Сравнительный анализ организации систем синтаксических парсеров

А.А. Харламов, Т.В. Ермоленко, Г.В. Дорохина

ВВЕДЕНИЕ

Автоматический анализ естественно-языковых текстов является востребованной технологией, которая находит применение в текстовых процессорах (например: Microsoft Word, OpenOffice.org Writer) и поисковых системах, системах реферирования, системах классификации и кластеризации текстов [1] и, наконец, в системах поиска дубликатов в текстах. Анализ текста микроблогов узла социальной сети широко используется для исследования психосемантического профиля пользователя [2], направленного на повышение эффективности предоставления контекстной рекламы, агитационных и прочих материалов. Технология автоматического анализа текста необходима также для создания, разметки и выравнивания корпусов параллельных текстов, которые широко используются системами памяти перевода.

Естественный язык является многоуровневой структурой, в которой чаще всего выделяют следующие уровни: фонетический; морфологический; лексический; синтаксический; семантический; прагматический. По этой причине, системы для автоматического анализа естественно-языковых текстов решают в процессе работы те или иные задачи анализа информации этих уровней. Наиболее применимы анализаторы трех уровней, а именно - морфологические анализаторы, синтаксические парсеры, анализаторы смысла. Причем, если говорить об анализе смысла отдельного предложения, то синтаксический анализ исчерпывает все вопросы выявления основной смысловой структуры предложения, будь то дерево зависимостей, или предикатная структура. В процессе семантического анализа целого текста также важную роль играет этап синтаксического анализа. Другими словами, качество синтаксического парсера определяет во многих случаях качество решения задачи, стоящей перед системой анализа текста.

Современные системы синтаксических парсеров [3-6] успешно реализуют диаметрально противоположные методы синтаксического анализа. Авторами было выполнено исследование лингвистических информационных технологий в области систем обработки информации, в результате которого проведен анализ организации синтаксических парсеров и трудностей, с которыми сталкиваются их разработчики. В результате чего была разработана архитектура системы синтаксического анализа в составе лингвистического процессора, осуществляющего семантико-синтаксический анализ предложений русско- и англоязычных текстов.

1. АНАЛИЗ ОРГАНИЗАЦИИ СОВРЕМЕННЫХ СИНТАКСИЧЕСКИХ ПАРСЕРОВ

Рассмотрим доступные данные об организации систем, принимавших участие в соревновании синтаксических парсеров, полученные по материалам форума «Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка» [7]. Среди них системы, использующие различные методы синтаксического разбора: грамматику зависимостей; грамматику составляющих; грамматику связей (Link grammar parser). Лучшие результаты показали ABBYY Syntactic and Semantic Parser, Парсер грамматики связей, ЭТАП-3, SyntAutom, SemSin.

1.1. ABBYY Syntactic and Semantic Parser

ABBYY Syntactic and Semantic Parser [3] при анализе текста использует словарь синтаксических парадигм слов, задающий правила употребления лексемы в зависимости от её класса, а также - дерево универсальных семантических значений и отношений между ними. Словарь синтаксических парадигм слов включает в себя данные о морфологической парадигме и о множестве «синтаксических уровней». «Синтаксический уровень» представлен множеством «синтаксических форм», каждая из которых определяет специфическую «синтаксическую конфигурацию», определяющую: грамматическое выражение, которое сопоставляется с грамматическим значением компонента ноль или более заполненных

«поверхностных слотов». Для каждого из таких выражений задается множество семантических слотов, которые рассматриваются как семантические интерпретации.

Судя по описанию, изложенному в [3], система основана на лексическом подходе, который использует грамматику управляемых вершинами фразовых категорий – Head-driven Phrase Structure Grammar (HPSG). По данным работы [8] этот метод использует:

– лексикон с иерархической организацией, где каждая лексическая единица языка описывается иерархической структурой свойств, содержащей грамматическую и семантическую информацию;

– унификацию «как базовый механизм построения синтаксической структуры».

Здесь унификацией, согласно [8] называется наиболее общий метод, позволяющий двум совместимым дескрипциям структуры свойств соединять информацию, которую они содержат, в одну (обычно большую) дескрипцию. Две дескрипции являются совместимыми в том случае, если они не содержат в своих структурах конфликтующих типов или разных атомарных значений одних и тех же свойств.

В HPSG вводится два универсальных синтаксических принципа, а именно:

– принцип вершины HFP (Head Feature Principle) Для любой фразовой категории, где определена вершина, значение свойства HEAD материнского узла и значение свойства HEAD дочернего узла должны быть унифицированы;

– принцип модели управления (The Valence Principle), означающий, что значения свойств SPR (спецификатор) и COMPS (комплементы) материнского узла идентичны значениям аналогичных свойств вершинного дочернего узла.

Аналогичным образом метод унификации используется и при построении семантической структуры (свойство SEM), для чего в грамматике определяются дополнительные принципы.

Базовый компонент грамматики HPSG в упрощенном виде состоит из четырех максимально общих синтаксических правил [I. Sag, T. Wasow, 1999]:

1. Правило комплемента вершины (Head-Complement Rule)

[phrase: COMPS <>] → H[word: COMPS <(1),…,(n)>] (1) … (n) , где n – идентификатор комплемента.

Фразовая категория может состоять из лексической вершины и следующих за ней комплементов; в частном случае список комплементов пуст.

2. Правило спецификатора вершины (Head-Specifier Rule)

[phrase: SPR <>] → (1) H[phrase: SPR <(1)>]

Фразовая категория может состоять из фразовой вершины и предшествующего ей спецификатора.

3. Правило модификатора вершины (Head-Modifier Rule)

[phrase] → H(1)[phrase] [phrase: MOD (1)]

Фразовая категория может состоять из фразовой вершины и следующего за ней совместимого фразового модификатора.

4. Правило сочинения (Coordination Rule)

[SYN (0); IND s0] → [SYN (0); IND s1] … [SYN (0); IND sn-1] [HEAD conj; IND s0] [SYN (0); IND sn], где семантическое свойство IND - индекс некоторой ситуации.

Любое число вхождений элементов с одинаковой синтаксической структурой (свойство SYN) могут быть соединены в один сочинительный элемент той же структуры.

Приведенный базовый компонент грамматических правил обладает тремя недостатками:

(a) жесткий линейный порядок составляющих в правой части правила, что не позволяет использовать такого рода правила в языках с относительно

свободным порядком синтаксических составляющих, каким является русский (то же относится и к структурным свойствам лексикона HPSG, где строго определен порядок следования комплементов лексемы, так [COMPS <NP, PP>] означает, что в линейной цепочке предложения именная группа, управляемая данной лексемой, должна стоять перед предложной);

(б) правила не способны анализировать слабо проективные структуры, грамматически допустимые во многих языках;

(в) абсолютная зависимость синтаксических правил от правильности и полноты структур свойств отдельно взятого словарного входа лексикона.

Характеристика метода HPSG [8] указывает на ряд трудностей, с которыми пришлось столкнуться разработчикам данной системы:

– трудоёмкость разработки лексикона для русского языка;

– «отсутствие разделения анализа на уровни и словари (морфологический, синтаксический и семантический) лишает архитектуру лексикона прозрачности»;

– «лексикализм и успешность работы грамматик, построенных на унификации, целиком зависят от полноты лексикона»;

– правила грамматики HPSG затруднительно использовать для языка «с относительно свободным порядком синтаксических составляющих, каким является русский»; они «не способны анализировать слабо проективные структуры, грамматически допустимые во многих языках».

