И.В. Смирнов, А.О. Шелманов

Семантико-синтаксический анализ естественных языков
Часть I. Обзор методов синтаксического и семантического анализа текстов[1]

Аннотация. Рассмотрены задачи семантико-синтаксического анализа текстов на естественных языках. Приведен обзор подходов и методов синтаксического и семантического анализа текстов. Сделаны выводы о применимости существующих подходов к разработке методов семантико-синтаксического анализа текстов

Ключевые слова: синтаксический анализ, семантический анализ, формальная грамматика, машинное обучение.

Введение

Востребованные задачи автоматической обработки текстов на естественных языках включают в себя поиск по запросу, классификацию и кластеризацию текстов, извлечение знаний и фактов из текстов, поиск близких текстов и многие другие задачи. Качество результатов решения перечисленных задач напрямую зависит от применяемых подходов к представлению и обработке текстов. Такие подходы условно можно разделить на два класса: статистические и лингвистические. При первом подходе (наиболее часто используемом в настоящее время) текст представляется как упорядоченное множество последовательностей символов (слов), а обработка текстов сводится к статистической обработке встречаемости слов. Статистические подходы хотя и решают многие задачи обработки текстов, но являются подходами типа «грубой силы» и не позволяют принципиально решать многие задачи, такие как машинный перевод или извлечение фактов, на высоком уровне качества. Лингвистические подходы используют методы компьютерной лингвистики, которые представляют текст как набор более сложных структур (морфологических, синтаксических, семантических) и позволяют решать задачи обработки текстов на значительно более высоком уровне.

Последнее время всё больше появляется исследований, посвященных автоматическому семантическому анализу текстов. Многие разработчики информационно-аналитических и поисковых систем заявляют о применении методов семантического анализа в своих решениях, однако большинство современных подходов к семантике текста состоят в учете статистических характеристик слов и их сочетаемости, учете семантических классов слов. При этом из лингвистической информации используется только морфологические признаки и словарные формы слов, в некоторых случаях выполняется синтаксический анализ. Эффективность использования таких подходов видится недостаточной, поэтому весьма актуальным становится разработка методов автоматического семантического анализа текстов, которые основываются на развитой лингвистической теории семантики и позволяют весьма эффективно решать многие задачи обработки текстов с помощью компьютерных программ.

[1] Работа выполнена при поддержке РФФИ (проект № 12-07-33068) и Минобрнауки России по государственному контракту № 07.514.11.4134 от 08.06.2012г.

Существующие лингвистические подходы к представлению семантики текстов на естественных языках часто не обеспечивают достаточно эффективного решения задач автоматической обработки текстов. Связано это, прежде всего, с незаконченностью лингвистических исследований в области семантики языка, а также с многозначностью конструкций самого естественного языка. Современные исследования естественного языка подвели к идее о неразрывной связи синтаксиса и семантики. Развиваемый в данной работе подход к описанию семантики языка основан на понятии значения в контексте смысла высказывания. Реляционно-ситуационный анализ текста [1], опираясь на положения коммуникативной грамматики русского языка [2], оперирует значениями синтаксем – минимальных синтаксических единиц. Значение синтаксемы передаёт элементарный смысл высказывания. Центральной идеей теории коммуникативной грамматики является принцип тесной взаимосвязи синтаксиса и семантики, однако в существующих реализациях анализаторов это не учитывается. Фазы синтаксического и семантического анализа принципиально разделены, что, на наш взгляд, порождает не только логические противоречия при проектировании архитектуры анализатора, но и сложности реализации, т.к. для двух указанных фаз анализа разрабатываются различные программные структуры и алгоритмы, зачастую никак не согласованные между собой. Всё это приводит к неробастности анализатора и невысокой скорости анализа.

Целью настоящей работы является исследование взаимодействия синтаксиса и семантики и решение задачи их интеграции в рамках единого семантико-синтаксического анализатора. Первая часть работы посвящена исследованию существующих подходов к синтаксическому и семантическому анализу текстов на естественном языке.

1. Синтаксический анализ

Синтаксический анализ занимает одно из важнейших мест в цепочке обработки текстов на естественном языке. Среди центральных задач компьютерной лингвистики синтаксический анализ является «узким местом», поскольку до сих пор не предложено достаточно робастного и эффективного, как с точки зрения сложности вычислений, так и с точки зрения качества обработки, подхода для его проведения. Поэтому на сегодняшний день синтаксический анализ является актуальной и активно исследуемой проблемой компьютерной лингвистики.

1.1. Синтаксическая структура текста

Задачей синтаксического анализа является эксплицитное описание синтаксической структуры текста. Большинство моделей представления синтаксической структуры опираются либо на грамматику зависимостей [3], либо на грамматику непосредственно-составляющих [4].

