УДК 004.822 DOI: 10.17587/it.27.267-274

С. М. Салибекян, канд. техн. наук, e-mail: ssalibekyan@hse.ru,
Национальный исследовательский университет "Высшая школа экономики",
Московский институт электроники и математики, г. Москва

Объектно-атрибутный подход для семантического анализа естественного языка*

Описывается методика семантического анализа естественного языка (ЕЯ) и семантического поиска в нем, включающая в себя: основные этапы анализа ЕЯ, формат семантической сети для представления смысла текста, работу с полисемией (многозначностью) слов, семантико-синтаксическое согласование слов и т. д. Методика основывается на применении объектно-атрибутного принципа организации вычислений и структур данных, относящегося к классу dataflow (вычислительные системы с управлением данными).

***Ключевые слова**: семантический анализ естественного языка, семантическая сеть, граф-трансформирующая система, вычислительная система с управлением потоком данных, семантико-синтаксический анализ естественного языка*

Введение

Предметом рассмотрения данной статьи является методика семантического анализа естественного языка (ЕЯ) с помощью объектно-атрибутного (ОА) подхода к организации вычислительного процесса, разрабатываемая с 2013 г. Проблематика семантического анализа ЕЯ становится все более актуальной: объем информации в мире увеличивается экспоненциальными темпами, а широко применяемый сейчас поиск по ключевым фразам (или частотный анализ текста) уже не в состоянии дать релевантные (соответствующие замыслу пользователя) результаты; требуются все бóльшие объемы автоматического перевода иностранных текстов, но повсеместно используемые сейчас статические алгоритмы не обеспечивают необходимого качества; все в бóльших объемах требуется автоматическое реферирование текстов и т. д. Однако реализация подобных систем связана с большими трудностями, такими как многозначность (полисемия) слов ЕЯ, сложные синтаксические и семантические конструкции, необходимость применения многостадийного анализа и т. д. За всю историю компьютерной лингвистики было предложено достаточно много методик семантического анализа ЕЯ: генеративные грамматики Хомского, теория "Смысл <-> текст" Игоря Мельчука, концепция "семантический Web" и т.д. Хороший обзор методов синтаксического и семантического анализа ЕЯ приведен в работах [1, 2]. Появляются различные программные продукты, из которых особо следует отметить ABBYY Compreno [3], осуществляющий глубинный синтаксический и семантический анализы ЕЯ. Однако до сих пор нет общепризнанной методики семантического анализа ЕЯ, где на всех этапах разбора языка применяется единый подход к анализу и единые форматы данных. Одна из причин этого — разделение анализа синтаксиса и семантики ЕЯ, а ведь на практике было доказано, что без учета семантики невозможно провести полноценный синтаксический анализ, и ошибки, допущенные на изолированном этапе синтаксического анализа, затем вносят искажения на этапе семантического анализа. Еще одной причиной является тот факт, что получаемые в процессе синтеза итоговые информационные конструкции имеют ограничения по топологии и информационному содержанию (т. е. ограничения в описании смыслов). Например, генеративная грамматика Н. Хомского, основы которой изложены в работе [4], получила на Западе большую популярность. Однако оказалось, что она не очень хорошо подходит для задач анализа ЕЯ. Причины тому следующие: семантическая конструкция языка представляется в виде дерева (а не семантической сети), а такая структура не может описывать все многообразие семантических конструкций ЕЯ; существуют ограничения в описании смысла текста (узлы и дуги графа-дерева, описывающего синтаксис и семантику предложения, помечаются только метками, но не могут содержать структурированную информацию). По этой же причине оказались неэффективными и другие виды грамматик: грамматик-

*Работа выполнена при финансовой поддержке гранта РФФИ в рамках научного проекта №20-07-00958, реализуемого проектной группой "Децентрализованные данные" НИУ ВШЭ.

ки зависимостей, грамматика непосредственно составляющих, категориальная грамматика, вершинная грамматика составляющих (Head-driven Phrase Structure Grammar, HPSG) [5], лексико-функциональная грамматика (Lexical Functional Grammar, LFG), вероятностная контекстно-свободная грамматика (Probabalistik ContextFree Grammar, PCFG) и многие другие. Не подтвердила свою эффективность и LinkGrammar [6], являющаяся дальнейшим развитием генеративной грамматики, хотя данной тематике было посвящено множество исследований. Единственная грамматика, работающая со сложным информационным содержанием семантической сети — это атрибутная транслирующая грамматика Н. Хомского (АТ-грамматика) [7]. В отличие от обычной грамматики к символам языка приписывается один или несколько атрибутов. Атрибутом может быть: символ, число, строка и т. д. Грамматика описывает преобразование не только строки символов, но и информационного содержания символов. Однако АТ-грамматика применяется в основном для семантического анализа ЕЯ, так как она не в состоянии проводить синтез семантической сети, описывающей смысл текста.

Применение объектно-ориентированного (ОО) программирования (ООП) для распознания ЕЯ не дало ожидаемого эффекта. Так, ООП не нашла широкого применения даже в области баз данных, не говоря уже о базах знаний (надежды на то, что ООП составит конкуренцию реляционной парадигме, сошли на нет в 1990-х гг.) [8]. Причина тому — громоздкость и негибкость ООП [9]: класс как структура данных создается заранее и не изменяется во время вычислительного процесса. ООП успешно применяется при представлении статической семантической сети, например ОО-язык OWL, используемый в Semantics Web. Семантический же анализ предполагает синтез семантической сети с заранее неизвестной структурой.

Для анализа ЕЯ также применяются и методы машинного обучения, например, кластеризация, нейронные сети и т. д. Один из подходов к машинному обучению — это вероятностная комбинаторная категориальная грамматика (Probabilisti Combinatorial Categorial Grammars, CCG) [10], которая обеспечивает обучение вычислительной системы для задач кластеризации предложений ЕЯ. Однако такие методы неспособны осуществить полный семантический анализ, и ограничиваются только поверхностным — выделением сегментов текста (группы слов, связанной по смыслу глаголом или другим языковым элементом). Еще один недостаток таких методик анализа — необходимость обучения на больших обучающих выборках.

Достаточно удачной в области семантического анализа ЕЯ можно считать теорию "Смысл <-> текст" И. Мельчука. Теория [11, 12] предполагает анализ текста на нескольких уровнях, с постепенным переходом от одного уровня к другому: фонологический (уровень текста), поверхностно-морфологический, глубинно-морфологический, поверхностно-синтаксический, глубинно-синтаксический и семантический (уровень смысла). Морфологические и фонологические конструкции языка представляются с помощью линейных конструкций, синтаксическая — в виде дерева, остальные — виде семантической сети. Особенностью этой методики является то, что она объединяет в себе все стадии анализа языка от синтаксического до семантического. Недостаток такого подхода состоит в том, что на каждом уровне анализа конструкций языка используются свои форматы представления семантической сети, что вызывает необходимость применения сложных алгоритмов преобразования форматов представления информации при переходе с одного уровня на другой.