Несмотря на указанные недостатки подхода лексикализма и недостатки базового компонента унифицирующей грамматики, необходимо признать большой экспериментальный потенциал построенной на HPSG модели для исследователей в области искусственного интеллекта. Метод анализа текста, используемый ABBYY Syntactic and Semantic Parser, очевидно, позволяет выполнять полный анализ предложений с высокой точностью. Однако данный метод использует базы данных, исчерпывающе описывающие перечень синтаксических конструкций, в которых употребляется лексема, и её соответствующие написания, а также дерево универсальных

семантических значений и отношений между ними. Себестоимость создания таких ресурсов и специфика коммерческой деятельности, в рамках которой они были созданы, позволяет предположить, что в свободном доступе эти ресурсы не появятся, и указывает на проблематичность воссоздания подобных ресурсов за обозримое время каким-либо научным коллективом, коммерческой организацией или научно-производственным объединением. Это делает невозможным реализацию метода отдельными научными коллективами, его использование в научных исследованиях и при создании инновационных технологий, связанных с обработкой текстов.

1.2. Парсер грамматики связей (LinkParser)

В отличие от HPSG, абстрактной и универсальной синтаксической теории ЕЯ, LinkParser с самого начала создавалась как аппарат для автоматической системы анализа предложения, что позволило авторам отойти от академических представлений, принятых в лингвистической традиции. Базовое отличие LinkParser состоит в том, что используемая модель анализа является контекстно-свободной грамматикой.

Каждая единица словаря грамматики описывается формулой, состоящей из соединителей (коннекторов connector). Коннектор состоит из имени типа связи (например, S – субъект, O – объект, CL – сегмент и т.д.), в которую может вступать рассматриваемая единица анализа, и суффикса, определяющего вектор направления соединения (‘+’ право-направленный коннектор и ‘-’ лево-направленный коннектор). Лево-направленный и право-направленный коннекторы одного типа образуют связь (соединение link). Так, два слова W1 и W2, имеющие словарные входы W1: A- и W2: A+, образуют соединение A в линейной последовательности W2W1, но не связаны в цепочке W1W2.

Язык формул, оперирующий коннекторами, использует четыре связки: оператор конъюнкции &, оператор дизъюнкции or, фигурные скобки {} для обозначения факультативности и неограниченность повторения @ (эквивалент оператора + Клини). Так, в формуле слова W: D- & {@A-}

выражение ‘@A-’ означает, что должна быть реализована связь с дескриптором A слева от W хотя бы один раз, и может повторяться неограниченное число раз; выражение ‘{@A-}’ означает, что связь A факультативна. Конъюнкция несимметрична для однонаправленных коннекторов и задает строгий порядок слов в предложении: в формуле W: A+ & B+ слово, реализующее соединение A, должно находиться ближе к W в линейной последовательности предложения, чем слово, реализующее соединение B, в той же последовательности. Для разнонаправленных коннекторов конъюнкция симметрична: формулы W: A- & B+ и W: B+ & A- эквивалентны.

Проблема избыточности словаря решается в системе LinkParser путем разбиения слов английского языка на 23 класса, где каждому такому классу приписывается своя формула. Разумеется, существует слова и подмножества слов-исключений, которые получают отдельную от основных классов формульную интерпретацию (к ним относятся, например, описание модальных глаголов или референциальных местоимений). Слова обобщаются в классы по селективным и субкатегориальным признакам. В ходе анализа словам в системе приписываются значения их базовых классов – селективных признаков (‘cat.n ran.v’).

Тип коннектора задается именем, где начальные заглавные буквы являются базовым дескриптором, а нижний составной индекс, как правило, задает значение граммемы, что позволяет косвенно проверять согласование или необходимое управление при установлении связи (например, ‘S+’ – существительное, ‘dogs ideas: Sp+’ – существительное во множественном числе, ‘dog idea: Ss+’ - существительное в единственном числе). Таким образом, могут соединяться либо равные коннекторы, либо два коннектора, один из которых выше уровнем: ‘Spa+’ может соединяться с ‘S-’, ‘Sp-’ или ‘Spa-’, но не с ‘Ss-’ или ‘Spb-’.

В анализаторе LinkParser используется около ста различных коннекторов, которые различаются преимущественно нижнем индексом. Число базовых дескрипторов при этом сравнительно небольшое.

В LinkParser вводятся следующие общие структурные ограничения:

– проективность, которая констатирует, что связи между словами в предложении не пересекаются;

– полнота связей, которая диктует, что все слова в линейной последовательности должны быть соединены между собой;

– порядок, означающий, что в линейной цепочке предложения должен выполняться порядок реализации соединений, заданный в формуле несимметричной конъюнкцией для однонаправленных коннекторов;

– исключение, суть которого заключается в том, что для одной пары слов не может быть проведено больше одной связи.

Нетрадиционный характер модели, используемой анализатором LinkParser, заключается также в том, что разработчики отказались от системы составляющих, столь популярной для представления синтаксической структуры английского языка. Они используют формализм, в концептуальном плане близкий к теории зависимостей, описанной в работах создателя лингвистической теории «Смысл ↔ Текст» И. Мельчука. В отличие от деревьев зависимостей, бинарные связи, строящиеся LinkParser, не содержат вершины и не имеют направления.

Используя информацию о селективных дескрипторах, приписанную терминальным единицам предложения, а также тип коннекторов, маркирующих соединения, можно транслировать построенную LinkParser проективную структуру (linkage) в классическое дерево зависимостей. Такая же трансляция возможна, когда рассматривается вложение соединений в дерево непосредственных составляющих, определенных в выходной структуре анализатора.

Алгоритм синтаксического анализа в процессоре LinkParser основан на методе динамического программирования [8]. Его суть в том, что в ходе

анализа предложения все множество синтаксических единиц, входящих в предложение S, разбивается на перекрывающиеся подмножества (подзадачи) с сохранением исходного линейного порядка. В рамках такого порядка каждое такое подмножество является (в случае успешного построения связей между его элементами) поддеревом полного графа S и называется частичным решением (partial solution).

Для ускорения работы алгоритма синтаксического анализа в LinkParser предложен ряд решений, в том числе и эмпирических. Перед началом анализа устанавливается фильтр, удаляющий все дизъюнкты, содержащие «непарные» коннекторы: если для некоторого коннектора X- дизъюнкта D, принадлежащего словоформе W, слева в линейной последовательности S не найдено X+, то D будет удален, аналогично для право-направленного коннектора X+. Другой метод ускорения вводит эмпирическое ограничение на длину возможного соединения в зависимости от типа связи. Несмотря на применяемые методы оптимизации, тестирование системы показывает, что в большинстве случаев анализ сложных предложений, длина которых превышает 25-30 слов, приводит к комбинаторному взрыву. Результатом работы анализатора в этом случае становится “панический” граф, как правило, случайный вариант синтаксической структуры, зачастую несвязанной.

