13. *Баранюк В. В., Богорадникова А. В., Смирнова О. С.* Определение семантического содержания предметной области на основе формирования тезауруса // International Journal of Open Information Technologies. 2016. Т. 4. № 9. С. 74–79.

14. *Cross T.* After Moore's Law // The Economist Technology Quarterly. 2016. № 3.

15. *Линев А. В., Боголепов Д. К., Бастраков С. И.* Технологии параллельного программирования для процессоров новых архитектур / Под ред. В. П. Гергеля. – М.: Изд-во Московского университета, 2010. 160 с.

16. *Ding N., Melloni L., Zhang Hang, Tian X., Poeppel D.* Cortical tracking of hierarchical linguistic structures in connected speech // Nature Neuroscience. 2015. Vol. 19. Iss. 1. P. 158–164.

17. *Мелихов А. А.* Применение дерева синтаксического разбора предложений для повышения релевантности результатов частотного анализа текста // Нейрокомпьютеры: разработка, применение. 2016. № 3.

18. *Prezioso M., Merrikh-Bayat F., Hoskins B. D., Adam G. C., Likharev K. K., Strukov D. B.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors // Nature. 2015. Vol. 512. Iss. 7550. P. 61–64.

19. *Свердлик А. Г.* Как эмоции влияют на абстрактное мышление и почему математика невероятно точна / Науч. ред. Ф. Абрамовича и С. Шрейдера. – М.: ЛЕНАНД, 2016. 256 с.

20. *Furber S.* Large-scale neuromorphic computing systems // Journal of Neural Engineering. 2016. Vol. 13. Iss. 5. 14 p.

21. *Галушкин А. И.* На пути к нейрокомпьютерам с использованием мемристоров // Информационные технологии. 2014. № 4. С. 2–19.

22. *Panov P., Soldatova L. N., Džeroski S.* Generic ontology of datatypes // Information Sciences. 2016. Vol. 329. P. 900–920.

Neurocomputing technologies` implementation in decision support system for bionic research

***Alexandr Alexandrovich Melikhov**, post-graduate student, assistant of the Department*
Federal State Budget Education Institution of Higher Education «Moscow Technological University» (MIREA)

The former article regards an intellectual system which provides information support for bionic technologies` research and development. The main focus is on its integrative capacity what can be achieved by specific integration modules which utilize the potential of non-classic computing architectures. The premises for such integration are provided, the basic concepts of integrative modules` architecture are discussed.

Keywords: information system, innovation support, thesaurus, natural language processing, knowledge engineering, bionic information resources.

УДК528.88; 551.465; 551.463.8; 551.463.6; 528.873.044.1; 629.78

МЕТОДЫ И АЛГОРИТМЫ ИНФОРМАЦИОННОЙ ИНТЕРПРЕТАЦИИ

***Евгений Евгеньевич Чехарин**, зам. начальника центра информатизации МИРЭА, ст. преподаватель кафедры инструментального и прикладного программного обеспечения института информационных технологий,*
e-mail: tchekharin@mirea.ru,
Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский технологический университет» (МИРЭА),
https://www.mirea.ru

DOI: 10.21777/2312-5500-2016-5-39-49

Статья описывает методы и алгоритмы информационной интерпретации. Описаны методы и алгоритмы анализа текста. Информационная интерпретация раскрывается как технология применения моделей информационных ситуаций, информационных конструкций и информационных единиц. Показаны разные виды интерпретации, связанные с анализом текста. Описан процесс информационного взаимодействия как один из видов информационной интерпретации.

Ключевые слова: знание; информация; информационная интерпретация; анализ текста; лингвистика; компьютерная лингвистика; семантика; информационное взаимодействие; информационная ситуация; информационная конструкция; информационные единицы; информационное поле; семантическое поле.

Введение

Интерпретация трактуется как [1, 2] совокупность значений, придаваемых элементам какой-либо системы. Информационная интерпретация – это технология в информационном поле [3]. Кроме того, эта интерпретация отличается от интеллектуальной интерпретации, поскольку связана с алгоритмами. Информационная интерпретация может быть рассмотрена как информационное взаимодействие «объект – объект». Интерпретация информационных моделей, информационных единиц и информационных конструкций осуществляется в информационной области [1]. Информационная область – это пространство, где находится информация, информационные ресурсы, при этом информация и информационные ресурсы создаются, транспортируются, обрабатываются и используются. Информационная конструкция [4–6] – широкое понятие, которое может описывать обычный текст, информационную модель, информационную систему, информационную ситуацию, информационную позицию объекта в информационной ситуации, алгоритм, правила и т. п. Информационная конструкция, как текст, имеет семантику и форму. Если рассматривать текст как информационную конструкцию, то возникает обязательное требование структуры в тексте, как минимум семантической структуры.