Теория "Смысл <-> текст" относится к классу систем, которые производят синтез семантической сети, описывающей смысл, заложенный в тексте. Это достаточно перспективный подход, поскольку такую сеть можно применять затем для различных целей: семантического поиска информации, трансляции на другой ЕЯ и т. д. В настоящем исследовании предлагается использование ОА подхода к организации структуры данных и организации вычислительного процесса [13], который также нацелен на синтез семантической сети (или ОА графа в терминологии ОА) в качестве результата анализа текста. Подход обеспечивает выполнение всех стадий анализа ЕЯ: представление правил анализа ЕЯ, построение семантической сети и поиск информации в ней, процесс преобразования текста в семантическую сеть, а также использование специализированного языка программирования для описания алгоритма преобразования и т. д. Благодаря тому, что все уровни анализа ЕЯ работают по единым принципам, значительно повышается эффективность системы в целом, так как сокращается расход вычислительных мощностей на преобразование формата данных при переходе от одного уровня к другому. ОА подход, относящийся к классу dataflow (вычисления с управлением потоком данных) [14, 15], обладает массой положительных качеств: объектный принцип построения программы и данных, способность синтеза и модификации семантической сети непосредственно во время вычислительного процесса, удобство организации параллельных и распределенных вычислений и т. д. Данные качества позволяют эффективно реализовывать системы семантического анализа ЕЯ. Еще одним преимуществом ОА подхода можно считать разработанный для него формализм, описанный в работе [16].

Итак, целью настоящего исследования является разработка методики семантического анализа текста на ЕЯ на базе ОА подхода к организации вычислительного процесса и структур данных. Отличительной особенностью такой системы являет-

ся единообразие представления данных и методики их обработки на всех стадиях анализа текста. Задачами исследования является разработка:

- форматов данных и методов их обработки;
- методики представления информации в семантической сети;
- методики анализа полисемических (многозначных) слов,
  а также
- выделение стадий анализа ЕЯ;
- выделение основных информационных примитивов и конструкций и их формализация для них;
- программная реализация прототипа системы анализа ЕЯ.

1. Синтез семантической сети

Данные на всех уровнях анализа представляют собой ОА граф. Эта конструкция несколько напоминает фреймовую структуру данных [17] или классовую структуру в ООП. Единицей данных является информационная пара (ИП), представляющая собой двойку $c = <a,l>$, где a — атрибут, представляющий собой идентификатор операнда (нагрузки); l — нагрузка (данные или указатель). ИП похожа на слот во фрейме. ИП собираются во множество, называемое информационной капсулой (ИК). Это — аналог фрейма. В нагрузках ИП могут располагаться указатели на другие ИК, и таким образом ИК могут объединяться в семантическую сеть, где каждая ИК является ее вершиной, а ссылки в нагрузках ИП выполняют роль семантических связей. В самом начале анализа текст представляется в виде ОА графа, представляющего собой список толкований слов, а далее в процессе многостадийного анализа список трансформируется в семантическую сеть, хранящую смысл текста. Преобразования ОА графа осуществляются с помощью граф-трансформирующей системы [18], а правила трансформации описываются посредством нотации ОА грамматики [19], специально разработанной для описания преобразований структуры и информационного содержания ОА графа. Граф-трансформирующая система, основанная на ОА подходе, осуществляет процесс синтеза семантической сети на всех уровнях (этапах) семантико-синтаксического анализа ЕЯ, что обеспечивает однородность анализа ЕЯ без необходимости преобразования данных при переходе от одного уровня анализа к другому.

Основой системы анализа ЕЯ на базе ОА подхода является семантико-морфологический словарь, хранящий описания лексем (слов). Описание одной лексемы представляет собой ОА список [20] всех возможных ее толкований, ведь, лексемы ЕЯ обладают свойством полисемии (многозначности). Каждое толкование лексемы — это совокупность трех связанных ссылками ИК: ИК с описанием морфологических свойств толкования лексемы (падеж, род, число и т. п.), ИК семантических свойств толкования лексемы и ИК с признаками для семантического согласования толкований слов (именно благодаря такому согласованию происходит выбор нужного толкования полисемичного слова исходя из его контекста). Семантическое согласование слов включает основные признаки для согласования: "контейнер", "вместилище", "абстрактный объект", "физический объект" и т. д. Семантические признаки толкования слова могут включить в себя и различные семантические связи с другими объектами.

Во время анализа текста осуществляется поиск описаний лексем в этом словаре, и из них формируется список толкований слов анализируемого текста (рис. 1, см. вторую сторону обложки): в ИК, содержащей атрибут POS (Part Of Spech — часть речи), находится перечисление синтаксических свойств толкования слова, а в нагрузке ИП с атрибутом SemProp (семантические свойства) находится указатель на ИП с описанием семантических свойств толкования слова. Наиболее оптимально формировать и затем осуществлять семантико-синтаксический анализ одного предложения текста, затем формировать список толкований для следующего предложения текста, анализировать его и т. д. После формирования списка толкований слов происходит многостадийное преобразование списка в семантическую сеть (семантический ОА граф), описывающую смысл текста (ОА граф текста). Эту семантическую сеть впоследствии можно будет использовать для поиска необходимой информации, автоматического перевода, автоматического реферирования и т. п. Преобразование в семантическую сеть происходит в несколько этапов (стадий), на каждом из которых осуществляется анализ определенной части речи, синтаксической или семантической конструкции — таким образом, подобный анализ ЕЯ можно назвать семантико-синтаксическим. Для русского и английского языков нами было выделено порядка 20 этапов. Анализ происходит от простого к сложному: сначала анализируются самые зависимые части речи (для русского и английского языков — это наречие степени), после анализа частей речи начинается анализ синтаксических и семантических конструкций. Приведем более подробное описание процесса анализа ЕЯ при ОА подходе.

На каждом этапе анализа происходит поиск второстепенных лексем определенного типа, а затем выполняется их "склейка" с близлежащим главным словом. Операция склейки подразумевает, что семантические свойства зависимого слова (синтаксической конструкции) присоединяются к семантическому описанию главного слова (синтаксической конструкции), а описание толкования зависимого слова удаляется из списка толкований слов. Так, на первом проходе анализа ЕЯ осуществляется "склейка" наречий степени с качественными наречиями и глаголами. Например,

для словосочетания "очень зеленый" в ИК с семантическим описанием слова "зеленый" добавится информационная пара (ИП) с атрибутом "степень" и нагрузкой, в которой будет храниться описание свойства степени "зелености" (для задания степени свойства можно, например, применить лингвистическую переменную [21]). Описание слова "очень" удаляется из списка описания слов. Данный процесс иллюстрируется на рис. 1 (см. вторую сторону обложки), где применяются обозначения: POS (Part Of Spech) — атрибут части речи), SemProp — атрибут семантических свойств толкования слова, DEGREE — атрибут степени, ADVERB_DERGREE — обозначение наречия степени, ADJECT — обозначение прилагательного.