К сожалению, использование грамматики LinkParser для русского языка представляется невозможным по ряду причин. К их числу относятся следующие:

– основная идея грамматики, а именно - использование лево- и право-ветвящихся коннекторов, теряет свою силу для языка с относительно свободным направлением связей (особенно для глагольных групп);

– если предположить, что каждое возможное направление связи можно маркировать отдельным типом коннектора, то в этом случае резко возрастет как число базовых коннекторов, так и число дизъюнктов словоформ, что негативно сказывается на скорости работы процессора;

– избыточность и значительно возрастающий объем словаря, которые возникают в силу морфологической развитости флективного языка, когда каждая морфологическая форма описывается отдельной формулой, где нижний индекс входящего в нее коннектора должен будет обеспечить процедуру согласования, что приведет к усложнению составления коннекторов и к увеличению их общего числа в грамматике.

Тем не менее, LinkParser по праву считается одним из самых элегантных и детально проработанных решений задачи синтаксического анализа английского языка, а лингвистическая прозрачность грамматики и программная реализация алгоритмов на языке С придают процессору полную завершенность.

1.3. Синтаксический парсер лингвистического процессора ЭТАП-3

Синтаксический парсер лингвистического процессора ЭТАП-3 [12] определяет синтаксическую структуру фразы в виде дерева зависимостей, которое строится с помощью специальных правил (синтагм). Этих правил для каждого из рабочих языков системы (в данном случае - русского и английского) насчитывается несколько сотен. Все они бинарны. Этот факт означает, что любая синтагма позволяет связать некоторым синтаксическим отношением два слова фразы, если все условия этой синтагмы, описывающие контекст данной пары слов во фразе, выполнены. Более строго, синтагма связывает синтаксическим отношением не слова фразы, а некоторую пару омонимов этих слов, если они представлены в начале синтаксического анализа несколькими (морфологическими и/или лексическими) омонимами. Таким образом, омонимы слов фразы могут связываться синтаксическими отношениями независимо друг от друга.

В результате работы синтагм на первом этапе синтаксического анализа возникает граф гипотетических синтаксических связей (синтаксических гипотез). На дальнейших этапах синтаксического анализатора посторонние связи различными средствами отфильтровываются. Из графа синтаксических гипотез выделяется дерево синтаксической структуры фразы. Иными

словами, в основе алгоритма синтаксического анализа системы ЭТАП-3 лежит так называемый “фильтровый метод”.

Проблемные вопросы, возникающие при работе парсера заключаются в следующем.

1. Посторонние интерпретации. Рассмотрим это на примере предложения *Что делает правительство?* слово *правительство* здесь является субъектом, подлежащим, а слово *что* – прямым дополнением глагола *делает*. С точки же зрения парсера это предложение допускает и другие интерпретации, например:

- слово *что* может интерпретироваться как подлежащее, а *правительство* – как дополнение при глаголе *делает*;
- слово *что* может интерпретироваться как союз, вводящий неполное предложение.

2. Избыточность. Если лингвист, обслуживающий систему, встречает в тексте синтаксическую конструкцию, не учтенную в синтагмах, то ему достаточно подправить одну из соответствующих синтагм или создать новую, чтобы возникло недостающее синтаксическое отношение. Однако часто бывает, что некоторая языковая конфигурация (скажем, последовательность словоформ, принадлежащих определенным лексико-грамматическим классам), будучи погружена в другие контексты, образует другую синтаксическую конструкцию и должна анализироваться уже иначе. Предусмотреть все эти контексты при написании синтагм, по-видимому, невозможно в принципе. Отсюда следует, что синтагмы неизбежно будут порождать в ряде случаев лишние, неверные синтаксические гипотезы. Как показывает опыт эксплуатации парсера ЭТАП'а-3, для больших фраз количество гипотез может достигать величины 20-30 n, где n – число слов фразы.

Система ЭТАП -3 использует следующие лингвистические ресурсы.

1. Корпус текстов. Система ЭТАП-3 находится в экспериментальной эксплуатации уже довольно давно, были синтаксически размечены десятки

тысяч фраз из разного рода текстов (сейчас в корпусе текстов около 37 000 фраз). Все синтаксические структуры этих фраз сначала «начерно» строились системой ЭТАП-3, а затем вручную редактировались специалистами-лингвистами.

2. Для преодоления избыточности и оптимизации процесса выделения правильной синтаксической структуры из графа гипотетических связей применяют ранжирование синтаксических гипотез, порождаемых синтагмами, с помощью внедрения в правиловую систему обучающего статистического компонента. Таким образом, синтаксический анализатор ЭТАП-3 использует эмпирическую статистику, порожденную лингвистом-экспертом, который извлекает уроки из работы пусть несовершенной, но живой синтаксической системы и производит все более тонкую настройку этой системы. Этим достигаются две цели: расширяются рамки возможностей построенной лингвистом действующей модели языка; точнее определяются границы этих возможностей. Это приводит к тому, что правильная структура выделяется первой или одной из первых.

1.4. SyntAutom

SyntAutom [4] – система, основанная на правилах, построенных вручную. Система использует:

– морфорологический словарь;

– словарь валентности глаголов (создан вручную, насчитывает 12 тыс. глаголов);

– базу частотности морфологическиих интерпретаций слов, базу частотности бинарных отношений зависимости между парами лексических единиц, (вычисляются по большому автоматически размеченному корпусу);

– эмпирические веса, добавляемые, когда автомат пересекает некоторые состояния автомата.

Отличительная черта этой системы в том, что она имеет тенденцию непосредственно связывать значимые слова, тогда как вспомогательные слова переносятся на более низкие уровни дерева зависимостей.

Ограничения и особенности работы системы [4]:

– связи, которые отражаются в дереве зависимостей в ряде случаев описывают зависимости семантические, а не синтаксические;

– предлоги система подчиняет существительным, которыми они управляют;

– главный предикат подчиненной клаузы считается подчиненным главному предикату главной клаузы (клауза - простое предложение в составе сложного);

– считается, что предикат может быть выражен только глаголом, предложения без предиката разбираются как бессвязные;

– жертвуют некоторыми потенциальными разборами для отсечения ложных анализов и роста комбинаций;

– конструкции с выразительным союзом «и» система не разбирает;

– допускается контекстная субстантивация прилагательных (*«Коричневый идёт вашим глазам»*);

– выполняют винительно-родительную трансформацию в отрицательных предложениях (*«Я вижу собаку ->Я не вижу собаки »*).

Преимущества применяемого в системе [4] метода:

– синтаксическая и морфорологическая неоднозначность разрешаются одновременно в рамках унифицированного подхода;

– явное описание переходом автомата обеспечивает гибкий способ управления процессом парсинга текста;

– состояния автомата, реализующего парсинг текста в данной системе, зачастую предоставляет больше информации, чем в состоянии обеспечить контекстно-свободная грамматика;

– к системе можно легко добавлять «локальные» функции, которые вызываются только в специфических условиях.

Системе присущи общие трудности, характерные для большинства систем, основанных на правилах:

– трудно согласовать эмпирические веса с весами, которые формируются статистической моделью;

– существуют пределы, за которыми трудно увеличить грамматическое покрытие, что обусловлено комбинаторным ростом вариантов синтаксического разбора и падением точности синтаксического анализа.

1.5. SemSin

SemSin [6] – это семантико-синтаксический анализатор, в задачи которого входит снятие частеречной и морфологической омонимии, построение синтаксического дерева зависимостей и частичное снятие лексической неоднозначности. Система создана небольшим коллективом в «достаточно сжатые» сроки.

Использует следующие лингвистические ресурсы.

