного решения в пользу одной из рабочих гипотез. Адекватным ответом

САР на слитную речевую последовательность будем считать ту допустимую фразу или предложение, преобразованный прототип которой оказался самым близким (в смысле заданного критерия) к предъявленному для распознавания речевому сигналу. Необходимым условием реализации функционала САР являются структуризация и формализованное представление знаний предметной области.

Как показывает практика компьютерной лингвистики, фонемные эквиваленты графемных представлений фраз и предложений конкретной предметной области, также как и в случае генерации фонемных представлений слов, до настоящего времени формируются в ручном или автоматизированном режимах. При этом строится сетевая модель переходов, в которой учитываются возможные комбинации слов всех предложений всех текстов, описывающих предметную область. Целью настоящего исследования является выявление рациональных путей формализации этого процесса.

В ряде опубликованных работ [1, 2, 5-7] исследуются различные подходы к анализу семантики предметной области и формированию её ассоциативной модели. Для определённости исследования примем допущение, что свойства предметной области отображаются в форме текстовой информации.

В задачах смыслового анализа текстовых документов хорошо зарекомендовал себя латентный семантический анализ (*Latent semantic analysis*) (ЛСА) [7, 8]. В основе метода ЛСА лежат принципы факторного анализа, и, в частности, процедура выявления латентных связей изучаемых явлений или объектов. Отметим, что данный метод положительно себя зарекомендовал при извлечении контекстно-зависимых значений лексических единиц за счёт статистической обработки больших корпусов текстов. Метод ЛСА опирается на линейный алгебраический подход и использует приведение матриц к каноническому виду. Здесь изучается прямоугольная матрица данных с числом столбцов n , равным числу разных слов, и со строками, которые представляют семантически обособленные фрагменты текста (концепции), представленные предложениями, фразами или синтагмами.

Число повторений слова в «концепциях» характеризует их статистическую значимость и интерпретируется как мера смысла. На столбцах и строках могут быть введены априорные целевые функции (функции интереса) и изучены условия диффузии интереса при движении по матрице. Далее применяется алгебраическая процедура, которая формирует сингулярное разложение прямоугольной матрицы (*Singular value decomposition*). В ходе разложения матрица разбивается оптимальным образом на сумму декартовых произведений векторов строк на векторы слов с весами, равными собственным значениям матрицы. Тем самым в неявной форме решается классическая задача кластеризации в пространстве слов и концепций, что позволяет определить формальные решения для целого ряда типовых задач смыслового анализа.

Основные недостатки метода ЛСА заключаются в формально-математическом подходе, в сложности строгой интерпретации численных характеристик и основанных на них заключений, в существенной трудоёмкости вычислений, кубически зависящей от объёма исходного текста. Выделение струк-

турных элементов (СЭ) освобождает текст от случайных (шумовых) вкраплений, однако информация, которую несут СЭ, может быть неактуальной. Это имеет место в случаях, если конкретный СЭ используется в более широких контекстах или представляет субъективно авторское изложение или типовую фразу (штамп).

В рамках используемой нами информационной концепции смысл каждой фразы, каждого предложения и документа определяется лишь на фоне предыдущего (или объемлющего) текста и измеряется количеством новой информации, которую этот фрагмент несет. Несомненный практический интерес представляют приложения метода ЛСА к задачам идентификации семантических моделей знаний предметной области [7]. Однако применение известного метода в каноническом виде здесь связано с большими трудозатратами.

Авторы работы [6] обосновали технологию статистической обработки текстовой информации TextAnalyst, позволяющая автоматически выявлять ключевые понятия в тексте на основе идентификации частотного портрета предъявленного текста. Можно выделить несколько этапов обработки информации. На первом этапе осуществляется первичная обработка: из текста удаляется нетекстовая информация, текст сегментируется на слова и предложения, из текста удаляются стоп-слова, рабочие и общеупотребимые слова, а оставшиеся слова подвергаются морфологической обработке. Морфологическая обработка производится с использованием традиционного морфологического словаря – словаря первого уровня – $\{B_i\}^1$. Далее формируется словарь второго уровня – $\{B_i\}^2$ – словарь корневых основ и устойчивых словосочетаний. На следующем этапе строится частотный портрет текста, в результате выявляются частоты p_i встречаемости корневых основ понятий (полученных в результате морфологического анализа) и их устойчивых сочетаний, и частоты p_{ij} их попарной встречаемости в предложениях текста, то есть формируется словарь третьего уровня $\{B_i\}^3$. Полученная сеть является исходной для формирования сети переходов.