2. ОА-грамматика

Для описания преобразований ОА графа была разработана нотация ОА грамматики. Формально ОА грамматика представляет собой четверку: OAG = {*A,L,P,G*}, где *A* — алфавит атрибутов; *L* — алфавит нагрузок ИП (алфавит включает в себя не только числа и строки, но и ссылки на ИК); *G* — ОА граф (список описаний лексем исходного языка); *P* — правила преобразования ОА графа (продукция). ОА грамматика, по сути, является разновидностью атрибутной графовой грамматики [18], где к вершинам графа приписываются один или несколько атрибутов. ОА грамматика же оперирует с ИП и ИК ОА графа: с помощью операций добавления ИП к ИК, удаления ИП из ИК, создания ИК, изменения атрибута или нагрузки ИП можно осуществлять модификацию как структуры, так и информационного содержания ОА графа.

Для ОА грамматики была разработана следующая нотация. Знак "=" в нотации описания правила преобразования обозначает ИП: слева от него указывается атрибут, справа — нагрузка. С помощью знаков "{"и"}" выделяется ИК. ИК или нагрузка могут быть именованы (имя в ОА вычислительной системе является мнемоникой указателя на ИК или нагрузку). В левой части правила преобразования (продукции) перед ИК может указываться тип множества ИК шаблона искомого подграфа. Если тип не указывается, то он по умолчанию устанавливается как "И". Множество "И" подразумевает, что все ИП из данной ИК шаблона поиска должны совпадать с ИП из ИК графа текста. Перечислим лишь некоторые типы ИК (их было выделено порядка 20): "ИЛИ" ("OR") — хотя бы одна ИП должна совпадать с ИП из ИК графа текста; "обратное И" ("INVERS AND") — все ИП из ИК графа-текста должны совпадать с ИП из ИК шаблона; "исключающее ИЛИ" ("XOR") — только одна ИП должна совпадать в ИК текста и ИК шаблона и т. д. С помощью аппарата множеств ИП можно, например, описывать поиск среди различных вариантов пространственно-временных отношений объектов физического мира [22].

Например, преобразование на рис. 1 (см. вторую сторону обложки) описывается с помощью продукции ОА-грамматики следующим образом:

{POS = ADVERB_DERGREE SemProp = temp} {POS = ADJECT SemProp = temp2} -> {POS = ADJECT SemProp = *temp}. (1)

Правая часть правила преобразования (1) задает шаблон искомого подграфа семантической сети, который необходимо преобразовать; после знака "->" идет описание трансформированного подграфа. Значком "=" разделяются атрибут и нагрузка ИП, фигурными скобками обозначается ИК. В данном случае описание наречия степени удаляется из списка толкований слов, а в ИК с описанием семантических свойств (адрес этого описания хранится в указателе "temp") добавляется в ИК с описанием семантических свойств прилагательного с атрибутом DERGREE (степень). Знак "*" обозначает конкатенацию (соединение) ИК, что было в нагрузке с ИК по указателю "temp". Второй проход анализа осуществляет "склейку" прилагательных с существительными. Например, при разборе выражения "очень зеленый стул" на первом проходе происходит склейка слов "очень" и "зеленый" (описание слова "очень" удаляется из списка, а его семантические свойства "склеиваются" со свойствами слова "зеленый"). После такой операции в списке остаются описания слов "зеленый" и "стул". На втором проходе осуществляется следующее преобразование: к описанию семантических свойств слова "стул" добавляется ИП с семантическими свойствами прилагательного, а именно, атрибутом "цвет" ("Color"), нагрузкой "зеленый" и степенью "очень", а описание слова "зеленый" удаляется из списка. В результате получается следующий семантический ОА граф: {POS = NOUN SemProp = {Color = GREEN DEGREE = Very}}, где GREEN — обозначение зеленого цвета, Very — обозначение степени "очень". При склейке к главному слову добавляется только ИК с описанием семантических свойств второстепенной лексемы, ИК с описанием ее морфологических свойств удаляется за ненадобностью, так как оно не понадобится для дальнейшего семантико-синтаксического анализа. На заключительных этапах анализа осуществляется разбор синтаксических конструкций с союзами, склейки существительных и глаголов и, наконец, склейка предикативных частей сложносочиненного или сложноподчиненного предложений. В конце концов в семантической сети остаются только ИК с описанием семантических свойств объектов и отношений между ними.

3. Обработка многозначности лексем

Анализ смысловых связей между предложениями в ОА системе осуществляется с применением

так называемого тематического словаря. Он представляет собой список, куда помещаются описания объектов, упоминающихся в тексте ранее. Например, при анализе первого предложения из фрагмента "В комнате стоял **стул**. **Стул** был зеленым." описание объекта "стул" передается в тематический словарь; при анализе же второго предложения описание слова "стул" будет обнаружено в тематическом словаре, и ссылка на это описание будет включена в список толкований слов. И впоследствии свойство "зелености" будет добавлено именно в описание семантических свойств того слова, которое встретилось в первом предложении. Описания объектов, попадающие в тематический словарь, через некоторое время удаляются из него, поскольку они становятся уже неактуальными по причине устаревания (читатель уже забывает о тех словах, которые он встретил в тексте достаточно давно) или иной причине. Процесс синтеза семантического графа из текста на ЕЯ приведена на рис. 2 (см. вторую сторону обложки).

ОА анализ успешно решает проблему работы с полисемией (многозначностью) слов. На рис. 3 (см. третью сторону обложки) представлен процесс обработки двух лексем с двумя толкованиями каждая. Пусть согласно правилам преобразования списка толкований слов необходимо провести "склейку" 2-го и 3-го толкований лексем (выделено штриховым прямоугольником в левой части рис. 3). Тогда происходит так называемое расщепление списка толкований. В результате образуется список из всех возможных альтернативных толкований этого фрагмента текста, содержащих все возможные комбинации толкований этих двух слов, точнее, четыре ветки с описанием двух лексем. На ветви, соответствующей склеиваемым толкованиям, осуществляется модификация списка толкований слов, согласно применяемому правилу преобразования (на рис. 3 справа внизу выделено штриховым прямоугольником). Альтернативные ветви толкований фрагментов текста могут удаляться из списка в процессе дальнейшего анализа, если обнаруживается синтаксическое или семантическое несогласование лексем в них. Таким образом, процесс преобразования списка толкований слов в семантическую сеть представляет собой процесс многократного создания и уничтожения альтернативных ветвей толкований фрагментов текста. Семантико-синтаксическое согласование представляет собой сравнение синтаксических и семантических атрибутов слов. Так, после союза "в" ("in") должно идти слово с атрибутом "вместилище" ("container"), а за союзом "на" ("Under") должно следовать слово с атрибутом "поверхность" ("surface") и т. д. В том случае, если не будет согласовано ни одно из толкований слова, текст считается несогласованным и не может быть проанализирован.