1. Словарь и классификатор В. А. Тузова, созданный на основе морфологического словаря А.А. Зализняка. При определении семантики использовался словарь С.А. Кузнецова. В нём каждая лексема содержит морфологические характеристики, а также номер своего класса и модели управления слов (актанты вызываемых ею лексем в виде падежей или предлогов с соответствующими падежами). Словарь содержит общеупотребительные слова, названия и имена собственные.

2. База фразеологизмов обеспечивает разбор трех типов словосочетаний: неизменяемых (*несмотря ни на что*, *вдалеке от*), с изменяемым первым словом (*гвоздь программы*) и полностью изменяемых (*белая ворона*).

3. База предлогов, хранящая классы существительных, с которыми они взаимодействуют, и названия связей с хозяевами предложных групп («хозяин» - главное слово в синтаксической группе).

4. База продукционных правил (около 210).

В процессе анализа предложения система сегментирует его, устанавливает главное слово сегмента («центр сегмента»), может объединять

сегменты, подчинять их. Исходное предложение разбивается по знакам пунктуации на отдельные сегменты. Каждому сегменту при этом присваивается свой тип, исходя из наличия/отсутствия подчинительного союза или глагольной формы. После завершения работы сегментации проводится построение именных и предложных групп внутри сегментов Таким образом в первой фазе синтаксического анализа определяется топологическая структура предложения (выделение глагольных групп и сегментов), во второй фазе происходит выделение фразовых категорий в пределах, определенных границами сегментов. Следовательно, в первой фазе анализ предложения проводится «сверху вниз», во второй – «снизу вверх», но на фрагментах меньше длины предложения. Следует отметить, что идея необходимости разделения сегментационного и непосредственно синтаксического (в смысле установление связей между отдельными словами) анализа – параллельное построение сверху и снизу структуры предложения – существовала в московской прикладной лингвистике еще в 1970-ые годы. Такая стратегия позволяет значительно снизить объем необходимых для ее реализации вычислений.

В описание процессора не включена информация о построении или разрешении синтаксической омонимии на уровне сегментов, то есть возможность рассмотрения структурных вариантов сегментации предложения с разными границами сегментов. Нет также упоминания о сочинении предикатов – важной составляющей анализа для правильного определения границ сегментов. Следует также отметить, что время анализа линейно зависит от длины предложения.

1.6. Анализ ответов систем: проблемные точки

Организаторами форума «Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка» в работе упомянутых систем выделены следующие «проблемные точки разбора» [7].

1. Если в предложении находится несколько потенциальных хозяев, то системы выбирают либо линейно предшествующее существительное,

либо вершинный глагол, либо ближайший финитный глагол в дереве. Однако не все такие варианты будут семантически оправданы.

2. Большинство систем не смогло справиться с примером, в котором присутствуют три однородных определения вида X, Y и Z, относящихся к одному существительному.
3. Многие системы ошибаются при обработке идиоматических конструкций «малого синтаксиса», если срабатывают альтернативные, характерные для русского языка шаблоны, (Например, в предложении *Что такое обучение* – ошибочно приписывают атрибутивную связь в паре *обучение → такое*).
4. Часто наблюдаются трудности, связанные с нахождением вершины в предшествующей клаузе.
5. В сложных предложениях, безусловно, ошибок больше. Часто наблюдаются трудности с нахождением вершины в предшествующей клаузе. Могут оставаться незамеченными вершины–существительные или связки типа *есть*.

В числе наиболее частых случаев, в которых у систем наблюдаются расхождения, отмечены [7]: «неоднословные союзы и предлоги, сложные слова с дефисным написанием; связь между однородными членами, между главной и подчиненной клаузой, между сочиненными клаузами (включая интерпретацию союзов), союз в начале главной клаузы; глагол-связку с инфинитивами, именами, прилагательными, причастиями; группы с количественными и порядковыми числительными (включая предложные и с модификаторами типа более, минимум); связь подлежащего с именным сказуемым; связь в группах вида ‘прилагательное + прилагательное + существительное’».

2. ОРГАНИЗАЦИЯ АВТОРСКОЙ СИСТЕМЫ СИНТАКСИЧЕСКОГО АНАЛИЗА ТЕКСТОВ

В ходе выполнения исследований по рассматриваемой теме авторами предложен единый подход к обработке неструктурированных текстов на русском и английском языках. В рамках этого подхода объединены в едином

комплексе морфология и синтаксис, а также утилиты статистического анализа текста с целью создания программного обеспечения для понимания неструктурированной текстовой информации. Создана система семантико-синтаксического анализа предложений русского и английского языка, которая позволяет выделить предикатные структуры предложений текста и построить деревья синтаксического подчинения предложений. На всех этапах работы системы используется многоуровневое представление текста (слова, словосочетания, предложения), допускающее несколько интерпретаций элементов текста, каждое из которых сохраняется. Также представление несколько избыточно. Однако оно даёт возможность изменить интерпретацию отдельных структурных элементов текста (лексические и нелексические единицы, словосочетания, предикатные структуры предложений) или их совокупности, если такая необходимость возникнет на более поздних этапах обработки текста (синтаксический, семантический, прагматический). Это обстоятельство делает лингвистический процессор более гибким и надёжным. Предложенное представление структурных элементов текста позволяет также отследить употребление в тексте неизвестных системе лексических единиц (регистр, наличие потенциальных словоизменительных форм), что даёт возможность выделить кандидатов на роль имен собственных, аббревиатур и сокращений.

Рассмотрим организацию отдельных модулей системы более подробно.

2.1. Модуль морфологического анализа

В ходе работы модуля морфологического анализа лексических единиц текста эти единицы последовательно подвергаются следующим видам анализа.

1. Декларативный морфологический анализ [9], использующий такие базы словоформ, как
   - общеупотребительные слова;
   - фамилии, имена и отчества.

2. Морфологический анализ слов с дефисным написанием [10] на основе декларативного морфологического анализа и правил согласования частей составного слова.
3. Бессловарный морфологический анализ [11], результаты которого фильтруются с помощью словарей начальных форм:
   - словарь географических названий;
   - пользовательский словарь имен собственных;
   - пользовательский словарь административных названий;
   - пользовательский словарь общеупотребительных слов.
4. Бессловарный морфологический анализ, результаты которого уточняются на основе анализа частоты употребления в тексте «несловарных» лексических единиц с учетом регистра и потенциальных словоизменительных форм.

Анализ лексических единиц выполняется в изложенной последовательности. В случае, если на некотором шаге получена одна или несколько интерпретаций слова (леммы и морфологической информации), то последующие шаги не выполняются. На шаге 3 «Бессловарный морфологический анализ, результаты которого фильтруются с помощью словарей начальных форм», интерпретациями слова считаются только те результаты бессловарного морфологического анализа, которые принадлежат хотя бы одному словарю начальных форм. Используемый при этом список словарей начальных форм является открытым. Это означает, что наряду с приведенными словарями начальных форм могут использоваться аналогичные словари для специфических предметных областей.

Средства декларативного морфологического анализа общеупотребительных слов программно реализованы и апробированы для слов русского и английского языка. Используемые на остальных шагах средства морфологического анализа связаны с анализом слов русского языка.