На третьем этапе обработки частоты встречаемости преобразуются в смысловые веса с использованием специальной итеративной процедуры, в результате чего наибольшие веса получают понятия, которые «стягивают» на себя смысловую структуру текста. Полученные таким образом смысловые веса ключевых понятий показывают значимость этих понятий в тексте. В дальнейшем эта информация используется для выявления группы предложений, содержащих наиболее важную информацию в тексте.

В результате исследований строится искомая однородная семантическая (ассоциативная) сеть N как совокупность несимметричных пар понятий $\langle c_i, c_j \rangle$, где c_i и c_j – понятия, связанные между собой отношением ассоциативности (совместной встречаемости в некотором фрагменте текста).

В последние годы новый импульс к развитию получили методы лингвистического анализа текстов на естественном языке [1, 6]. Классический лингвистический подход к анализу текста предполагает существование нескольких относительно независимых уровней анализа: морфологического, синтаксиче-

ского и семантического. В настоящее время успешно культивируются логико-лингвистические методы автоматического анализа текстов, которые основываются на эвристических правилах, разработанных экспертами-лингвистами [1]. Используемые при этом механизмы разработки лингвистических ресурсов весьма наукоёмки, поскольку для создания автоматических систем необходима разработка модели представления знаний части естественного языка, что требует согласованных усилий высококвалифицированных лингвистов, системных программистов и операторов.

Рассмотрим сущность предлагаемого нами подхода к формализованному построению ассоциативной модели предметной области для случая семантического анализа заданной текстовой конструкции [3].

Предположим, что в ходе семантического анализа исходного текста выделены и отобраны ключевые понятия (слова, термы) $Z_j, j = \overline{1, r}$. Для автоматизации процесса формирования сети допустимых переходов между ключевыми словами в заданной предметной области воспользуемся апробированной технологией автоматического построения семантической сети текста на основе корпуса текстов конкретной предметной области [6, с. 370].

Основным инструментом формализованного представления семантики предметной области (текста) выступает семантическая сеть предметной области (текста). Далее с учётом выявленных особенностей и дополнительной информации о тематике и структуре текста могут быть осуществлены типизация ключевых слов и отбор их допустимых парных сочетаний $(Z_i - Z_j), i, j = \overline{1, r}, j \neq i$. Для отображения парных сочетаний ключевых слов введём матрицу инцидентий $C = \{c_{i,j}, i, j = \overline{1, n}\}$, элементы которой определяются по правилу:

$$c_{i,j} = \begin{cases} 1, & \text{если элементы } (Z_i, Z_j) \in D_k; \\ 0, & \text{если } j = i; \\ 0, & \text{если элементы } (Z_i, Z_j) \notin D_k. \end{cases}$$

Здесь $D_k \in D$ – подмножество парных сочетаний ключевых слов.

Обязательным этапом автоматического формирования сети переходов является обоснование механизма выявления существенно важных для описания предметной области ключевых фраз и предложений [6]. Поэтому на основании имеющихся данных сформулируем задачу синтеза семантической модели знаний, которая заключается в построении обобщённого функционального графа $G(P, I)$ по моделям $P_v, v = \overline{1, m}$, выявленных цепочек ключевых слов.