4. Семантическая сеть и методика поиска информации в ней

Теоретической основой семантического ОА графа является теория семантических падежей (валентностей) Ч. Филмора [23], в которой вводится перечень возможных типов связей между объектами, описываемыми в тексте. Так, Филлмором было выделено 9 ролей (связей), а Ю. Д. Апресяном — 25 [24], например, "объект", "субъект", "инструмент". Так, для предложения "Мальчик ударил большую змею палкой." будет синтезирован следующий ОА-граф: {**Объект** = мальчик **Субъект** = { **Объект** = змея **Свойство** = {**РазмерОтносительный** = большой}} **Действие** = ударить **Инструмент** = палка} (жирным выделены атрибуты ИП, описывающие семантические роли/валентности). Обзор основных семантических связей между объектами приведен в работе [25]. Нами во время разработки методики семантического анализа русского языка было выделено более ста различных семантических ролей (валентностей по Ю. Д. Апресяну), так как ролей, выделенных Филлмором, Апресяном и другими лингвистами, не хватало для описания семантики текста. И, скорее всего, в дальнейшем их список и еще пополнится.

Также был разработан формат семантической сети, представляющей смысл текста. Такая сеть должна, по сути, представлять собой универсальную сетевую базу данных, способную описывать реальный мир. В результате анализа была разработана структура сети, описанная в работе [22]. Сеть такого формата содержит три уровня: описания множеств объектов, описания свойств и состояний объектов и описание (в том числе и пространственно-временных) отношений объектов. В ней можно описать не только объекты, их свойства и отношения, но и множества, которые они образуют.

В ходе исследования также была разработана методика семантического поиска в тексте. В сетевых базах данных информационный поиск представляет собой поиск определенного подграфа. Для этого пользовательский запрос (вводится на ЕЯ) преобразуется в ОА граф и используется в качестве шаблона для поиска (рис. 4, см. третью сторону обложки). Назовем его шаблоном. Если в ОА графе текста находится подграф, изоморфный шаблону, то пользователю выдается сообщение об успехе поиска, причем в изоморфном подграфе должны совпадать с шаблоном не только узлы и связи, но информационное содержание в узлах.

Нами было выделено три типа вопросительных предложений которые могут выступать в качестве информационного запроса: от данных типов зависит формат представления графа-шаблона. Первый — общий вопрос, например "Мама уже ходила в магазин?". Ответ на данный вопрос может быть только "Да" или "Нет". На такой вопрос синтезируется простой граф-шаблон, и далее система проверяет,

имеется ли в графе-тексте подграф, совпадающий с шаблоном. Если хотя бы один такой подграф есть, то выдается сообщение "Да", иначе "Нет". Второй — запрос с "дыркой". "Дыркой" (рис. 4, см. третью сторону обложки) называется та часть шаблона, где подразумевается ответ на вопрос. Например, "Где Маша потеряла мячик?". Вопросительный союз "где" подразумевает локатив (т. е. место, где происходило действие). Поэтому в семантической сети шаблона на месте локатива будет находиться узел, обозначающий пустоту. Данный узел будет совпадать с любым локативом в графе текста. При нахождении подграфа ответом на вопрос как раз и будет тот локатив, который совпадает с "дыркой". Третий тип — вопрос с шаблоном ответа. Такие вопросы задают шаблон ответа (довольно часто в них применяются частицы "ли", "или", "и"), например, "Коля или Вася пошли гулять?". По результату такого запроса будет сформирован запрос с "дыркой" на месте главного действующего лица, и также будет задан шаблон ответа, представляющий собой список из двух имен. На такой вопрос может быть три варианта ответа: "Коля", "Петя" или ответ не найден. Последний вариант выдается в том случае, если в графе текста не будет найден подграф, изоморфный шаблону, или ответ на месте "дырки" не будет соответствовать шаблону.

Семантический поиск также необходим и в процессе синтеза семантической сети. Дело в том, что правая часть правила преобразования ОА графа описывает шаблон поиска подграфа, который необходимо модифицировать (правила сравнения информационного содержания узлов ОА графа описываются в левой части правила преобразования ОА графа). Поэтому наиболее рационально разделить трансформирующую граф систему на

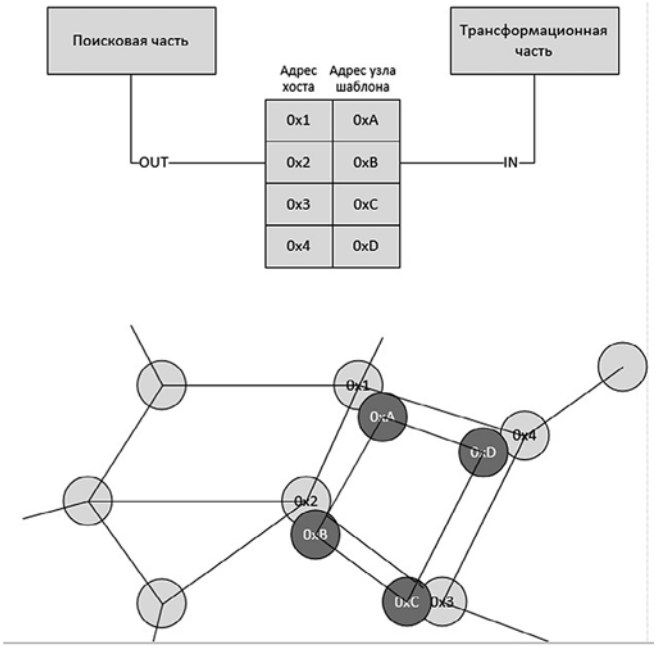

Рис. 5. Граф-трансформирующая система

две части: устройство поиска и устройство трансформации графа. Первое из них находит подграф, соответствующий шаблону из левой части правила преобразования, передает его на второе устройство (рис. 5). Но здесь возникает одна трудность: продукция описана относительно шаблона. Например, в правиле (1) ссылки temp и temp2 указывают на граф-шаблон, а трансформировать необходимо найденный подграф графа текста. Решением проблемы стало применение таблицы преобразования адресов — устройство поиска формирует ее после нахождения подграфа и передает устройству трансформации. В этой таблице прописываются соответствия адресов узлов шаблона и найденного подграфа. Во время преобразования подграфа устройство трансформации заменяет адреса узлов шаблона на адреса узлов подграфа и проводит модификацию структуры информационного содержания подграфа. Устройство трансформации графа было реализовано программно. Оно способно изменять ИП, вставлять новые ИП в ИК, удалять ИП из ИК, создавать новые ИК. Данный инструментарий дает возможность модификации любых ОА графов. Описание продукций осуществляется с помощью специализированного языка программирования (ОА язык). Посредством его команд можно задать шаблон поиска, все необходимые ссылки на ИК и ИП шаблона, а также задать действия, которые необходимо сделать в случае нахождения подграфа, соответствующего левой части продукции ОА грамматики.