Следует отметить положительный эффект использования метода морфологического анализа, реализованного в системе. Он позволяет сочетать

средства декларативного и бессловарного морфологического анализа, правил морфологического анализа слов с дефисным написанием при условии сохранения всех интерпретаций слов. Такой вывод основывается на анализе результатов форума «Оценка методов автоматического анализа текста: морфологические парсеры русского языка» [12]. Организация одной из участвовавших на этом форуме систем (РДМА_ИПИИ) была принята в качестве основы для системы, которая рассматривается далее. Используемые в ней словарные базы откорректированы после устранения ошибок и неточностей, обнаруженных в РДМА_ИПИИ, и дополнены после совершенствования алгоритмов морфологического анализа слов с дефисным написанием.

Результаты бессловарного морфологического анализа слов русского языка в системе РДМА_ИПИИ с большой вероятностью содержали правильную интерпретацию отдельного слова. Однако они содержали и ряд «побочных» интерпретаций, не являющихся словами русского языка. Это создавало определенные трудности, поскольку рост количества интерпретаций слова замедляет анализ текста на последующих этапах его обработки. Перечень результатов бессловарного морфологического анализа удалось в значительной мере сократить по сравнению с системой РДМА_ИПИИ за счёт использования словарей начальных форм (специфических и пользовательских), а также уточнения результатов вероятностными методами.

Для реализации предложенного метода потребовалось создать средства декларативного морфологического анализа слов английского языка. Метод декларативного морфологического анализа слов состоит в явном задании парадигмы слова как набора словоформ, каждая из которых представлена написанием и морфологической информацией. При этом впервые сделана попытка описания слов английского языка с помощью предложенной системы представления отдельных значений грамматических категорий и их сочетаний. Набор значений грамматических категорий, описывающих

некоторое слово, в дальнейшем будем называть морфологической информацией (МИ).

Морфологическая информация хранится в виде набора битовых полей, что отвечает требованиям компактности, однозначности и простоты извлечения отдельных морфологических характеристик словоформы. Таблица 1 содержит перечень обозначений с помощью чисел и макроопределений, используемых в системе для задания морфологической информации слова английского языка. Эти обозначения значений подобраны так, чтобы совпадали одинаковые значения одних и тех же категорий для русского и английского языка. В столбце «Совпадает с русским» такие обозначения помечены символом ‘+’. Морфологическую информацию словоформы формируем применением побитового «или», например: _Noun_en | _Nominative_en | _Singular_en.

Значение определенной грамматической категории для слова по его морфологической информации находятся с помощью масок категорий (см. табл. 2). Отметим, что численные значения масок категорий для русского и английского языка совпадают.

Таблица 1

Значения грамматических категорий для английского языка

Обозначение в программе		***Грамматические категории***		***Совпадает с русским***
Число	Макроопределение	Категория	Значение	
0x00000001	_Nominative_en	Падеж	Именительный	+
0x00000002	_Prityag_en		Притяжательный	–
0x00000003	_Objekt_en		Объектный падеж	–
0x00000004	_PritagAbsol_en		Притяжательный абсолютный	–
0x00000008	_Masculine_en	Род	Мужской	+
0x00000010	_Feminine_n		Женский	+
0x00000018	_Neuter_en		Средний	+
0x00000020	_Singular_en	Число	Единственное	+
0x00000040	_Plural_en		Множественное	+
0x00000080	_Pres_en	Время	Настоящее	+
0x00000100	_Future_en		Будущее	+
0x00000180	_Past_en		Прошедшее	+
0x00000200	_FaceFir_en	Лицо	1-е	+
0x00000400	_FaceSec_en		2-е	+
0x00000600	_FaceThi_en		3-е	+
0x00000800	_Active_en	Залог	Действительный	+
0x00001000	_Passive_en		Страдательный	+
0x00002000	_ComparativeFormOfAdj_en	Степень сравнения	Сравнительная	+
0x00004000	_ExellentFormOfAdj_en		Превосходная	+
0x00008000	_Verb_en	Часть речи	Глагол	+

0x00010000	_Participle_en		Причастие	+
0x00018000	_Gerund_en		Деепричастие	+
0x00020000	_Adjective_en		Прилагательное	+
0x00028000	_Noun_en		Существительное	+
0x00030000	_Pronoun_en		Местоимение	+
0x00038000	_Numeral_en		Числительное	+
0x00040000	_AdVerb_en		Наречие	+
0x00048000	_Preposition_en		Предлог	+
0x00050000	_Conjunction_en		Союз	+
0x00058000	_Particle_en		Частица	+
0x00060000	_Interjection_en		Междометие	+
0x00070000	_Article_en		Артикль	–
0x00078000	_ComparativeWord_en		Сравнительное слово	+
0x00080000	_Animate_en	Одушевленность	Одушевленное	+
0x00100000	_NotAnimate_en		Неодушевленное	+
0x00200000	_ReturnPron_en		Возвратно-усилительное местоимение	–
0x00400000	_2st_Verb_form_en	Форма глагола	Прошедшее неопределенное время действительного залога	–
0x00800000	_3st_Verb_form_en		Причастие прошедшего времени	–
0x00C00000	_4st_Verb_form_en		Причастие настоящего времени	–
0x01000000	_Count_en	Тип числительного	Количественное	+
0x02000000	_Ordinal_en		Порядковое	+
0x04000000	_DefiniteArt_en	Тип артикля	Неопределенный	–
0x08000000	_IndefiniteArt_en		Определенный	–
0x10000000	_IndefiniteT_en	Группы времен	Простое	–
0x20000000	_Continuous_en		Длительное	–
0x30000000	_Perfect_en		Совершенное	–
0x40000000	_PerfectContinuous_en		Совершенное длительное	–

Применив операцию побитового «и» к значению морфологической информации словоформы и маски определенной категории, можем получить значение этой грамматической категории для словоформы. Если словоформе категория присуща – получим ненулевое значение. Например, определение значения категории числа происходит путем применения операции побитового «и» для значения морфологической информации и маски категории. Если словоформе категория не присуща, то результат этой операции равен 0. Приведем пример определения значения категории «число» для слова, морфологическая информация которого хранится в переменной MI:

MI & count_mask_en

Результат: _Singular_en, _Plural_en или 0.

Таблица 2

Маски категорий морфологической информации

Числовое	***Макроопределение***	***Маска категории***

значение		
0x00000007	case_mask_en	Падеж
0x00000018	rod_mask_en	Род
0x00000060	count_mask_en	Число
0x00000180	time_mask_en	Время
0x00000600	face_mask_en	Лицо
0x00001800	active_passive_mask_en	Залог
0x00006000	adjfrm_mask_en	Степень сравнения, краткость
0x00078000	part_of_speech_mask_en	Часть речи
0x00180000	animate_mask_en	Одушевлённость
0x00C00000	aspect_of_verb_mask_en	Вид глагола
0x03000000	number_type_mask_en	Тип числительного
0x0C000000	article_type_mask_en	Тип артикля
0x70000000	tence_group_mask_en	Группы времен

Все словарные формы, включенные в парадигму, состоят из одного слова и могут быть отличны друг от друга по написанию. Из одной словоформы состоят парадигмы следующих частей речи: наречие, союз, междометие, предлог. Примеры парадигм остальных частей речи приведём в таблице 3.