Грамматика зависимостей предполагает, что предложения текста представляют собой деревья зависимостей, в которых слова связаны ориентированными дугами, обозначающими синтаксическое подчинение [5-7]. Считается, что этот формализм хорошо отражает специфику языков с произвольным порядком слов, в которых между словами может присутствовать значительное количество непроективных[2] связей. К таким языкам, относится немецкий, чешский, русский, а также другие восточнославянские языки.

В грамматике непосредственно-составляющих предложения текста представляются в виде иерархии составляющих (синтаксических групп): все предложение разбивается на непересекающиеся проективные группы, которые в свою очередь состоят из более мелких групп и т.д. вплоть до атомарных групп – слов предложения. В грамматике составляющих, по сути, не допустимы непроективные синтаксические отношения между словами. Поэтому этот формализм считается подходящим для языков с фиксированным порядком слов, например, для английского, где проективность соблюдается более строго. В английском языке встречаются и непроективные связи: «John saw a dog yesterday which was a Yorkshire Terrier». В этом предложении пересекаются связи между «saw – yesterday» и «dog – was». Тем не менее, несмотря на ряд исключений для большинства конструкций в английском свойство проективности соблюдается.

Надо заметить, что эти два формализма применяются для синтаксического анализа различных языков примерно в равной степени.

[2] Свойство проективности синтаксического дерева означает, что если представить дерево графически, то связи между собой не пересекаются. Непроективные связи нарушают свойство проективности дерева.

В случае, когда все связи в дереве зависимости являются проективными, оно является изоморфным соответствующему дереву составляющих [8, 9]. Как отмечается в работе [10] только 10% предложений в корпусе СинТагРус (корпус синтаксически размеченных предложений, входящий в состав НКРЯ [11]) содержат непроективные связи, а их общая доля по отношению к общему числу связей не превышает 1%. Таким образом, в каких-то задачах можно, не сильно пожертвовав полнотой, учитывать только проективные связи и с одинаковой эффективностью применять оба формализма.

Известны работы, в которых для синтаксического анализа русского языка применяется грамматика связей [12, 13]. Эту модель нельзя отнести ни к грамматике зависимостей, ни к грамматике составляющих. Она отличается тем, что текст в ней представляется не в виде дерева, а в виде сети, в которой отсутствует корневой элемент [14]. Несмотря на то, что существует ряд проектов использующих LinkGrammar [15], широкого распространения этот формализм не получил.

1.2. Методы синтаксического анализа

Синтаксический анализ разделяют на глубокий (полный) анализ (deep parsing) и поверхностный (shallow parsing) [16, 17]. Задачей глубокого синтаксического анализа является построение полного синтаксического дерева предложения с максимальной связанностью с учетом дальних связей, а также определение грамматических функций слов предложения (подлежащее, сказуемое, обстоятельства места, времени и т.д.). Для поверхностного синтаксического анализа нет четкого определения, это понятие объединяет в себе различные подходы, работающие на уровне синтаксиса, которые направлены на построение неполной (частично связанной) синтаксической структуры текста разной сложности. Поверхностный синтаксический анализ охватывает такие задачи как разделение предложения на рекурсивно невложенные синтаксические группы (chunking), сегментацию (выделение в предложении различных оборотов и простых предложений в составе сложного), а также построение поверхностного синтаксического дерева. Заметим, что поверхностные анализаторы могут строить довольно связанные деревья, которые близки к результатам работы глубоких синтаксических анализаторов [18]. Однако поверхностные анализаторы обычно не предназначены для установления всех синтаксических связей в предложении, не учитывают дальние связи и не предназначены для определения грамматических функций слов предложения. Подобное упрощение задачи синтаксического анализа по сравнению с глубоким анализом позволяет использовать вычислительно и алгоритмически более простые и робастные методы. Кроме того в рамках упрощенной задачи (например, для выделения синтаксических групп) удается достичь высоких показателей качества.

Методы как полного, так и поверхностного синтаксического анализа находят применение в широком спектре различных прикладных задач обработки текстов. Глубокий синтаксический анализ традиционно используется в системах машинного перевода [9, 19], поверхностный анализ в качестве альтернативы глубокому применяется в информационно-поисковых и информационно-аналитических системах. Кроме того в некоторых лингвистических анализаторах поверхностный анализ является этапом предобработки текста перед проведением глубокого анализа, что позволяет в целом упростить и ускорить процедуру проведения синтаксического анализа [20-22].

1.2.1. Методы поверхностного синтаксического анализа

Первоочередная задача, решаемая в рамках поверхностного синтаксического анализа – это задача выделения непересекающихся рекурсивно невложенных синтаксических групп (например, именных или глагольных), зачастую без установления связей внутри этих групп (chunking). С решением подобных задач справляются простые конечные автоматы [23].