Заключение

Разработанная методика ОА анализа ЕЯ реализована на практике — создана экспериментальная база знаний для анализа русского и английского языков, включающая в себя словарь из описаний порядка двух сотен слов и около ста правил преобразования списка исходных лексем. Это доказало работоспособность предложенной методики анализа ЕЯ. В дальнейших планах работы стоит расширение семантико-морфологического словаря и числа правил преобразования списка толкований слов, а также критериев семантического согласования лексем в экспериментальной базе данных. Кроме того, планируется добавление вероятностного анализа при семантическом согласовании слов.

Разработанная методика обладает однородностью представления и обработки данных на всех уровнях семантического анализа ЕЯ, что сводит к минимуму преобразования информационных конструкций при переходе от одного уровня анализа к другому. Все преобразования осуществляются с помощью одной и той же трансформирующей граф системы, обеспечивающей целостность данных при их изменении. Трансформирующая система способна изменять не только структуру семантической сети, но и ее информационное содержание. Для описания как исходных данных, так и правил преобразования

семантической сети существует такой инструмент, как ОА язык — специализированный язык программирования для описания данных и алгоритма работы ОА вычислительной сети. Сама же ОА система может быть реализована как программно, так и аппаратно. Еще одно преимущество ОА анализа ЕЯ в том, что он позволяет работать в одной парадигме буквально всем специалистам, занимающимся созданием системы семантического анализа ЕЯ: разработчикам аппаратной части, программистам, лингвистам, математикам.

В планы на будущее входят разработка алгоритма и программная реализация поиска изоморфного подграфа. Семантический поиск необходим не только для реализации пользовательских поисковых запросов, но и является неотъемлемой частью ОА трансформирующей граф системы, необходимой для осуществления левой части продукции правила ОА грамматики.

Эффективность работы создаваемой системы анализа ЕЯ требует проверки. Эту проверку можно будет провести путем статистического анализа работы системы на больших текстах (в настоящее время реализована только экспериментальная система анализа ЕЯ). Для этого будет необходимо программно реализовать алгоритм поиска изоморфного подграфа в семантической сети (в настоящее время применяются различные "заглушки", эмулирующие работу поисковой системы). Аналитические расчеты сложности алгоритма поиска показывают, что в худшем случае временная сложность алгоритма равна N^n, где N — число узлов в графе текста, n — число узлов в графе-шаблоне. Данная сложность полиномиальна при условии ограниченности n (т. е. когда поисковые запросы пользователя ограничены по объему текста, а на практике так и происходит), и поисковый алгоритм выполним на ЭВМ. Но ожидается, что на практике сложность будет намного меньше, чем данная оценка.

Результаты исследования найдут применение в области семантического анализа текста, семантического поиска, автоматического перевода и реферирования. Для автоматического перевода необходима разработка методики синтеза текста из семантической сети; однако в настоящее время исследования по данной тематике нами не проводятся.

Список литературы

1. **Смирнов И. В., Шелманов А. О.** Семантико-синтаксический анализ естественных языков часть I. Обзор методов синтаксического и семантического анализа текстов // Искусственный интеллект и принятие решений. 2013. № 1. С. 41—54.

2. **Смирнов И. В., Шелманов А. О., Кузнецова Е. С., Храмоин И. В.** Семантико-синтаксический анализ естественных языков. Часть II. Метод семантико-синтаксического анализа текстов // Искусственный интеллект и принятие решений. 2014. № 1. С. 11—24.

3. **Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.** Syntactic and semantic parser based on abbyy compreno linguistic technologies, computational linguistics and intellectual technologies // Proceedings of the International Conference Dialog. 2012. P. 90—103.

4. **Chomsky N.** Three models for the description of language // IRE Transactions on Information Theory. 1956. Vol. 2, N. 3. P. 113—124.

5. **Müller S.** Unifying everything: Some remarks on simpler syntax, construction grammar, minimalism and HPSG // Language. 2013. Vol.89, N.4. P.920—950.

6. **Link** grammar. 2013. feb. URL: http://www.abisource.com/projects/link-grammar/

7. **Ахо А., Сети Р., Ульман Дж.** Компиляторы. Принципы, технологии, инструменты. М.: Издательский дом "Вильямс", 2008. С. 383—398.

8. **Дейт К. Дж.** Введение в системы баз данных. М.: Издательский дом "Вильямс", 2005.

9. **Gabriel R.** Objects Have Failed: Notes for a Debate. (retrieved 17 May 2009). URL: http://www.dreamsongs.com/Files/ObjectsHaveFailed.pdf.

10. **Luke S.** Zettlemoyer and Mi hael Collins. Learning to map sentences to logical form: Structured lassication with Probabilisti Categorial Grammars // Proc. of 21th Conf. on Uncertainty in Articial Intelligence (UAI-2005). Edinburgh, Scotland, 2005. P. 658—666.

11. **Мельчук И. А.** Опыт теории лингвистических моделей "СМЫСЛ <—>ТЕКСТ". М.: Школа "Языки русской литературы", 1999.

12. **Мельчук И. А.** Русский язык в модели "СМЫСЛ <—>ТЕКСТ". Москва-Вена: Школа "Языки русской культуры", Венский славистический альманах, 1995.

13. **Салибекян С. М., Панфилов П. Б.** Объектно-атрибутная архитектура — новый подход к созданию объектных систем // Информационные технологии. 2012. № 2. С. 8—13.

14. **Data** flow computing: theory and practice / edited by John A. Sharp. Ablex Publishing Corp. Norwood, NJ, USA, 1992. 569 c.

15. **Milutinovic V., Trifunovic N., Salom J., Giorgi R.** The guide to dataflow supercomputing. USA: Springer; 2015.

16. **Салибекян С. М., Панфилов П. Б.** Вопросы автоматно-сетевого моделирования вычислительных систем с управлением потоком данных // Информационные технологии и вычислительные системы. 2015. № 1. С. 3—9

17. **Minsky M.** A framework for representing knowledge. MIT AI Laboratory Memo 306, June, 1974.