Таблица 3

Примеры парадигм изменяемых частей речи английского языка

Часть речи	***Напи-сание***	***Лемма, 1/0***	***Морфологическая информация***
Прилага-тельное	angry	1	_Adjective_en
	angrier	0	_Adjective_en | _ComparativeFormOfAdj_en
	angriest	0	_Adjective_en | _ExellentFormOfAdj_en
Сущест-вительное	project	1	_Noun_en | _Singular_en
	project's	0	_Noun_en | _Singular_en | _Prityag_en
	projects	0	_Noun_en | _Plural_en
	projects'	0	_Noun_en | _Plural_en | _Prityag_en
Местоиме-ние	you	1	_Pronoun_en | _Singular_en | _FaceSec_en | _Nominative_en
	you	0	_Pronoun_en | _Singular_en | _FaceSec_en | _Objekt_en
	your	0	_Pronoun_en | _Singular_en | _FaceSec_en | _Prityag_en
	yours	0	_Pronoun_en | _Singular_en | _FaceSec_en | _PritagAbsol_en
	yourself	0	_Pronoun_en | _Singular_en | _FaceSec_en | _ReturnPron_en
	you	1	_Pronoun_en | _Plural_en | _FaceSec_en | _Nominative_en
	you	0	_Pronoun_en | _Plural_en | _FaceSec_en | _Objekt_en
	your	0	_Pronoun_en | _Plural_en | _FaceSec_en | _Prityag_en
	yours	0	_Pronoun_en | _Plural_en | _FaceSec_en | _PritagAbsol_en
	yourselves	0	_Pronoun_en | _Plural_en | _FaceSec_en | _ReturnPron_en
	I	1	_Pronoun_en | _Singular_en | _FaceFir_en | _Nominative_en
	me	0	_Pronoun_en | _Singular_en | _FaceFir_en | _Objekt_en
	my	0	_Pronoun_en | _Singular_en | _FaceFir_en | _Prityag_en
	mine	0	_Pronoun_en | _Singular_en | _FaceFir_en | _PritagAbsol_en
	myself	0	_Pronoun_en | _Singular_en | _FaceFir_en | _ReturnPron_en
Глагол	go	1	_Verb_en
	went	0	_Verb_en | _2st_Verb_form_en
	gone	0	_Verb_en | _3st_Verb_form_en
	going	0	_Verb_en | _4st_Verb_form_en
	goes	0	_Verb_en | _Pres_en | _IndefiniteT_en | _FaceThi_en | _Singular_en

Каждая словоформа парадигмы описывается тремя значениями: написанием; пометкой, указывающей, является словоформа леммой (1) или нет (0); значением морфологической информации как совокупности значений отдельных грамматических категорий, объединенных операцией побитового «или» (в табл. 3 обозначена символом ‘|’).

Только для местоимений используются следующие значения грамматической категории «падеж»: «Именительный», «Объектный падеж», «Притяжательный абсолютный».

Для русского и английского языков результаты морфологического анализа дополняются результатами графематического анализа. На этапе графематического анализа получают интерпретацию элементы или последовательности элементов текста, которые на последующих этапах анализа текста будут рассматриваться как единое целое (обобщенный базовый элемент) определённого типа:

- нелексическая единица (телефонный номер, обозначение даты и времени, адрес электронной почты, адрес интернет, имя файла, комбинация клавиш, смайлик);

- сокращение, аббревиатура;

- слово, написанное «вразрядку»;

- группа лексических единиц (фамилия с инициалами; географические и административные названия, состоящие из одного и более слов; имена собственные, состоящие из одного и более слов; устойчивые словосочетания и обороты, неоднословные предлоги; идиоматические выражения).

Объединение лексических единиц в группы выполняется на основе анализа написаний этих единиц и написаний их лемм. Если некоторая последовательность написаний лексических единиц/написаний лемм лексических единиц принадлежит словарю географических названий, словарю имен собственных, словарю административных названий, базе устойчивых словосочетаний и оборотов или базе идиоматических

выражений, то эта последовательность объединяется в группу лексических единиц.

Интерпретация слов, написанных вразрядку, и упомянутых групп лексических единиц включает в себя одно или несколько значений морфологической информации. Это позволяет на этапах синтаксического и семантического анализа оперировать не несколькими лексическими единицами, а одним обобщенным базовым элементом. Во многих случаях количество интерпретаций обобщенного базового элемента значительно меньше количества интерпретаций входящих в него лексических единиц. За счет уменьшения, таким образом, количества анализируемых единиц предложения и количества их интерпретаций ускоряется процесс анализа и снижает неоднозначность его результатов, поскольку упомянутые единства в тексте зачастую упоминаются как одна смысловая единица. При этом подобные объединения лексических единиц не приводят к потере данных. Согласно предложенному методу используется многоуровневое представление текста. Оно, с одной стороны, сохраняет все данные, полученные на более ранних этапах обработки, а с другой – позволяет получить представление о структурных элементах текста: лексических единицах, группах лексических единиц и нелексических единицах текста. Такое представление позволяет упростить и ускорить синтаксический анализ текста. Кроме того, оно даёт возможность изменить интерпретацию отдельных структурных элементов текста или их перечень, если такая необходимость возникнет на более поздних этапах обработки текста (синтаксический, семантический, прагматический), что сделает лингвистический процессор более гибким и надёжным.

2.2. Модуль семантико-синтаксического анализа

Кратко опишем особенности метода синтаксического анализа, реализованного в данном модуле, и их влияние на эффективность его работы. При этом будем оперировать понятием «обобщенный базовый элемент», понятием *сегмент*.

Сегмент определяется следующими значениями:

– последовательность элементов, составляющих «тело» сегмента каждый из которых является или обобщенным базовым элементом или сегментом;

– левая и правая границы сегмента;

– № элемента сегмента, являющегося главным словом в сегменте (этот элемент при анализе сегмента, являющегося родительским по отношению к данному, будет «представлять» весь сегмент); перечень интерпретаций главного слова сегмента, которые представляют сегмент.

Структуры данных, описывающие сегмент и обобщенный базовый элемент, также содержат поля для хранения альтернативной интерпретации сегмента. Альтернативная интерпретация (написание и его морфологическая информация) может «представлять» весь сегмент в родительском сегменте. В качестве написания альтертернативной интерпретации используем местоименные слова: *тогда-то; потому-то; там-т;, такой-то; тот-то; так-то; тому-то; то-то; затем-то.* Это нужно, если обобщенный базовый элемент представляет собой идиоматическое выражение или сегмент соответствует неморфологизированному члену предложения (член предложения, представленный неспециализированной формой для занимаемой синтаксической позиции [13], например: … *нашел угол за недорого.* (А. Рыбаков) в значении *такой-то угол*).

При анализе предложения в рамках рассматриваемого метода выполняются следующие шаги.

1. Выделение в отдельные сегменты последовательностей слов фрагмента:

а) сложных числительных;

б) последовательностей наречий, предшествующих прилагательному;

в) групп слов, состоящих из существительного и последовательности предшествующих ему прилагательных/причастий/порядковых числительных, согласующихся с существительным;

г) групп слов, состоящих из глагола и последовательности предшествующих ему наречий.

Эти операции выполняются в указанной последовательности. При этом в выделенной последовательности слова не должны быть разделены ни знаками препинания, ни союзами или другими словами.