Однако для более робастного анализа применяются подходы, основанные на машинном обучении. Наиболее известный подход заключается в том, чтобы представить задачу выделения синтаксических групп в виде задачи классификации слов предложения. Классы определяют принадлежность слова к некоторой группе. Существует несколько постановок задачи такого рода. В одной из известных постановок слова предложения разделяются на три класса: B (before) – «начало группы»; I (inside)

– «внутри группы»; O (outside) – «конец группы» (данный подход упоминается часто под названием BIO [24-25]). В другой постановке слова классифицируются по четырем классам: «начало группы», «конец группы», «и начало и конец», «вне группы» [26]. Анализатор обучается на примерах, взятых из корпуса, в котором соответствующие группы уже размечены. В качестве признаков классификации обычно используется информация о части речи слова и позиция в предложении. После классификации каждого слова ищется лучшее разбиение предложения на группы. При этом учитывается, что группы пересекаться не могут, а также некоторые другие ограничения.

Еще одна задача, которая рассматривается в рамках поверхностного синтаксического анализа – это сегментация предложения. Под сегментацией будем понимать, согласно с [27], разбиение сложных предложений на простые, а также выделение любых обособляемых оборотов: причастных и деепричастных оборотов, согласованных определений, вводных оборотов и т.д. Для русского языка подход к сегментации предложения описывается в работах [28, 29]. Сегментация использует результаты морфологического анализа и разбиение предложения на синтаксические группы. Она состоит из двух этапов. Сначала проводится фрагментация предложений по знакам пунктуации и сочинительным союзам. Затем, фрагменты склеиваются или вкладываются внутрь других фрагментов на основании заранее созданной системы правил. Эти правила опираются в основном на предположение о том, что каждый сегмент не должен содержать больше одного неоднородного подлежащего, сказуемого, причастия или деепричастия. При этом границы сегментов не должны разбивать синтаксические группы. В результате склейки фрагментов формируется набор сегментов, которые можно отождествлять с простыми предложениями или оборотами.

Многие синтаксические анализаторы (парсеры), которые принято относить к поверхностным, строят весьма связные синтаксические деревья, которые по качеству установления синтаксических связей могут соперничать даже с результатами глубоких анализаторов [18]. Методы, которые используются в подобных анализаторах, можно условно разделить на методы, основанные на ручном составлении грамматик, и статистические методы с применением машинного обучения.

В первом случае вручную строится формальная грамматика (например, контекстно-свободная грамматика), которая в дальнейшем используется анализатором для построения поверхностных синтаксических деревьев. Одна из ключевых особенностей естественного языка – неоднозначность не позволяет напрямую использовать парсеры, созданные для анализа контекстно-свободных искусственных языков (например, языков программирования). Для естественного языка невозможно построить однозначную детерминированную грамматику (т.е. грамматику, которая порождает строчки только одним способом), которая описывала бы его в достаточной мере. Синтаксический анализ детерминированным анализатором по неоднозначной грамматике затруднителен, в то время как применение классических недетерминированных анализаторов приводит к значительному проигрышу по скорости из-за необходимости возвратов (backtracking).

Исходя из вышесказанного, следует, что для синтаксического анализа естественных языков необходимо применять специальные приемы, чтобы учесть их неоднозначность. Для уменьшения количества альтернатив при синтаксическом анализе применяют различные эвристики. Они отсеивают бесперспективные варианты, а также определяют стратегию поведения парсера. Кроме того, для решения этой проблемы применяют парсеры и формализмы, созданные специально для обработки неоднозначных грамматик. Наиболее известный подход, который называется chart parsing, предложен Martin Kay [30]. Chart парсер разбирает неоднозначные контекстно-свободные грамматики. Главная идея по уменьшению вычислительной сложности заключается в том, чтобы при наличии неоднозначности не проводить каждый раз анализ всего предложения для каждой альтернативы заново, а сохранять результаты для уже обработанных альтернатив и учитывать их повторно для анализа новых. Этот подход можно рассматривать и как кэширование результатов вызовов функций вместе с параметрами вызова. Функция для одинаковых наборов параметров возвращает одинаковые результаты, следовательно, можно вернуть кэшированный результат и избежать ее повторного вычисления [9]. Подход chart parsing реализован в алгоритмах

Earley [31] (проводит разбор сверху вниз) и CYK [32] (проводит разбор снизу вверх). Оба этих алгоритма имеют в общем случае сложность $O(n^3)$, но могут иметь лучшую сложность в частных случаях.