Е.Е. Чехарин

Интерпретация текстовых конструкций. Среди информационных конструкций большое место занимают текстовые информационные конструкции. Для интерпретации информационной конструкции используем понятие информационной ситуации [7, 8], но введем дифференциацию этого понятия. Будем различать: информационную ситуацию по структуре, информационную ситуацию по семантике, морфологическую информационную ситуацию. Исследование текста для последующей интерпретации ведет к необходимости анализа текста, а в современной трактовке к интеллектуальному анализу текста (text data mining) [9].

Термин «анализ текста» описывает набор лингвистических, статистических методов и методов машинного обучения. В совокупности эти методы создают в большом модель и структуру семантического информационного содержания текстовых источников для бизнес-аналитики, исследовательского анализа данных, исследования или инвестирования [10]. В совокупности эти методы в малом создают структуру семантического информационного окружения информационной конструкции или информационной единицы [11].

Под этой технологией понимают процессы получения информации высокого качества (high-quality information) из текстовых массивов. Информацию высокого качества получают путем использования неких эталонов (шаблонов) с помощью различных средств, одним из самых распространенных среди которых является статистический анализ.

Интеллектуальный анализ текста обычно включает в себя: процесс структуриро-

вания входного текста, применение правила разбора, получение шаблонов, оценку, анализ и интерпретацию результата. Информационная интерпретация, основанная на интеллектуальном анализе текста, включает два этапа: предварительную обработку и интерпретационную обработку

Предварительная обработка включает: текстовую категоризацию, текстовую кластеризацию, стратификацию текста, концептуализацию, извлечение семантической сущности, технологию гранулирования таксонов, семантический анализ, обобщение информационных конструкций, моделирование отношений сущностей (или построение отношений между сущностями).

Второй этап обработки текста включает: извлечение семантической информации, лексический анализ, статистический анализ, распознавание образов, аннотирование, интеллектуальный анализ текста, анализ ссылок, анализ ассоциаций, визуальный (когнитивный) анализ, предикативный анализ.

Главной целью второго этапа является превращение информационной конструкции в данные для анализа или управления, с помощью обработки на естественного языке и аналитических методов.

В широком смысле процессы анализа текста выходят за рамки информационной интерпретации. Они включают следующие основные технологии, большая часть которых входит в задачи интерпретации.

- Информационный поиск или идентификация «корпуса текста» представляет собой подготовительный этап: сбора или определение набора текстовых материалов, которые находятся в Интернете или хранятся в файловой системе, в базы данных.
- «Распознавание сущности по имени» – метод, который используют в газетирах или в статистических методах для выявления названных текстовых функций: люди, организации, географические названия, символы биржевых сводок, определенные сокращения. В геоинформатике этот метод называется георефернция [12].
- Распознавание объекта по идентификационному образцу, такому как телефонные номера, адреса электронной почты, почтовые адреса, номера автомашин и пр.
- Распознавание по кореференции. Определенные словосочетания относятся к объекту и могут его характеризовать. Это понятие также близко к георефернции. Например, Москва, или столица России.
- Анализ отношений, фактов и событий. Идентификация этих понятий и подобной информации в тексте.
- Анализ семантических [13] и информационных [14] отношений в тексте.
- Количественный анализ текста представляет собой набор методов, в которых человек или компьютер извлекает семантические или грамматические отношения между словами, чтобы выяснить смысл или когнитивный стиль.
- Сопоставление текста с личностью, психологическое профилирование или психологический портрет.

Кластерные принципы интерпретации текста. Один из принципов информационной интерпретации состоит в том, что текст рассматривается не как аморфная совокупность, а как информационная конструкция. Второй принцип информационной интерпретации заключается в том, что процесс интерпретации рассматривается как информационное взаимодействие:

«исходный файл – алгоритм (программа) – результирующий файл»;
«исходная информационная конструкция – алгоритм (программа) – результирующая информационная конструкция».

Одним из направлений интерпретации текстов является метод отслеживания развития сцен и ситуаций в информационном потоке (Topic Detection and Tracking) [15]. Задачей этих методов является определение информационной ситуации, к которой относится текстовое описание.