18. **König B., Nolte D., Padberg J., Rensink A.** A tutorial on graph transformation // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics10800 LNCS. P. 83—104.

19. **Salibekyan S., Panfilov P.** Linguistic Processor Based on Object-Attribute Grammar // The 6th International Conference on Analysis of Images, Social Networks, and Texts (AIST-2017), Moscow, Russia, July 27-29, 2017, CEUR Workshop Proceedings, Vol-1975. P. 134—145. URL: http://ceur-ws.org/Vol-1975/paper15.pdf.

20. **Салибекян С. М., Панфилов П. Б.** Анализ языка с помощью объектно-атрибутного подхода к организации вычислений // Программная инженерия. 2013. № 1. С. 9—16.

21. **Заде Л.** Понятие о лингвистической переменной и его применение к принятию решений. М.: Мир, 1976.

22. **Салибекян С. М., Петрова С. Б.** Объектно-атрибутная модель представления пространственно-временных отношений между объектами // Прикладная информатика. 2016. Т. 11, № 3 (63). С. 103—115.

23. **Филлмор Ч.** Дело о падеже // Новое в.зарубежной лингвистике. 1981. Вып.X. С. 369—495.

24. **Апресян Ю. Д.** Избранные труды. Т. 1. Лексическая семантика. М.: Языки русской культуры, 1995. 472 с.

25. **Найханова Л. В.** Основные типы семантических отношений между терминами предметной области // Известия высших учебных заведений. Поволжский регион. Технические науки. 2008. № 1. С. 62—71.

S. M. Salibekyan, Assistant Professor, e-mail: ssalibekyan@hse.ru,
National Research University Higher School of Economics, Moscow, Russian Federation

Object-Attribute Approach for Semantic Analysis of Natural Language

The article describes the methodology of semantic analysis of natural language (NL) and semantic search in it, which includes: the general stages of analysis of NL, the format of the semantic network for representing the meaning of the text, words polysemy analysis, semantic and syntactic agreement of words, etc. The method is based on the object-attribute principle of organization of calculations and data structures, belonging to the dataflow class.

Keywords: *Semantic analysis of natural language, semantic network, graph-rewriting system, dataflow computing system, semantic-syntactic analysis of natural language, graph-transforming system*

Acknowledgements: This work was carried out with the financial support of the RFBR grant in the framework of the scientific project No. 20-07-00958

DOI: 10.17587/it.27.267-274

References

1. **Smirnov I. V., Shelmanov A. O.** Semantiko-sintaksicheskij analiz estestvennyh yazykov chast' I. Obzor metodov sintaksicheskogo i semanticheskogo analiza tekstov, *Iskusstvennyj Intellekt i Prinyatie Reshenij*, 2013, no. 1, pp. 41—54 (in Russian).

2. **Smirnov I. V., Shelmanov A. O., Kuznecova E. S., Hramoin I. V.** Semantiko-sintaksicheskij analiz estestvennyh yazykov. CHast' II. Metod semantiko-sintaksicheskogo analiza tekstov, *Iskusstvennyj Intellekt i Prinyatie Reshenij*, 2014, no. 1, pp. 11—24 (in Russian).

3. **Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.** Syntactic and semantic parser based on ABBY Compreno linguistic technologies, computational linguistics and intellectual technologies, *Proceedings of the International Conference Dialog*, 2012, pp. 90—103.

4. **Chomsky N.** Three models for the description of language, *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 113—124.

5. **Müller S.** Unifying everything: Some remarks on simpler syntax, construction grammar, minimalism and HPSG, *Language,* 2013, vol. 89, no. 4, pp.920—950.

6. **Link** grammar, 2013, feb., available at: http://www.abisource.com/projects/link-grammar/

7. **Aho A., Seti R., Ul'man J.** Kompilyatory. Principy, tekhnologii, instrument, Moscow, Vil'yams, 2008, pp. 383—398 (in Russian).

8. **Dejt K. Dzh.** Vvedenie v sistemy baz dannyh, Moscow, Vil'yams, 2005 (in Russian).

9. **Gabriel R.** Objects Have Failed: Notes for a Debate. (retrieved 17 May 2009), available at: http://www.dreamsongs.com/Files/ObjectsHaveFailed.pdf.

10. **Luke S.** Zettlemoyer and Mi hael Collins. Learning to map sentences to logical form: Structured lassication with Probabilisti Categorial Grammars, *Proc. of 21th Conf. on Uncertainty in Articial Intelligence (UAI-2005),* Edinburgh, Scotland, 2005, pp. 658—666.

11. **Mel'chuk I. A.** Opyt teorii lingvisticheskih modelej "SMYSL <—>TEKST", Moscow, Shkola "YAzyki russkoj literatury", 1999 (in Russian).

12. **Mel'chuk I. A.** Russkij yazyk v modeli "SMYSL <—>TEKST", Moscow,Vena, Shkola "Yazyki russkoj kul'tury", Venskij slavisticheskij al'manah, 1995 (in Russian).

13. **Salibekyan S. M., Panfilov P. B.** Ob"ektno-atributnaya arhitektura — novyj podhod k sozdaniyu ob"ektnyh system, *Informacionnye Tekhnologii*, 2012, no. 2, pp. 8—13 (in Russian).

14. **Data** flow computing: theory and practice, Ablex Publishing Corp. Norwood, NJ, USA, 1992, 569 p.

15. **Milutinovic V., Trifunovic N., Salom J., Giorgi R.** The guide to dataflow supercomputing, USA, Springer, 2015.

16. **Salibekyan S. M., Panfilov P. B.** Voprosy avtomatno-setevogo modelirovaniya vychislitel'nyh sistem s upravleniem potokom dannyh, *Informacionnye Tekhnologii i Vychislitel'nye Sistemy,* 2015, no. 1, pp. 3—9 (in Russian).

17. **Minsky M.** A framework for representing knowledge, MIT AI Laboratory Memo 306, June, 1974.

18. **König B., Nolte D., Padberg J., Rensink A.** A tutorial on graph transformation, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics10800 LNCS*, pp. 83—104.

19. **Salibekyan S., Panfilov P.** Linguistic Processor Based on Object-Attribute Grammar, *The 6th International Conference on Analysis of Images, Social Networks, and Texts (AIST-2017)*, Moscow, Russia, July 27—29, 2017, CEUR Workshop Proceedings, ISSN 1613-0073, Vol. 1975, pp. 134—145, available at: http://ceur-ws.org/Vol-1975/paper15.pdf (in Russian).

20. **Salibekyan S. M., Panfilov P. B.** Analiz yazyka s pomoshch'yu ob"ektno-atributnogo podhoda k organizacii vychislenij, *Programmnaya Inzheneriya*, 2013, no. 1, pp. 9—16 (in Russian).