Большой стимул развитию поверхностных парсеров дали методы машинного обучения с учителем [18]. Для разработки подобных анализаторов необходимы большие синтаксически размеченные корпуса текстов (банки синтаксических деревьев), например, такие как Penn TreeBank [33]. Они с одной стороны являются материалом для обучения статистических анализаторов, а с другой позволяют разным исследователям объективно оценивать и сравнивать друг с другом эффективность разработанных методов на одних и тех же данных. Подходы с применением машинного обучения хорошо развиты зарубежными исследователями для английского языка. Размеры и число доступных исследователям корпусов для русского языка весьма ограничено, поэтому работ, посвященных статистическим анализаторам русского языка, немного. Среди них надо отметить работы [34-36].

Наиболее известный подход статистического синтаксического анализа текста заключается в построении стохастической контекстно-свободной грамматики [37, 38]. На основе банка синтаксических деревьев автоматически строится грамматика, в которой для каждой группы неоднозначных правил установлена вероятность. В ходе синтаксического анализа текста, с помощью такой грамматики решается задача максимизации вероятности разбора (последовательности применения правил) для заданного предложения. Модели, предложенные разными исследователями, различаются в основном способом вычисления вероятности разбора и решения задачи максимизации.

Передовое направление исследований в области синтаксического анализа естественного языка состоит в использовании обучения без учителя. Это направление развивают, например, исследователи из Стендфордского университета [39, 40]. Главная идея их методов заключается в том, чтобы использовать зависимости, которые явно можно отыскать в тексте или разметке документа (пунктуация, форматирование), для простых случаев (например, для двусловий) и затем обобщать эти случаи на все более и более сложные выражения. В результате подобных обобщений формируется грамматика, которая затем может быть использована парсером для разбора текста. На сегодняшний день методы, использующие машинное обучение без учителя, сильно уступают по качеству разбора методам, использующим размеченные корпуса. Однако подход, основанный на обучении без учителя, в перспективе позволит уменьшить зависимость технологий по обработки естественного языка от дорогостоящих и трудоемких в создании синтаксически размеченных корпусов.

1.2.2. Глубокий синтаксический анализ

Методы глубокого синтаксического анализа в той или иной степени опираются на подходы поверхностного анализа. Однако при глубоком анализе применяются усовершенствования, направленные, во-первых, на построение наиболее полной синтаксической структуры текста с учетом дальних связей, а во-вторых, на определение грамматических функций слов в предложении.

Большинство анализаторов, выполняющих глубокий синтаксический анализ, основаны на вручную (или автоматизировано) составленной системе правил. Такие анализаторы используются в коммерческих системах ABBY Compreno [9] и Xerox XLE [41], в некоммерческих проектах RASP [42] и ENJU [43]. Наиболее известным в России исследовательским проектом такого рода является ЭТАП-3 [19].

Глубокие синтаксические анализаторы имеют обычно весьма сложную систему правил. В системах ABBY Compreno и ЭТАП-3 она представляет собой одну из разновидностей грамматик зависимостей. Системы XLE, RASP и ENJU используют грамматики составляющих. Грамматики в глубоких анализаторах, как правило, состоят из нескольких классов правил, выполняющих различные функции: установление связей между словами, фильтрация построенных связей, ранжирование синтаксических деревьев. Также подобные анализаторы используют большое количество эвристик, в соответствии с которыми ведется применение правил и построение синтаксической структуры предложения. Заметим, что в современных анализаторах, основанных на правилах, широко применяются и статистические методы. Однако они играют не главную роль, а используются скорее как дополнительно средство для управления разбором и ранжирования построенных деревьев.

Кроме непосредственно синтаксических правил и эвристик в системах, проводящих глубокий синтаксический анализ, широко применяются семантические знания. Это различные тезаурусы, словари, описывающие предикатно-аргументные структуры, а также словари семантической и синтаксической сочетаемости. Разработчики глубоких анализаторов, участвовавших в соревновании парсеров на семинаре, проведенном в рамках Диалог-2012 [44], отмечают, что построение полноценной синтаксической структуры текста не возможно без его семантической интерпретации и применения экстралингвистических знаний [9, 19, 45]. Таким образом, последнее время усиливается интеграция синтаксического и семантического анализа, и в таком случае говорят о семантико-синтаксическом анализе. Например, разработчики компании ABBY [9] заявляют, что в их синтаксическом анализаторе синтаксис и семантика рассматриваются как две грани одной структуры и строятся при анализе текста параллельно. Семантико-синтаксический анализ позволяет разрешать синтаксическую неоднозначность и отсеивать ложные конструкции за счет обеих интерпретаций.

Несмотря на очевидные минусы анализаторов, основанных на правилах, такие как трудоемкость разработки полной системы синтаксических правил, сложность портирования этой системы на другие языки, низкая эффективность при обработке «грязных» текстов, общая трудоемкость сопровождения и отладки, этот подход позволяет зачастую добиться наилучших показателей точности и полноты при синтаксическом анализе грамматически правильных текстов.