Модели таких интерпретаций предполагают, что в реальном мире вначале появляется первичная информация о ситуации, а затем появляются уточняющие факты о самой ситуации. Например, при крушении самолета вначале появится сообщение о самом факте крушения, затем будут появляться уточнения этого события.

Одной из проблем анализа ситуаций является возможный процесс перетекания (трансформации) одной ситуации в другою. Результатом анализа и обработки этих ситуаций являются информационные кластеры потоков, описывающие развитие ситуации. Задача отслеживания развития ситуаций схожа с задачей кластеризацией текстов. Как и задача кластеризации текстов, задача кластеризации информационных потоков обладает большой размерностью данных. Это связано с тем, что большинство методов кластеризации работает с данными, представленными в виде векторов в пространстве R^n [16]. Представление текстовых данных в таком пространстве обычно осуществляется с помощью процедуры сопоставления каждого признака с функцией-индикатором данного слова. Следовательно, общая размерность пространства задачи определяется общим количеством таких признаков и сопряжена с проблемой больших данных [17]. Поскольку в качестве признаков используются слова и их сочетания, то общая размерность пространства может достигать 10 млн.

Однако при этом вектор признаков каждой информационной ситуации «разрежен». Такая ситуация типична для многомерных баз данных. Она означает, что небольшое число параметров вектора обладает значениями. Стандартный подход кластеризации текстов описан во многих источниках. Наиболее часто описаны алгоритмы k-средних [18] и алгоритмы иерархической кластеризации [19].

Существуют отличия кластеризации данных [20] и текста. Во-первых, основные алгоритмы кластеризации работают на заранее статическом множестве данных. Информационный поток, описывающий трансформацию информационной ситуации, не является фиксированным. Использование алгоритмов кластеризации данных на информационном потоке означает необходимость при каждом новом сообщении проводить кластеризацию заново.

Многие алгоритмы кластеризации данных задают заранее число кластеров, на которые нужно разбить совокупность данных. Это предположение существует в алгоритмах k-means, k-medoid и их оптимизированных версиях [18]. В случае задачи обработки информационных потоков эта информация отсутствует. Алгоритмы кросс-валидации [21], которые инкрементно подбирают число кластеров, не применимы, поскольку они требуют много времени для подбора кластеров.

В ряде работ [22] для кластеризации текста предлагается линейный во времени алгоритм приблизительной кластеризации. Первоначально множество кластеров считается пустым. Затем для каждого нового сообщения в информационном потоке выполняются инкрементные операции.

Особенностью алгоритма является то, что решение о принадлежности какой-либо точки принимается только один раз. Другой особенностью алгоритма является то, что кластеризация производится только на основе текстового материала. Однако это приводит к тому, что не учитывается вспомогательная информация: даты публикации материала, наличие ссылок, дополнительная информация по теме. Все рассмотренные алгоритмы обладают существенными ограничениями и используют только «поверхностное» представление информационной конструкции.

Лингвистические методы анализа текста. В качестве альтернативы можно рассмотреть лингвистические методы анализа текста. Лингвистические методы анализа текста используют методы компьютерной лингвистики. Компьютерная лингвистика ориентирована на семантический анализ текстов, их структуры и содержания.

Основной подход к анализу текста в компьютерной лингвистике включает несколько уровней анализа. Стандартными считают следующие уровни: предморфологический, морфологический, синтаксический и семантический [23]. Такое разделение условно, по-

скольку существуют задачи, которые относятся сразу к нескольким уровням анализа.

Верхним уровнем анализа текста является предморфологический. На этом уровне выделяется общая структура текста, то есть выделяются отдельные фразы, абзацы, параграфы, границы условных слов и предложений. Этот уровень анализа является общим для задач информационного поиска и для задач глубокого анализа текста.

Следующий уровень анализа текста – морфологический. На этом уровне уточняются границы слов и предложений. Для каждого слова выделяются его грамматические характеристики и начальные формы. В задачах информационного поиска этот уровень анализа также присутствует, но используется с целью уменьшения размерности пространства термов, а не для извлечения дополнительной информации. В информационном поиске используется процедура лемматизации, или стемминга, которая выделяет неизменяемую часть слова.

Третьим уровнем анализа является синтаксический уровень. За последние несколько десятков лет он получил значительное развитие. Процедура анализа рассматривает каждое предложение по отдельности и конструирует его синтаксическую структуру. Такая структура отражает структуру подчинения слов в предложении, что показывает, как слова в предложении зависят друг от друга.