Пусть семантическая модель предметной области отображается неориентированным графом $G^*(P, I)$, где $P = \{P_v, v = \overline{1, m}\}$ – множество вершин укрупнённого графа G^* , представляющих выделенные пары слов, а $I = \{I_w, w = \overline{1, r}\}$ – множество дуг этого графа, представляющих информационные связи между концептами $P_v, v = \overline{1, m}$. Требуется построить граф $G^0(P, I)$, который обладает следующим свойством $G^0 = \bigcup_j G_j$ для определённого набора условий, отражающих требования представительности сетевой модели предметной области.

Для реализации сформулированной задачи синтеза можно воспользоваться методом, изложенным в работе [5, с. 41]. В основу известного метода положено рекуррентное соотношение

$$G_{k,l} = G_{k-1,l} \cup G_k, \quad G_{1,l} = G_1, \quad k = 2, 3, \dots, n.$$

Алгоритмы, построенные на основе этого соотношения, сводятся к пошаговому объединению графов. Пусть на $(k-1)$ -м шаге в результате объединения получен граф $G_{k-1,l}$. Для объединения этого графа с графом G_k в граф G_k выбирается очередная вершина графа G_k и сравниваются инцидентные ей дуги с дугами графа $G_{k-1,l}$.

По результатам такого сравнения принимаются решения о включении элементов графа G_k в граф $G_{k-1,l}$ с целью образования графа $G_{k,l}$. В качестве базовых операций таких алгоритмов используются: операции пересечения множеств и манипуляции с матрицами инцидентий, обеспечивающие замену модулей (вершин), дополнение информационных множеств новыми элементами (включение новых дуг), перевод элементов одних множеств в элементы других множеств и др.

Полученный в результате синтеза и цепочки преобразований $G \rightarrow \dots \rightarrow G^*$ искомый граф G^* имеет ярусно-параллельную форму, удобную для содержательной интерпретации и последующей алгоритмизации модели предметной области.

После выбора ключевых слов и значимых предложений текста на основе этих предложений строится ассоциативная сеть, которая является основой для идентификации сети переходов между ключевыми словами для этой группы предложений. Сеть, построенная на всех предложениях корпуса текстов, описывающего предметную область, ранга выше порогового, является сетью переходов между ключевыми словами для всей предметной области.

Список используемых источников:

1. Ваграменко Я.А., Фанышев Р.Г. Технология интеллектуального анализа текстовой информации в базах знаний образовательной экспертной системы // Педагогическая информатика. 2011. № 1. С. 57-62.
2. Васенин В.А., Афонин С.А., Козицын А.С. Автоматизированный анализ текстовой информации // Информационные технологии. 2009. №7. С. 56-57.
3. Надеждин Е.Н. Теоретические аспекты семантического анализа междисциплинарных знаний в интеллектуальных обучающих системах. Тульский государственный педагогический университет имени Л.Н. Толстого. Тула, 2013. 18 с. В ВИНТИ 17.12.2013 г.; № 374-B2013.
4. Надеждин Е.Н., Смирнова Е.Е. Методы моделирования и оптимизации интегрированных систем управления организационно-технологическими процессами в образовании: монография. Тула: ТулГУ, 2013. 250 с.
5. Парасюк И.Н., Сергиенко И.В. Пакеты программ анализа данных: технология разработки. М.: Финансы и статистика, 1988. 159 с.
6. Харламов А.А., Ермоленко Т.В. Семантическая сеть предметной области как основа для формирования сети переходов при автоматическом распознавании слитной речи // Открытые семантические технологии проектирования интеллектуальных систем. Минск: БГУИР, 2013. С. 369-374.
7. Thomas K., Landauer T., Harshman R. Latent semantic analysis, J. Amer. Soc. of Information Science, 1990, 41(6).

8. Olmos R., León J.A., Jorge-Botana G. & Escudero I. Using Latent Semantic Analysis to grade brief summaries some proposals. *International Journal of Continuing Engineering education and Long Learning* 21 (2/3). 2011. P. 192-209.

© 2015, Надеждин Е.Н.

Задача выявления цепочки ключевых слов и предложений при семантическом анализе текста

© 2015, Nadezhdin E.N.

The task of identifying the chain of keywords and proposals in the semantic analysis of text