Современные исследования в области глубокого синтаксического анализа направлены на разработку методов, целиком основанных на машинном обучении. В работах [46, 47] предлагается подход, который позволяет проводить синтаксический анализ близкий к глубокому. В этом подходе применяется парсер, работающий по методу «перенос-свертка» (shift-reduce parser), действия которого предсказываются с помощью классификатора, предварительно обученного на размеченном корпусе синтаксических деревьев. В работе [46] использовался memory-based классификатор, предсказания которого основываются на поиске в памяти сохраненных примеров, которые в соответствии с некоторой моделью близки к входным данным. В более поздней работе [47] авторы для обучения используют алгоритм SVM [48]. Парсер просматривает слова предложения слева направо и для каждого слова и его контекста получает предсказание следующего своего действия от классификатора: поместить / извлечь следующее слово в/из внутреннего стека; установить правую/левую связь между словом в тексте / словом в стеке. Парсер строит связные и проективные синтаксические деревья. Для установления непроективных связей используется специальная процедура. Анализатор также может быть обучен для назначения связям типов, отражающих грамматическую функцию слов в предложении.

Открытая реализация этого анализатора «MaltParser» сравнивалась с синтаксическим анализатором системы ЭТАП-3 на банке синтаксических деревьев SynTagRus [10]. Примечательно, что анализаторы показали весьма сходные показатели правильности установления хозяина (89% у «MaltParser» и 88% у ЭТАП-3), из чего можно сделать вывод о том, что подходы, основанные на обучении, пригодны и для глубокого анализа.

2. Семантический анализ

Существует несколько теорий семантики естественного языка, описание некоторых можно найти, например, в [49]. Как уже говорилось выше, в настоящей работе мы развиваем подход, основанный на значениях синтаксических единиц (синтаксем), несущих элементарный смысл в высказывании, и связей между ними. Семантический анализ в данном случае состоит в выделении именных синтаксем, установлении их значений и связей между ними [1]. Именная синтаксема выражена в предложении именной или предложной группой, она характеризуется предлогом, падежом и категориально-семантическим классом главного управляющего слова. Значение синтаксемы в случае наличия в предложении предикатного слова (глагола или деепричастия) определяется на основании словаря, где перечислены сочетания предикатных слов и возможных синтаксем со значениями. Если предикатное слово в предложении отсутствует, то для установления значений синтаксем используются специальные контекстные правила [50]. Таким образом, главную роль

в установлении значений здесь играют синтаксис и словарь. Рассмотрим зарубежные работы в данной области.

Идеи о значении высказываний в языке и способах его выражения были впервые систематизировано описаны в работах Ч. Филлмора [51, 52]. Ключевое место в подходе Филлмора занимает понятие падежа, который выражает роль – семантическое содержание аргумента при предикате. Согласно Филлмору, «смыслы падежей образуют набор универсальных, возможно врожденных, понятий, идентифицирующих некоторые типы суждений, которые человек способен делать о событиях, происходящих вокруг него, – суждений о вещах такого рода, как «кто сделал нечто», «с кем нечто случились», «что подверглось некоему изменению».

Эти идеи положили начало целому направлению зарубежных исследований в области «понимания текста» (text understanding), посвященному установлению семантических ролей (semantic role labeling) слов и словосочетаний.

2.1. Методы установления семантических ролей

Основополагающей для целого ряда работ в области semantic role labeling является работа [53], в которой представлена система определения семантических ролей, описываемых посредством семантического фрейма. Она приписывает слову в предложении или абстрактные семантические роли «Агенс» (Agent) или «Пациенс» (Patient), или более специфичные для некоторой предметной области роли, например, «Говорящий» (Speaker), «Сообщение» (Message). Для схематического представления ситуации, включающей участников, их свойства и взаимосвязи, используется фрейм. Роль является частью (слотом) такого фрейма.

Авторы предлагают методы для вероятностного оценивания появления того или иного значения фрейма. Эти подходы статистические, они основаны на обучении по размеченному корпусу и проверке точности предсказания значений на тестовой части размеченных данных. Тексты размеченного корпуса FrameNet [54] подвергались синтаксическому анализу, затем извлекались элементы фреймов и дополнительные свойства предложений. На основе значений слотов рассчитывались вероятности каждой семантической роли.