Существует две распространенных модели представления синтаксической структуры. Н. Хомский [24] предложил рассматривать синтаксическую структуру как построение составляющих или групп предложения. В такой модели группа предложения – структурная информационная единица предложения, сформированная из составляющих предложений меньшего размера. Элементарными составляющими являются семантические информационные единицы – отдельные слова предложения. Достоинство этого представления синтаксической структуры состоит в его связи с синтаксическим анализом формальных языков, например языков программирования, а также с теорией информационных единиц [25–27]. Особенно важно такое построение для интерпретации, поскольку оно позволяет создавать общие процедуры для интерпретации естественного и формальных языков.

Анализ формальных языков условно хорошо формализован и существуют эффективные процедуры анализа. Однако методы анализа формальных языков плохо применимы к анализу естественного языка. Формальные процедуры анализа, как правило, предполагают, что существует только одна корректная синтаксическая структура, и не предполагают, что может возникнуть неоднозначность разбора. При построении формальных языков в них закладывают изоморфность, что исключает анализ гомоморфных ситуаций. При использовании методов формального анализа неоднозначность разрешается за счет построения формального языка.

Другим ограничением формального подхода является предположение о том, что «составляющие предложения» (сложные информационные единицы) неразрывны в аспекте отражения всего предложения. Они как информационные единицы покрывают сплошной интервал слов предложения. Такое предложение справедливо не для всех языков.

Например, английский язык обладает строгим порядком слов в предложении. Для него предположение о семантической неразрывности допустимо. Исключения могут быть обработаны с помощью специальных средств построения

В русском языке нет строго порядка слов в предложении, поэтому ситуация иная. Установлено, что в корпусе русского языка до 10% предложений имеют разрывные группы. В результате такой подход слабо подходит для русского языка.

Другим подходом к представлению синтаксических структур является грамматика зависимостей, впервые предложенная Л. Теньером в [29]. Синтаксическая топологическая структура в этом подходе представляет собой граф, в котором вершины составляют слова предложения, а ребра – синтаксические связи между этими словами.

Такое представление обладает более выразительными средствами представления синтаксических структур. Если синтаксическая структура не имеет непроективных свя-

зей, которые пересекаются с другими, то такой граф имеет вид дерева и может быть однозначно преобразован в представление на основе составляющих.

Возможно и обратное – любая синтаксическая структура на основе составляющих может быть преобразована в дерево зависимостей. Этот механизм широко используется в теории реляционных баз данных

В общем виде синтаксическая структура анализа на основе дерева зависимостей задается следующим образом. Пусть дано предложение $x = \{w_1, w_2, \dots, w_n\}$, где w_i – i-е слово в предложении. Дополнительно вводится фиктивное слово w_0 для обозначения вершины синтаксической структуры предложения. $L = \{l_1, \dots, l_{|L|}\}$ – множество возможных типов связей. Тогда синтаксическая структура предложения – ориентированный маркированный граф $G = (V, A)$, в котором $V = \{0, 1, \dots, n\}$ – множество вершин, соответствующих словам предложения, $A \in V \times L \times V$ – множество связей графа.

Синтаксическая структура считается правильно построенной, если

- вершина под номером 0 является корнем синтаксического дерева;
- у каждой вершины не более одного родителя, т. е. если $(i, l, j) \in A$, то не существует таких вершин i' и меток l', что $(i', l', j) \in A$;
- в графе G нет циклов.

Задача синтаксического анализа формулируется как задача построения синтаксической структуры в виде ориентированного маркированного графа $G = (V, A)$ для заданного предложения $S = \{w_1, w_2, \dots w_n\}$, где n – число слов в предложении, V – вершины (слова), A – дуги (связи). Построенный граф должен удовлетворять приведенным выше условиям.

Поскольку основной информационной единицей анализа является связь между двумя словами, то необходимо определить ее характеристики:

- источник связи – слово (семантическая информационная единица, СИЕ), из которого связь выходит;
- зависимое – слово, в которое связь приходит;
- синтаксический тип связи – показывает роль этой связи в предложении. Для русского языка определено около 80 различных типов связей [30];
- направление связи – показывает, в какую сторону связь направлена;
- левое и правое слова связи – это слова связи относительно их порядка в предложении;
- имя связи – имя синтаксического отношения, связывающего СИЕ между собой.

Эти характеристики являются «внутренними» и относятся только к связи между информационными единицами. Однако есть и очень важная «внешняя» характеристика связи, которая значительно влияет на алгоритмы синтаксического анализа: проективность синтаксической связи.