	Выделение слов сегмента	Слова, «представляющие» элементы сегмента
	Заботливый папа немедленно подарил мальчику двадцать вторую очень маленькую красную машинку	Заботливый папа немедленно подарил мальчику двадцать вторую очень маленькую красную машинку
а)	Заботливый папа немедленно подарил мальчику (двадцать вторую) очень маленькую красную машинку	Заботливый папа немедленно подарил мальчику вторую очень маленькую красную машинку
б)	Заботливый папа немедленно подарил мальчику (двадцать вторую) (очень маленькую) красную машинку	Заботливый папа немедленно подарил мальчику вторую маленькую красную машинку
в)	Заботливый папа (немедленно подарил) мальчику (двадцать вторую) (очень маленькую) красную машинку	Заботливый папа подарил мальчику вторую маленькую красную машинку
г)	(Заботливый папа) (немедленно подарил) мальчику ((двадцать вторую) (очень маленькую) красную машинку)	папа подарил мальчику машинку

Такой подход позволяет сократить количество элементов, которые подлежат анализу далее. Как легко заметить здесь не возникает упомянутая в работе [7] трудность нахождения связи «в группах вида ‘прилагательное + прилагательное + существительное’».

2. Фрагментация предложения – разбиение предложения на сегменты по знакам препинания и определение типа каждого сегмента.

На этом шаге типу сегмента может быть приписано одно из значений: причастный оборот; деепричастный оборот; вводное слово или конструкция; обращение (слово или сочетание слов, называющее того, к кому или к чему обращаются с речью), неопределённый тип сегмента.

Между словами сегментов устанавливают связи на основе анализа их написаний и морфологической информации. Из перечня пар потенциально связанных слов формируют дерево зависимости [14]. Отметим что

предлагаемый в работе [14] набор типов связей предполагает наличие отдельных типов связей между словами, входящими в состав главных членов предложения (главных связей), и отдельных типов для связей со второстепенными членами предложения (второстепенных связей). Построение дерева зависимости при этом начинается с подбора с главных связей, соответствующих одной из минимальных структурных схем предложения [15]. Таким образом, удаётся избежать вопросов с главными членами предложения, в которых предикат выражен не формой глагола [5,7]. При этом для уменьшения количества вариантов разбора используют словарь валентности глаголов [16]. По нему выбирают предпочтительные связи между глаголом и его актантами, запоминают семантический класс глагола и типы предикатной связи присутствующих в предложении актантов.

Если удалось построить дерево зависимости, то сегмент помечают как «связный с потенциальным предикатом». Для остальных строят дерево зависимости словосочетаний по алгоритму из работы [17], в котором связи между словами могут иметь тип: согласование, управление, примыкание. В зависимости от того, удалось ли построить дерево, сегмент помечают либо как связный, либо как «несвязный».

Сегмент «водное слово или конструкция» определяются по базе шаблонов вводных слов и конструкций. Такие сегменты, а также обращения не считают членами предложения, их не связывают с другими сегментами и в дерево зависимости не включают (его не подчиняют предикату), а подчиняют специально предусмотренному во всех предложениях корневому узлу «*Top», которому подчиняется и главное слово предложения, как предложено в работе [5].

Сегменты, которым приписаны значения «причастный оборот», «деепричастный оборот», «вводное слово или конструкция», «обращение», в предложении с обеих сторон выделяются запятыми. При формировании сегмента эти знаки препинания включаются в состав сегмента в качестве его границ. То есть они не входят ни в тело сегмента, ни в оставшуюся часть

предложения и не будут помехой при поиске однородных членов предложения и границ простых предложений внутри сложносочиненного предложения.

Если предложение не содержит знаков препинания и союзов, то всё предложение будет представлено одним сегментом, а его будем анализировать по алгоритмам, разработанным для синтаксического анализа простых распространённых неосложнённых причастными и деепричастными оборотами предложений [14].

3. Поиск границ простых предложений в составе сложноподчиненного предложения и определение типа связи между ними

На этом шаге используется та часть базы шаблонов союзов, которая описывает подчинительные союзы. Обратимся к организации базы шаблонов [18] и преимуществам, которые она даёт как в части анализа сложных предложений, так и при анализе употребления союзов в остальных случаях. Шаблон союза описывается полями:

1) количество слов первой части союза;
2) количество слов второй части союза;
3) написание союза;
4) предполагается ли запятая перед союзом 1- да, 0 -нет, 2-может стоять перед союзом, если он во второй части;
5) возможно ли многократное повторение, разделённое запятыми (открытое употребление союза) 1- да, 0 – нет;
6) возможно ли разделение союзом двух однородных членов (закрытое употребление союза) 1- да, 0 – нет;
7) тип союза;
8) номер класса правил для поиска связанных союзом слов.

Номер класса правил определяет множество правил, каждое из которых представлено кортежами:

1) грамматическое выражение, которое сопоставляется с морфологической информацией первого из слов, связанных

посредством союза;

2) грамматическое выражение, которое сопоставляется с морфологической информацией второго из слов, связанных посредством союза;
3) перечень грамматических категорий, по которым должны согласоваться морфологические информации первого и второго слов элементы;
4) порядок следования в первого и второго слов в тексте (0 – безразличен, 1 – обязательно второе слово следует после первого, 2 - обязательно первое слово следует после второго);
5) требуемое положение первого слова по отношению к границам сегмента, в который он входит, и по отношению к границе между связанными сегментами;
6) требуемое положение второго слова по отношению к границам сегмента, в который он входит, и по отношению к границе между связанными сегментами.

Первые три элемента этих правил заимствованы из работы [3], где они описывали структуру правила согласования слов.

Для односоставных союзов «первым словом» считается слово, находящееся левее союза, а вторым – находящееся правее союза. Для двусоставных союзов первое слово принадлежит сегменту, в котором находится первая часть составного союза, а второе - сегменту, в котором находится вторая часть составного союза. Причем для подчинительных союзов первое первая часть союза принадлежит главной клаузе.

Описанная выше организация шаблонов правил позволяет решать отмеченные в работе [7] вопросы современных парсеров связанные с неодносложными союзами, с разбором трёх и более однородных членов, выбора слов главной и подчиненной клауз, связывающих их между собой, а также отмеченный в работе [5] вопрос с выразительным союзом «и».

Сегменты, между словами которых найдена связь, объединяют в один путем объединения их элементов. Предпочтение отдаётся сегментам, которые ближе расположены друг к другу.

4. Установление связей между однородными членами.

Для установления однородных членов используется база шаблонов сочинительных союзов, поля которой приведены выше. После этого анализируются оставшиеся знаки препинания.

5. Получение предикатной структуры предложения.

Предикатная структура представлена фрагментом дерева зависимости, в котором дугам приписаны типы семантической связи, а для глагола отмечен его семантический класс.

Из узлов фрагмента дерева зависимости (также как из самого дерева зависимости) доступны как написания соответствующих слов в тексте, так и их интерпретации (леммы и морфологическая информация), которые соответствуют синтаксическому разбору, что важно для последующего семантического и прагматического анализа.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований была разработана система синтаксического анализа, в которой реализованы методы автоматической обработки текста на русском и английском языках, позволяющие проводить его морфологический и семантико-синтаксический анализ. Это позволило сделать следующие выводы.

1. Предложено использовать многоуровневое представление текста, которое, с одной стороны, сохраняет все данные, полученные на более ранних этапах обработки, а с другой – позволяет получить представление о минимальных структурных элементах текста: лексических (слова) и нелексических (сокращения, аббревиатуры, адреса, даты и т.п.) единицах текста. Такое представление позволяет упростить синтаксический анализ текста. Кроме того, оно даёт возможность изменить интерпретацию отдельных структурных элементов текста или их перечень, если такая

необходимость возникнет на более поздних этапах обработки текста (синтаксический, семантический, прагматический), что сделает лингвистический процессор более гибким и надёжным.