В своей работе авторы делают вывод, что синтаксис и семантика связаны настолько, что существует возможность обучиться распознаванию семантических отношений, основываясь только на синтаксической информации. Поэтому большинство признаков для обучения отражают синтаксические характеристики компонентов семантических ролей, включая синтаксические отношения между предикатом и компонентами ролей. Среди этих признаков: тип синтаксической группы; категория управления; путь в синтаксическом дереве; позиция; залог; главное слово. Применялись также методы обобщения (generalization) для повышения точности обнаружения семантических ролей для новых (unseen) данных, т.е. для предикатов, не встречающихся в обучающих примерах. Были введены три обобщенных абстрактных роли Agent, Patient, Goal и составлена иерархия, относящая каждую специфичную роль к одной из этих трех. Затем в обучающих и тестовых примерах роли заменялись соответствующими обобщенными ролями. Система достигала 80% точности классификации на тестовом множестве.

Результатами описанного статистического подхода являются условные вероятности и их веса в линейной комбинации. Эти вероятности трудно интерпретировать словесными формулировками в грамматических категориях, что является, на наш взгляд, недостатком данного подхода. Еще одним недостатком является наличие признаков-лексем, что завязывает метод на тексты определенной предметной области. Применяемые методы обобщения, основанные на создании иерархий свойств и ролей, не всегда применимы.

Ряд работ [55-57] посвящен извлечению семантической информации из текстов (information retrieval, IR) на основе модели генеративного лексикона (generative lexicon, GL), в которой текст представляется набором слов и связей между ними, в частности, рассматриваются связи между именами существительными и глаголами. Задача состоит в определении тех существительных и глаголов в предложении, которые находятся в семантической связи.

Стоит отметить работу [55], в которой используется индуктивное логическое программирование (Inductive Logic Prigramming). В данной работе с помощью метода машинного обучения строятся правила, записанные на языке предикатов первого порядка. Для представ-

ления связей между словами текста используется модель генеративного лексикона, согласно которой каждому слову сопоставляется аргументная структура и так называемая qualia структура. Последняя состоит из qualia ролей (roles), рассматриваемых как функции, сопоставляющие слову следующие "атрибуты объекта": CONSTITUTIVE – состоит ли он из частей или сам является частью; FORMAL – свойства, которые выделяют слово из других слов; TELIC – его назначение и функции, цель; AGENTIVE – факторы, «ответственные» за его происхождение, источник происхождения.

Для обучения используется специальный корпус французского языка MATRA-CCR Aerospitiale, с морфологической (POS Tagging) и семантической разметкой. Отличительной особенностью данной работы является использование символических индуктивных методов машинного обучения в отличие от статистических методов, применяемых к GL модели текста другими исследователями. В качестве метода обучения используется метод ALEPH, реализованный на Прологе. На вход подаются положительные примеры – пары «существительное-глагол», образующие связь, и отрицательные примеры – пары, не образующие связь. Метод пытается найти правила, которые объясняют (покрывают) все или почти все положительные примеры и не покрывают большинство или все отрицательные примеры. В обучающих примерах (и, соответственно, в гипотезах) используется информация о взаимном расположении элементов пар «существительное-глагол» в предложениях, а также расстояние между ними. Точность предсказания полученных гипотез на тестовом множестве составляла 0.813, полнота составляла 0.890. Выбор индуктивного метода обучения обосновывается авторами тем, что он позволяет получать правила, объясняющие обучающие примеры и позволяющие интерпретировать результаты предсказаний семантических связей с точки зрения лингвистики. Особенно отмечается понятность правил для лингвистов.

В ряде работ [58, 59] решается задача обучения семантических анализаторов (semantic parsers) текста на корпусе предложений, каждое из которых сопровождается смысловым представлением на специальном формальном языке. В зависимости от решаемых конкретных задач под языком смыслового представления (meaning representation language) подразумевается язык запросов к базе данных по географии или язык команд для эмулирующих игру в футбол программ. В первом случае выражения языка представляют собой запросы на языке Пролог, во втором случае используется специальный язык CLang, описывающий тактику и поведение игроков с помощью правил "если-то". Задача состоит в том, чтобы с помощью семантического анализатора перевести выражения на естественном языке (английском) в один из указанных языков смыслового представления. Перед этим выполняется обучение анализаторов с помощью статистической обработки естественного языка. В работе используется подход, объединяющий синтаксис и семантику для представления обучающих примеров и результатов обучения. Фактически, обучающий пример представляет собой синтаксическое дерево с семантическими пометами для некоторых узлов.

Авторы применяют несколько методов для обучения анализаторов. Один из них индуктивный, основанный на технологии индуктивного логического программирования, другие 4 основаны на статистических подходах. Отметим, что для первой задачи, индуктивный метод достигал максимальной точности 0.77 при полноте 0.79, в то время как точность статистических методов при тех же обучающих примерах достигала максимума 0.95 при полноте 0.70. Для второй задачи, индуктивный метод показал максимальную точность 0.49 при полноте 0.11, максимальная точность статистических методов составляла 0.86 при полноте 0.54 для того же множества обучающих примеров.