Свойство проективности синтаксической связи можно показать графически – если построить дерево синтаксической структуры, то проективные связи между собой не пересекаются. Это классическая информационная модель.

Анализ текста на основе размеченных корпусов текстов. Иным подходом к анализу лингвистической информации является методика применения размеченных корпусов текстов [31]. Это направление в лингвистике называют корпусной лингвистикой. Корпусная лингвистика – относительно молодая область, поскольку первые корпусы с синтаксической разметкой представляли собой карточки, на которых была представлена синтаксическая структура предложения. Кроме того, основное предназначение корпусов заключалось в создании и проверке существующих лингвистических теорий. Такая организация корпусов резко снижала возможности прикладного использования этих корпусов из-за высокой трудоемкости.

Первым корпусом, представленным в электронном виде, содержащим синтаксическую разметку, был корпус Lancaster-Leeds Treebank [31], который был создан в 1989

году в рамках экспериментов по синтаксическому анализу. Сегодня электронные корпуса с синтаксической разметкой существуют не только для английского, но и для многих других языков. В частности, для русского языка существует Национальный корпус русского языка (НКРЯ) [30].

В настоящее время одним из основных применений размеченных корпусов текстов является их использование для построения и усовершенствования алгоритмов автоматической обработки текстов. При наличии в корпусе синтаксической разметки возможным становится построение процедур автоматического синтаксического анализа без привлечения экспертов-лингвистов. Корпус в таком подходе используется одновременно для двух целей. С одной стороны, он служит основным источником материала для самообучения автоматических алгоритмов синтаксического анализа, с другой – является эталонной моделью для оценки качества проводимого анализа.

Для эффективного использования размеченных корпусов в вышеуказанных целях необходимо соблюдение ряда условий. Наиболее важным является применение заранее определенной экспертами-лингвистами формальной схемы разметки. Во-вторых, необходимо, чтобы при разметке корпуса лингвисты последовательно придерживались этой схемы. Предполагается, что тексты являются реализацией модели языка, и если их лингвистическая разметка не будет последовательной, то алгоритмы автоматического машинного обучения не смогут выявить закономерности языка.

Другим важным условием является репрезентативность корпуса. Корпус должен содержать большое количество разнообразных текстов из различных областей и жанров, для того чтобы системы, построенные на его основе, могли качественно обрабатывать текст.

Синтаксический анализ на основе машинного обучения. Развитие корпусной лингвистики и появление размеченных корпусов текстов большого объема сделало возможным использование методов машинного обучения для построения автоматических синтаксических анализаторов [32]. В основе использованы алгоритмы распознавания образов с самообучением. При таком подходе корпус рассматривается не как языковая модель, а как обучающая выборка, поскольку для каждого предложения доступна его эталонная синтаксическая структура.

В процессе обучения производится выявление закономерностей, которые при обработке текстов (информационных конструкций) позволяют получать синтаксические структуры, максимально приближенные к эталонным. Для этого используются современные методы и модели машинного обучения.

Синтаксический анализ предложений с использованием алгоритма максимальных остовных деревьев. Алгоритм на основе максимальных остовных деревьев [32] преобразует задачу синтаксического анализа в задачу нахождения максимального остовного дерева на графе возможных связей. Для этого вводится функция оценки связи

$$s(i,j) = w \cdot f(i,j),$$

где $f(i,j)$ – функция векторизации признаков, на основе которых принимается решение о проведении связи из слова с индексом i в слово с индексом j, w – весовая модель оценки связи, полученная с помощью машинного обучения.

Алгоритм выбирает такое дерево, сумма оценок связей которого будет максимальна:

$$s(x,y) = \sum_{(i,j)\in y} s\,(i,j).$$

$$max_y s(x,y)$$

Задача обучения заключается в получении такой функции оценки связи, которая для максимального числа предложений из корпуса позволяла бы построить эталонную синтаксическую структуру. В этом случае задача обучения становится задачей ранжи-

рования – необходимо, чтобы правильная связь получала большую оценку, чем остальные потенциальные связи:

$$s(i,j) > s(k,j), \forall k \neq i,$$
$$s(i,j) > s(i,m), \forall m \neq j.$$

Таким образом, все связи можно представить как ранг, наиболее важным элементом которого является правильная (эталонная) связь. Для применения этого алгоритма необходимо определить правила для построения рангов. Для деревьев зависимостей их можно определить на основе следующих фактов:

- для каждого слова правильна только одна входящая связь;
- для данного слова потенциальными хозяевами могут быть все остальные слова предложения и вершина.