2. Введенные понятия структурных единиц предложения как обобщенный базовый элемент и сегмент, разработанные структуры данных для их описания, а также процедура выделения в отдельные сегменты последовательностей слов фрагмента предложения позволяют:

- эффективно выделять атрибутивные связи в словосочетаниях, включая проблемный случай нахождения связи в группах вида «прилагательное + прилагательное + существительное», упомянутый в работе [5];

- упростить процедуру синтаксического анализа – сократить после выделения сегментов количество анализируемых слов предложения за счет главного слова в сегменте.

3. Предложенная организация шаблонов союзов и правил для их выделения позволяет решать такие проблемные вопросы современных парсеров как интерпретация неодносложных союзов, разбор трёх и более однородных членов, выбор слов главной и подчиненной клауз, связывающих их между собой, а также проблему выразительного союза «и».

4. Использование минимальных структурных схем предложения позволяет избежать вопросов с главными членами предложения, в которых предикат выражен не формой глагола, а для русского языка – учитывать связку типа *есть*, кроме того, не допускать ошибок при обработке идиоматических конструкций «малого синтаксиса».

5. Использование словаря валентности глаголов позволяет уменьшить количество вариантов разбора, поскольку по нему выбирают предпочтительные связи между глаголом и его актантами, запоминают семантический класс глагола и типы предикатной связи присутствующих в предложении актантов.

6. Созданный синтаксический парсер представляет собой «легкий» инструментарий для анализа текста. Он предлагает универсальный подход к семантико-синтаксическому анализу, используя семантическую классификацию предикатов, адаптируемую под большинство языков. Это выгодно отличает представленную разработку от систем ABBYY Syntactic and Semantic Parser и ЭТАП-3, которые используют лингвистические ресурсы, создаваемые многими годами и большим количеством профессионалов, а следовательно, имеют огромную себестоимость и являются зависимыми от языка.

Представленная синтаксическая модель предложения является универсальной, описывает предложения русско- и англоязычных текстов и позволяет полностью выявлять как предикативные так и синтагматические отношения в виде дерева зависимостей, осуществлять первичный семантический анализ за счет учета семантического содержания актантов предиката, используя семантическую классификацию предикатов.

Работа была выполнена в рамках НИР «Исследование и разработка программного обеспечения понимания неструктурированной текстовой информации на русском и английском языках на базе создания методов компьютерного полного лингвистического анализа» (по контракту Минобрнауки от «07» июня 2012 г. 07.514.11.4133).

Список литературы:

1. Красников И. А., Никуличев Н. Н. Гибридный алгоритм классификации текстовых документов на основе анализа внутренней связности текста [Электронный ресурс] / И. А.Красников, Н. Н. Никуличев // «Инженерный вестник Дона», 2013, №3. – Режим доступа: http://ivdon.ru/magazine/archive/n3y2013/1773 (доступ свободный) – Загл. с экрана. – Яз. рус.

2. Носко В. И. Система автоматизированного построения графа социальной сети [Электронный ресурс] / В. И. Носко // «Инженерный

вестник Дона», 2012, №4 (часть 2). – Режим доступа: http://www.ivdon.ru/magazine/archive/n4p2y2012/1428 (доступ свободный) – Загл. с экрана. – Яз. рус.

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 // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11 (18): В 2 т. Т. 2: Доклады специальных секций – М.: Изд-во РГГУ, 2012. – С. 91-103.

4. Antonova A. A., Misyurev A. V. Russian dependency parser SyntAutom at the DIALOGUE-2012 parser evaluation task // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11 (18): В 2 т. Т. 2: Доклады специальных секций – М.: Изд-во РГГУ, 2012. – С. 104-118.

5. Iomdin L., Petrochenkov V., Sizov V., Tsinman L. ETAP parser: state of the art // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11 (18): В 2 т. Т. 2: Доклады специальных секций – М.: Изд-во РГГУ, 2012. – С. 119-131.

6. Каневский Е.А., Боярский К.К. Семантико-синтаксический анализатор SemSin // Электронный ресурс: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Kanevsky.pdf

7. Толдова С. Ю., Соколова Е. Г., Астафьева И., Гарейшина А., Королева А., Привознов Д., Сидорова Е., Тупикина Л., Ляшевская О. Н. Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11 (18): В 2 т. Т. 2: Доклады специальных секций – М.: Изд-во РГГУ, 2012. – С. 77-90.

8. И.М. Ножов Морфологическая и синтаксическая обработка текста (модели и программы) // Электронный ресурс: http://www.aot.ru/docs/Nozhov/msot.pdf

9. Дорохина Г.В., Павлюкова А.П. Модуль морфологического анализа слов русского языка // Искусственный интеллект. – 2004. – № 3. – С. 636-642.

10. Дорохина Г.В. Исследование алгоритма морфологического анализа слов с дефисным написанием / Г.В. Дорохина, А.О. Журавлёв, Е.А. Бондаренко // Системы и средства искусственного интеллекта. ССИИ-2012 : материалы международной научной молодёжной школы (пос. Кацивели, АР Крым, Украина, 1-5 октября 2012). - Донецк : ИПИИ «Наука і освіта», 2012. - С.17-24.

11. Дорохина Г. В. Модуль морфологического анализа без словаря слов русского языка / Г. В. Дорохина, В. Ю. Трунов, Е. В. Шилова // Искусственный интеллект. – №2. – 2010. – С.32-36.

12. Ляшевская О. Н., Астафьева И., Бонч-Осмоловская А. А., Гарейшина А., Гришина Ю., Дьячков В., Ионов М., Королева А., Кудринский М., Литягина А., Лучина Е., Сидорова Е., Толдова С. Ю., Савчук С. О., Коваль С. А. Оценка методов автоматического анализа текста: морфологические парсеры русского языка//Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2010) Т. 9. Вып. 16. М.: РГГУ, 2010, С. 318-326.

13. Лущай В.В. Заполнение позиционного состава предложения по принципу функциональной эквивалентности: интроспективный анализ в русле экспликационной грамматики / В.В. Лущай – Донецк: ДонНУ. – 2010. – 255с.

14. Дорохина Г. В. Автоматическое выделение синтаксически связанных слов простого распространенного неосложненного предложения / Г.В. Дорохина, Д. С. Гнитько // «Сучасна інформаційна Україна: інформатика, економіка, філософія»: матеріали доповідей конференції, 12 - 13 травня 2011 року, Донецьк, 2011. Т. 1. – с. 34-38.

15. Современный русский язык: Учеб. Для филол. спец. высших учебных заведений. Под ред. В.А. Белошапковой. – 3-е изд. – М.: Азбуковник, 1997. – 928 с.

16. Бондаренко Е. А. Принципы автоматической обработки естественно-языковых текстов: валентностный подход / Е. А. Бондаренко, О. А. Каплина // Искусственный интеллект. — 2013. — N1. — С. 80-90.

17. Дорохина Г.В. Ограничение количества гипотез фразы при распознавании слитной речи // Известия ТРТУ – 2005. – № 10. – С. 54-60.

18. Харламов А.А. Метод выделения главных членов предложения в виде предикативных структур, использующих минимальные структурные схемы / А.А Харламов, Т.В. Ермоленко, Г.В. Дорохина, Д.С. Гнитько // Речевые технологии. — 2012. — №2. — С.75-84.