В работе [60] для установления семантических ролей предлагается использовать фреймы и методы обучения, основанные на энтропии. Под установлением семантических ролей подразумевается приписывание семантических ролей аргументам глаголов. Предложенный подход состоит из двух шагов. На первом шаге для глагола в предложении определяются типы его аргументов. Аргументы могут быть трех типов: обязательный, необязательный и т.н. нуль-аргумент, который не является семантически значимым для глагола. На втором шаге каждому обязательному или необязательному аргументу приписывается семантическая роль на основании множества фреймов. Каждый фрейм описывает комбинацию глагола с несколькими

аргументами разных типов, при этом каждому аргументу приписана роль. Сопоставляя фрейм-шаблон с конкретной комбинацией глагола и его аргументов, можно приписать этим аргументам соответствующие шаблону роли. Поскольку для одного аргумента при глаголе может подойти несколько фреймов-шаблонов, необходимо разрешение многозначности. Для определения типов аргументов, а также для снятия многозначности при установлении семантических ролей используется подход, опирающийся на понятие энтропии. На этапе обучения вычисляются вероятности появления аргументов с определенными значениями признаков в определенных слотах фреймов. При классификации вычисляется вероятность появления некоторой роли в позиции рассматриваемого аргумента с учетом других аргументов и ролей. Эксперименты показали, что средняя точность предложенного авторами подхода составляла 0.72 при полноте 0.64.

Несколько работ основывается на применении принципа максимальной энтропии [61, 62]. Цель обучения состоит в подборе распределения условных вероятностей, дающих максимальное значение условной энтропии [63]. На этапе классификации выбирается метка, для которой значение энтропии максимально. Эксперименты показали, что точность предсказания на тестовом множестве составила в среднем 0.805 при полноте 0.728.

В работе [64] предлагается сочетание фреймовой модели и аппарата функциональных грамматик (systemic-functional grammar) для представления смысловой модели текста. В то время как фреймовая модель позволяет описать структурную составляющую модели текста – участников ситуации, их роли, функциональная грамматика организует языковые выражения в классы, независимые от предметной области и языка. Авторы работы ставят цель реализовать семантический анализатор, который приписывает роли (значения) словам любого текста, не зависимо от предметной области. Авторы выделяют несколько типов процессов, описываемых функциональной грамматикой. Каждый тип процесса порождает множество ролевых шаблонов, каждый из которых состоит из самого процесса и участников данного процесса (события). Задача заключается в отображении текста предложения в один из таких ролевых шаблонов, но при этом необходимо учитывать только грамматические характеристики слов самого предложения. С помощью методов машинного обучения авторы выделяют множество разделяющих характеристик, позволяющих приписать синтаксическим единицам ту или иную роль с учетом характеристик окружающих их слов в предложении. Точность анализа составила от 0.70 до 0.98.

В работе [65] для установления семантических ролей используются Марковские цепи энтропии. Кроме стандартных признаков (часть речи, позиция относительно глагола) учитывались суффиксы глаголов, залог глагола, а также попарное сочетание некоторых признаков. На тестовом множестве алгоритм в среднем по разным ролям показал точность предсказания 0.71 при полноте 0.50.

В работе [66] говорится, что семантические роли в предложении взаимосвязаны и не существуют сами по себе, в связи с этим предлагается рассматривать роли в предложении как связанную структуру и устанавливать роли не для каждого слова по отдельности, а для всех слов в предложении в совокупности. Авторы используют те же признаки, что и в работе [53], описанные ранее, и добавляют свои признаки, отражающие синтаксические зависимости и связи элементов предложения. Для отдельных ролей точность классификации на тестовом множестве достигала 0.96 при полноте 0.95.

Все описанные выше работы опираются на размеченные корпуса. На сегодняшний день семантически размеченные репрезентативные корпуса существуют лишь для небольшого числа языков (русский язык, к сожалению, не входит в их число). Для многих языков подобные корпуса отсутствуют вовсе, что является препятствием для развития методов семантического анализа, основанных на машинном обучении с учителем. Эта проблема стимулировала исследователей искать подходы, которые ослабили бы зависимость анализаторов от семантически размеченных корпусов. Так активно стали использоваться методы с частичным привлечением учителя (semi-supervised).

Ряд semi-supervised подходов известен под названием «проекция аннотаций» (annotation projection). Их главная идея заключается в том, что семантическая структура размеченных предложений на одном языке сопоставляется схожим по некоторым признакам неразмеченным предложениям, переведенным на другой язык. По-

пытки перевести английские семантически размеченные корпуса предпринимались для немецкого [67], шведского [68], китайского [69].