Другим важным вопросом является выбор функции векторизации признаков связи. Эта функция определяет преобразование информации о связи и ее участниках в числовой вектор. Классический вариант алгоритмов максимальных остовных деревьев предполагает независимость ребер графа – каждая синтаксическая связь проводится независимо от уже существующих.

На каждое слово в среднем приходится mn примеров, n – среднее число слов в предложении, m – число возможных имен связи. Алгоритмы на основе максимальных остовных деревьев различаются способом построения дерева на основе полученных оценок. Свойства проективных и непроективных алгоритмов сильно различаются. В частности, не существует вычислительного эффективного алгоритма, который мог бы построить непроективные связи и использовал бы при этом информацию об уже существующих связях. Для проективных связей такой алгоритм существует [32].

Семантическая интерпретация. Семантическая интерпретация включает сопоставление форме текста смыслового значения. Примером такой процедуры является дефиниция [34]. Основой для такой процедуры служит естественная формальная семантика, или естественная семантика. Естественная семантика – это семантика внешнего мира, сформулированная на естественном языке [34].

Компьютерная семантика может быть рассмотрена как некий внутренний уровень формальной семантики. Она обычно представляется ориентированным графом, аналогично представлению когнитивных карт. Это дает основание подчеркнуть связь между семантическим описанием и когнитивным описанием. Различие состоит в интерпретаторе. Если интерпретатор компьютерный, то задача интерпретации решается в формализованном (перенесенном на компьютер) семантическом поле терминологических отношений. Если интерпретатор включает человека, то задача интерпретации решается в когнитивном пространстве. В этом случае компьютерной семантике присваивается внешняя семантика с помощью человеческой интерпретации.

В практической деятельности любое представление задач реального мира требует перевода контекстных знаний эксперта (приложений высокого уровня) в воспроизводимые операции вычислительной машины (низкий уровень) или в модели обработки информации. Многообразие естественного языка позволяет описать задачи, которые невозможно адекватно описать на формальном языке. Поэтому для уменьшения семантического разрыва применяют различные средства. Одним из таких средств является информационное взаимодействие [35, 36].

Для интерпретации необходимо семантическое информационное взаимодействие. По существу, информационное взаимодействие структурно повторяет процесс управления с обратной связью. При интерпретации один объект является исходным (концептом), второй – порожденным (дефиницией). Интерпретация определяет качественное равенство между ними. Исходный объект имеет неполноту описания. При таком взаимодействии участвует субъект, поэтому в информационном поле необходимо рассматривать триаду

«исходная информационная конструкция – интерпретация –
расширенная информационная конструкция».

Если исходная информационная конструкция есть некий номен или объект, то расширенная информационная конструкция есть номен плюс дефиниция или объект плюс дефиниция. При информационном взаимодействии субъект принимает информацию о текущем описании информационной конструкции. Если текущее описание неадекватно исходному объекту, то формируется новое формальное описание в рамках того языка, на котором оно выполнено. Этот цикл повторяется, пока не будут исчерпаны возможности языка формального описания. Процесс информационного взаимодействия приведен на рис. 1.

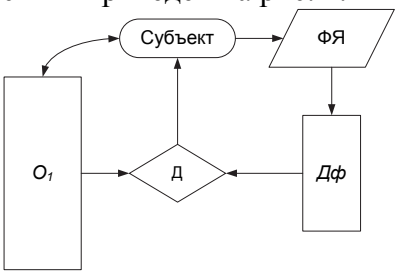

Рис. 1. Процесс информационного взаимодействия

Исходный объект O_1 служит основой интерпретации. Субъект использует O_1 как эталон и с помощью формального языка ФЯ формирует или находит дефиницию (Дф). При наличии информационного соответствия процесс интерпретации заканчивается. При наличии информационного взаимодействия в информационном поле осуществляется сравнение семантики O_1 и дефиниции. В дискриминаторе Д осуществляется такое сравнение, результат сравнения направляется субъекту. Субъект осуществляет семантический анализ. Семантический анализ служит для выявления информационного соответствия или семантического разрыва и формирования дополнительных действий для преодоления разрыва. Такой процесс повторяется, пока не будет достигнут баланс возможностей формального языка и описания объекта. По существу, Д выполняет функции формального интерпретатора. В процессе информационного взаимодействия происходят следующие изменения расширенной информационной конструкции (РИК) по отношению к исходной ИК:

- увеличение сложности РИК;
- модификация структуры РИК;
- повышение адаптивности РИК в сравнении с ИИК;
- интеграция информационных объектов в РИК;

Важной характеристикой такого информационного взаимодействия является инкрементность [37] и ресурсность [38]. Инкрементность интерпретации как взаимодействия выражается в том, что дескриптивный ресурс предшествующей стадии включается в дескриптивный ресурс последующей стадии. Ресурсность интерпретации как взаимодействия выражается в том, что субъект или объект в процессе интерпретации накапливают опыт или прескриптивный ресурс [39]. Этот опыт представляет собой неявное знание [40], которое при помощи когнитивной трансформации применяется как когнитивный информационный ресурс при формировании интерпретируемых объектов.

Заключение. Методы и алгоритмы информационной интерпретации находятся в состоянии развития. Поэтому они в большой степени опираются на методы и алгоритмы анализа текста, компьютерную лингвистику, технологии информационного поиска. Развитием этих подходов является применение обобщенной модели «информационная конструкция». Информационная конструкция может описывать текст на естественном языке и текст на формальном языке. Такое обобщение позволяет формализовать методы анализа естественного языка и формальных языков. Еще одними из методов обобщения интерпретации являются информационные единицы. Выделение информационных единиц на уровне структуры и семантики позволяет разделять структурную и семантическую ситуацию. Третьим отличием информационной интерпретации является введение модели информационной ситуации как среды, в которой происходит интерпретация. Информационная ситуация также является универсальной моделью, которая может подвергаться дифференциации и быть обобщенной. Применение моделей информационной ситуации,

информационной конструкции и информационных единиц является новым этапов в интерпретации вообще и интерпретации в информационной семантике в частности.

Литература

1. *Чехарин Е. Е.* Интерпретация информационных конструкций // Перспективы науки и образования 2014. № 6. С. 37–40.

2. *Чехарин Е. Е.* Интерпретация космической информации при исследовании Земли // Образовательные ресурсы и технологии. 2015. № 2 (10). С. 137–143.

3. *Цветков В. Я.* Естественное и искусственное информационное поле // Международный журнал прикладных и фундаментальных исследований. 2014. № 5 (часть 2). С. 178–180.

4. *Tsvetkov V. Ya.* Information Constructions // European Journal of Technology and Design. 2014. Vol. 5. Iss. 3. P. 147–152.

5. *Дешко И. П.* Информационное конструирование: Монография. – М.: МАКС Пресс, 2016. 64 с.

6. *Rozenberg I. N.* Information Construction and Information Units in the Management of Transport Systems // European Journal of Technology and Design. 2016. Vol. 12. Iss. 2. P. 54–62.

7. *Розенберг И. Н., Цветков В. Я.* Информационная ситуация // Международный журнал прикладных и фундаментальных исследований. 2010. № 12. С. 126–127.

8. *Цветков В. Я.* Информационные модели объектов, процессов и ситуаций // Дистанционное и виртуальное обучение. 2014. № 5. С. 4–11.

9. KDD-2000 Workshop on Text Mining – Call for Papers. http://www.cs.cmu.edu.

10. Archived November 29, 2009, at the Wayback Machine.

11. *Чехарин Е. Е.* Информационная модель семантического окружения // Перспективы науки и образования. 2014. № 4. С. 20–24.

12. *Цветков В. Я.* Георефернция как инструмент анализа и получения знаний // Науки о Земле. 2011. № 2. С. 63–65.

13. *Смулаковская Р. Л.* Лексико-семантические отношения в тексте (функционально-коммуникативный аспект). – М.: Просвещение, 1987.

14. *Tsvetkov V. Ya.* Information Relations // Modeling of Artificial Intelligence. 2015. Vol. 8. Iss. 4. P. 252–260.

15. *Park S. B., Zhang B. T., Kim Y.* Word Sense Disambiguation by Learning Decision Trees from Unlabeled Data // Applied Intelligence. 2003. Vol. 19. No. 1. P. 27–38.

16. *Feldman R., Sanger J.* The Text Mining Handbok. – Cambridge: Cambridge University Press, 2007.

17. *Чехарин Е. Е.* Большие данные: большие проблемы // Перспективы науки и образования. 2016. № 3. С. 7–11.

18. *Hartigan J. A., Wong M. A.* Algorithm AS 136: A K-Means Clustering Algorithm // Journal of the Royal Statistical Society. Series C (Applied Statistics). 1978. Vol. 28. No. 1. P. 100–108.