21. **Zade L.** Ponyatie o lingvisticheskoj peremennoj i ego primenenie k prinyatiyu reshenij, Moscow, Mir, 1976 (in Russian).

22. **Salibekyan S. M., Petrova S. B.** Ob"ektno-atributnaya model' predstavleniya prostranstvenno-vremennyh otnoshenij mezhdu ob"ektami*, Prikladnaya Informatika*, 2016, vol. 11, no. 3 (63), pp. 103—115 (in Russian).

23. **Fillmor Ch.** Delo o padezhe, *Novoe v.zarubezhnoj lingvistike*, Moscow, Progress, 1981, iss. H, pp. 369—495 (in Russian).

24. **Apresyan Yu. D.** Izbrannye trudy. Vol. 1. Leksicheskaya semantika, Moscow, Yazyki russkoj kul'tury, 1995, 472 p. (in Russian).

25. **Najhanova L. V.** Osnovnye tipy semanticheskih otnoshenij mezhdu terminami predmetnoj oblasti, *Izvestiya Vysshih Uchebnyh Zavedenij. Povolzhskij Region. Tekhnicheskie Nauki,* 2008, no. 1, pp. 62—71 (in Russian).

Рисунки к статье С. М. Салибекяна

«ОБЪЕКТНО-АТРИБУТНЫЙ ПОДХОД ДЛЯ СЕМАНТИЧЕСКОГО АНАЛИЗА ЕСТЕСТВЕННОГО ЯЗЫКА»

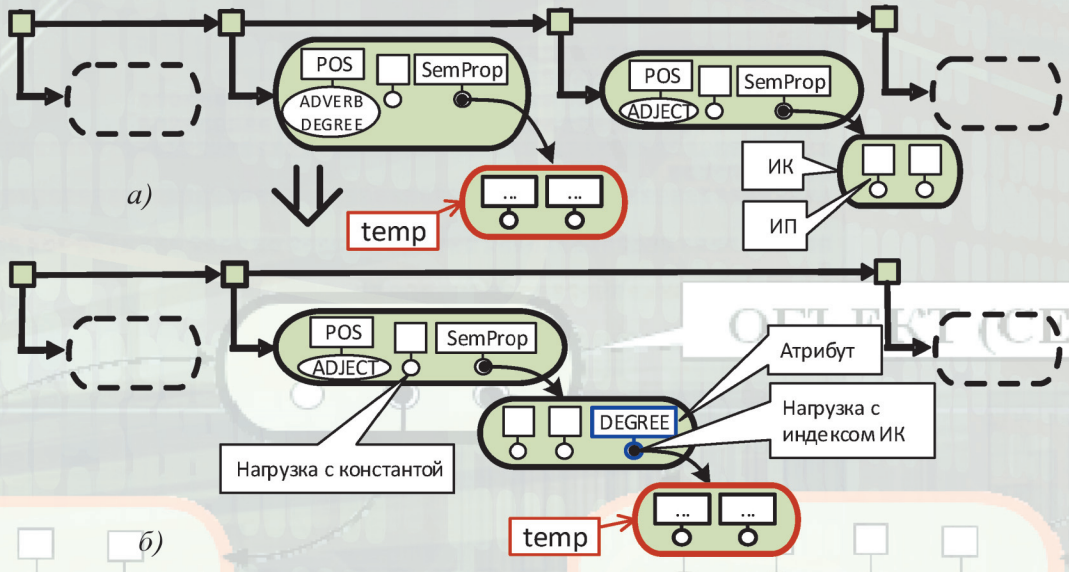

Рис. 1. Преобразование списка толкований слов

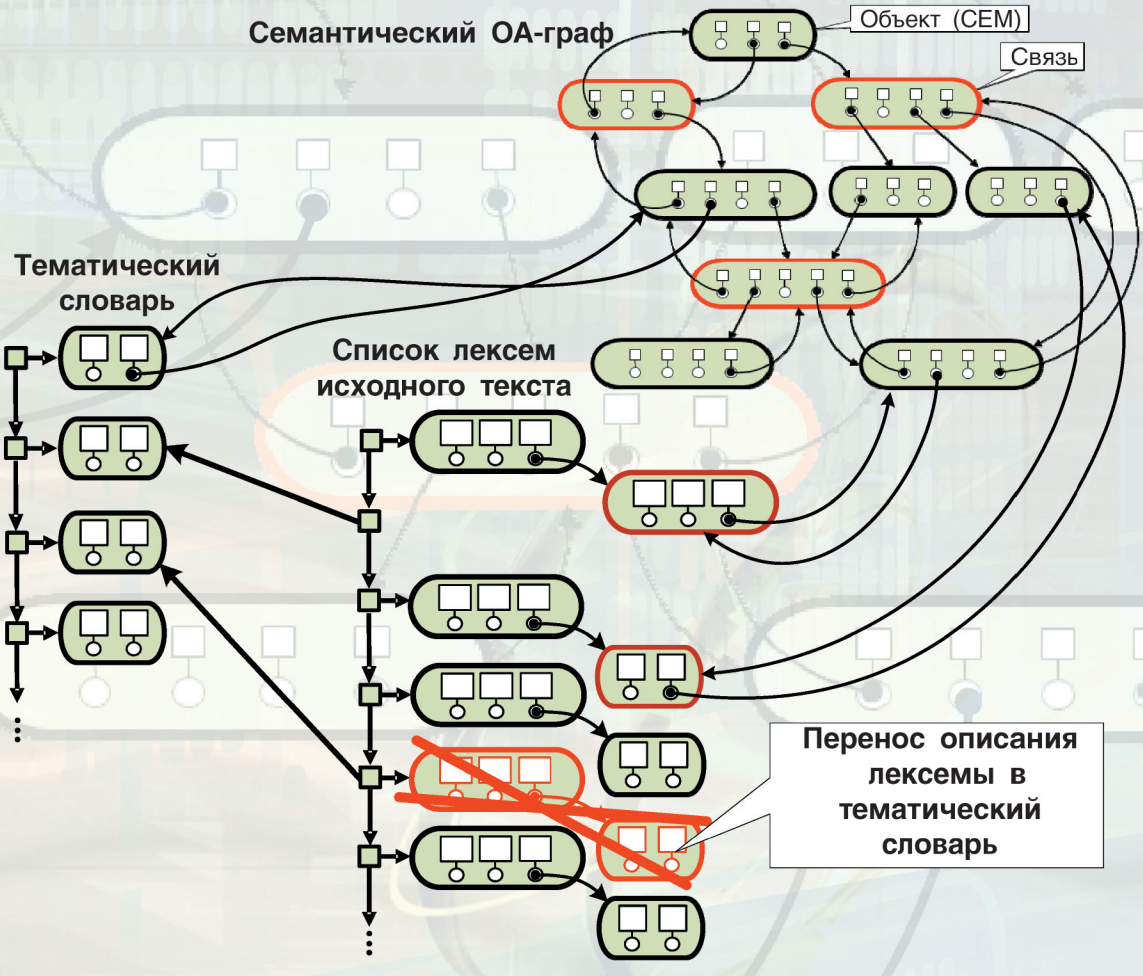

Рис. 2. Схема построения ОА-графа из списка лексем исходного текста

Рисунки к статье С. М. Салибекяна

«ОБЪЕКТНО-АТРИБУТНЫЙ ПОДХОД ДЛЯ СЕМАНТИЧЕСКОГО АНАЛИЗА ЕСТЕСТВЕННОГО ЯЗЫКА»