19. *Defays D.* An efficient algorithm for a complete link method // The Computer Journal (British Computer Society). 1977. Vol. 20. No. 4. P. 364–366.

20. *Joachims T.* A statistical learning model of text classification with support vector machines // Proceedings of SIGIR 2001. P. 128–136.

21. *Glance N., Hurst M., Tomokiyo T.* BlogPulse: Automated Trend Discovery for Weblogs // Proceedings of WWW'2004.

22. *Allan J.* Topic detection and tracking: event-based information organization. – Kluwer Academic Press, 2002.

23. *Мельчук И. А.* Опыт теории лингвистических моделелей «Смысл ↔Текст». – М.: Языки русской культуры, 1999. 346 с.

24. *Хомский Н., Миллер Дж.* Введение в формальный анализ естественных языков // Кибернетический сборник / Под ред. А. А. Ляпунова и О. Б. Лупанова. – М.: Мир, 1965.

25. *Tsvetkov V. Ya.* Information Units as the Elements of Complex Models // Nanotechnology Research and Practice. 2014. Vol. 1. No. 1. P. 57–64.

26. *Tsvetkov V. Ya.* Information objects and information Units // European Journal of Natural History. 2009. No. 2. P. 99.

27. *Чехарин Е. Е.* Интерпретируемость информационных единиц // Славянский форум. 2014. № 2 (6). С. 151–155.

29. *Теньер Л.* Основы структурного синтаксиса. – М.: Прогресс, 1988. 656 с.

30. *Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л., Иомдин Л. Л.* Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–2005. – М.: Индриг, 2005. С. 193–214.

31. *Jurafsky D., Martin M.* Statistical Speech and Language Processing. – Prentice Hall, 1999.

32. *McDonald R.* Discriminative Learning and Spanning Tree Algorithms: PhD diss. – University of Pennsilvania, 2006. 240 p.

33. *Цветков В. Я.* Формирование дефиниций // Международный журнал прикладных и фундаментальных исследований. 2016. № 3 (часть 3). С. 503–504.

34. *Чехарин Е. Е.* Алгоритмы интерпретации данных дистанционного зондирования // Славянский форум. 2015. № 3 (9). С. 301–308.

35. *Tsvetkov V. Ya.* Information interaction // European Researcher. 2013. Vol. 62. No. 11-1. P. 2573–2577.

36. *Чехарин Е. Е.* Информационное взаимодействие в компьютерной лингвистике // Славянский форум. 2016. № 3 (13). С. 334–339.

37. *Цветков В. Я., Железняков В. А.* Инкрементальный метод проектирования электронных карт // Инженерные изыскания. 2011. № 1. С. 66–68.

38. *Ожерельева Т. А.* Ресурсные информационные модели // Перспективы науки и образования. 2015. № 1. С. 39–44.

39. *Цветков В. Я.* Дескриптивные и прескриптивные информационные модели // Дистанционное и виртуальное обучение. 2015. № 7. С. 48–54.

40. *Сигов А. С., Цветков В. Я.* Неявное знание: оппозиционный логический анализ и типологизация // Вестник Российской академии наук. 2015. Т. 85. № 9. С. 800–804.

Methods and algorithms of information interpretation

***Evgenii Evgen'evich Cheharin**, Deputy Head of the Center of Information Technologies MIREA Senior lecturer of the Department Institute of Information Technology Moscow Technological University (MIREA)*

This article describes the methods and algorithms of interpretation of information. This article describes the methods and algorithms for text analysis. Information interpretation is revealed as a technology model application information situations, information structures and information units. This article describes the different types of interpretation associated with the analysis of the text. This article describes the process of information exchange as one of the types of information interpretation.

Keywords: knowledge, information, information interpretation, text analysis, linguistics, computational linguistics, semantics, communication, information situation, information design, information units, information field, semantic field.

УДК 004.051

ПОСТРОЕНИЕ ПРОФИЛЯ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ ОТЕЧЕСТВЕННОЙ РАЗРАБОТКИ

***Лев Юрьевич Никулин**, аспирант,*
e-mail: i@dnbdrive.ru,
Московский университет имени С. Ю. Витте,
https://www.muiv.ru,
***Сергей Николаевич Маликов**, канд. техн. наук, ст. науч. сотр.,*
зам. генерального директора по научно-конструкторской работе,
e-mail: sergej.malikov@bk.ru,
ОАО «НИИ супер ЭВМ»,
http://www.super-computer.ru

DOI: 10.21777/2312-5500-2016-5-49-56