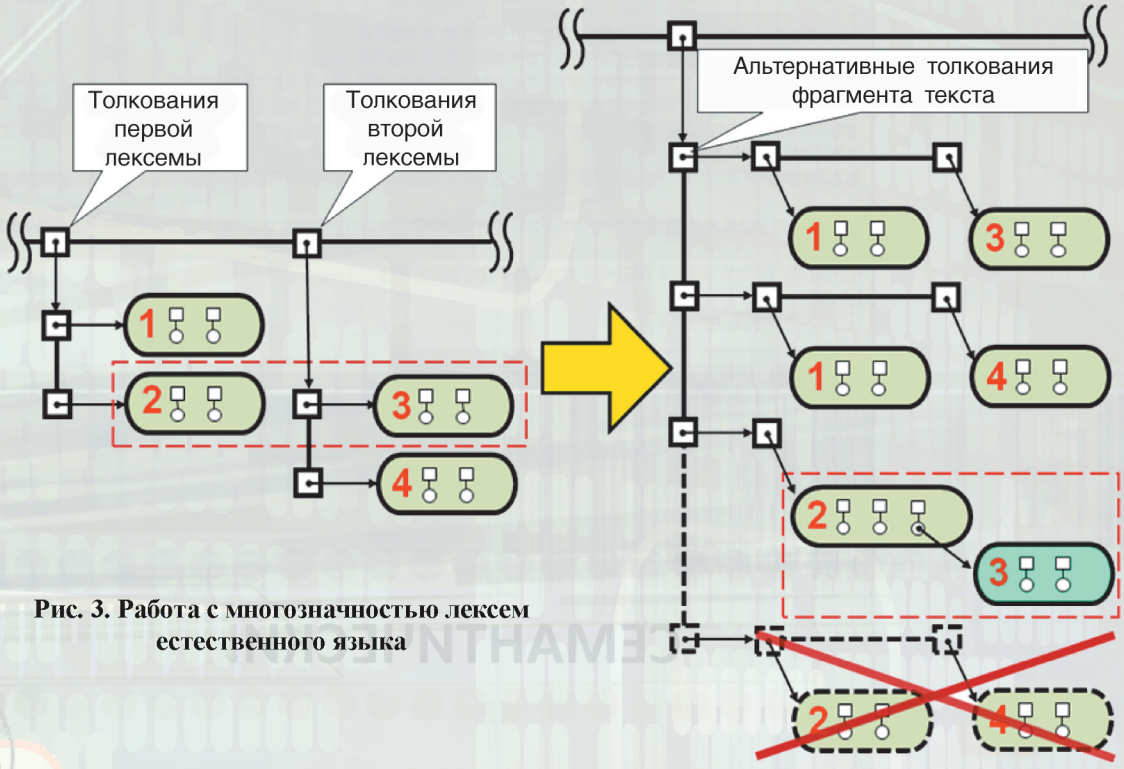

Рис. 3. Работа с многозначностью лексем естественного языка

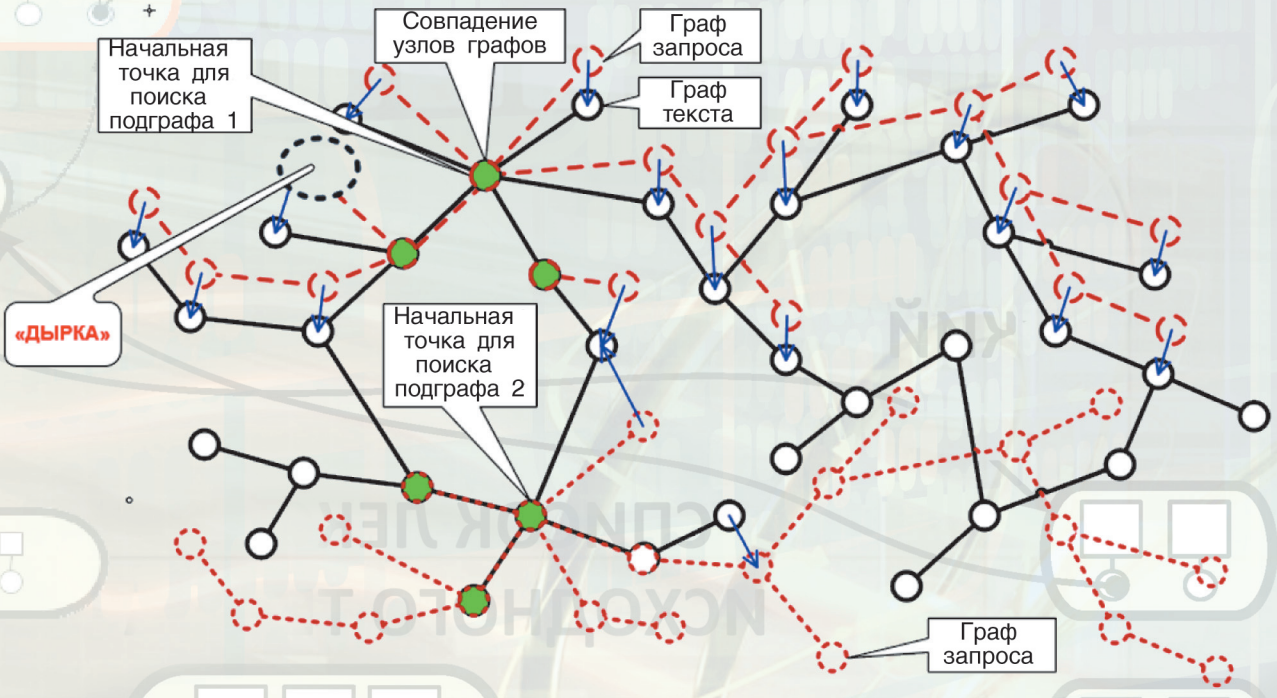

Рис. 4. Поиск подграфа в семантическом ОА-графе текста