Известны подходы по автоматическому расширению имеющихся семантических корпусов и предикатных словарей без привлечения параллельных корпусов и тезаурусов на других языках. Для определения семантического фрейма предикатов, отсутствующих в словаре, зачастую применяют тезаурус WordNet [70]. В работе [71] проводится обобщение фреймов за счет наличия гипонимии между известными и неизвестными предикатами в тезаурусе. Ряд исследователей [72, 73] трактуют задачу определения фрейма для неизвестного предиката как задачу классификации, в которой для расчета близости предикатов также используется WordNet. В работе [74] на основе небольшого семантически размеченного корпуса строится большой размеченный корпус, который используется для обучения семантических анализаторов. Новые аннотации строятся путем переноса уже известных аннотаций предложений из размеченного корпуса на близкие к ним предложения из неразмеченного корпуса. Близость предложений вычисляется путем сопоставления их синтаксических деревьев.

Самые последние работы в области семантического анализа сконцентрированы на проблеме обучения семантических анализаторов на неразмеченных корпусах. В ряде работ из этой области [75-76] предлагается подход, в котором задача установления семантических ролей представляется в виде задачи кластеризации. В этом подходе целью ставится разделить все аргументы семантической структуры (слова, или синтаксические поддеревья) на кластеры таким образом, чтобы каждый кластер соответствовал некоторой семантической роли. При этом нет заранее заданного набора ролей или их признаков. Набор семантических ролей и характерные им свойства формируются автоматически в процессе кластеризации. Полученные кластеры и их признаки, с одной стороны, могут быть использованы напрямую для решения прикладных задач, а с другой стороны, могут быть применены для быстрой семантической разметки корпуса, поскольку вручную экспертам достаточно разметить не весь текст, а лишь сопоставить роли самим кластерам.

Заключение

Современные методы синтаксического анализа показывают весьма высокую точность и полноту установления синтаксических связей: в среднем от 75% до 90% в зависимости от языка [18, 35, 36, 77].

На сегодняшний день для синтаксического анализа основополагающую роль играют статистические методы машинного обучения. Они успешно применяются как при поверхностном, так и при глубоком анализе, вытесняя или дополняя подходы, основанные на ручном построении правил и грамматик. Однако методы машинного обучения с учителем очень сильно привязаны к размеченным корпусам. Это ограничивает их применение только для тех языков, для которых существуют доступные репрезентативные размеченные корпуса. Кроме того результаты работы анализаторов, обученных на корпусе из одной предметной области, могут заметно ухудшаться на текстах из другой предметной области. Создание новых вручную размеченных корпусов очень дорого и по затратам сопоставимо с разработкой грамматики для естественного языка. Передовые исследования направлены на уменьшение зависимости от размеченных корпусов, на развитие методов машинного обучения без учителя.

Стоит также отметить возрастающую роль семантики при глубоком синтаксическом анализе. Для разрешения синтаксической омонимии необходима семантическая интерпретация текста. В связи с этим идут разработки анализаторов, в которых два уровня анализа интегрированы. Это новое направление можно назвать семантико-синтаксическим анализом.

В зарубежных исследованиях по семантическому анализу текстов, а именно установлению семантических значений, преобладает подход, основанный на автоматическом обучении грамматик и различного рода классификаторов. В качестве информативных признаков используются синтаксические и морфологические характеристики элементов анализируемого предложения, окружающих синтаксическую единицу, семантическое значение которой устанавливается. Это в первую очередь синтаксическая информация, включая: взаимное расположение элементов предложения друг относительно друга; подчинение; пути в син-

таксическом дереве от одной лексической единицы до другой; типы синтаксических групп.

Для обучения используют следующие типы методов анализа данных: индуктивные (логические); статистические. Эксперименты показывают, что статистические методы дают результаты, обладающие большей предсказательной силой по сравнению с результатами логических методов, поэтому исследователи чаще отдают предпочтение именно статистическим методам.

Все рассмотренные работы по установлению смысловых значений слов текста имеют дело с английским языком, который обладает отличным от русского языка строем. Русский язык флективный со свободным порядком слов, поэтому, многие признаки, используемые в рассмотренных работах, например, позиция относительно предиката или залог глагола являются не столь информативными для русского языка. Недостатком многих рассмотренных подходов является также опора на лексику, что ограничивает их применение предметной областью, к которой принадлежат обучающие корпусы.

Наконец, следует отметить, что для русского языка корпуса с семантической разметкой отсутствуют, что значительно затрудняет использование методов, основанных на автоматическом обучении.

Обзор методов синтаксического и семантического анализа текстов на естественном языке, выполненный в первой части работы, показывает необходимость и принципиальную возможность создания методов семантико-синтаксического анализа, объединяющих эти две фазы в одну.

При разработке методов семантико-синтаксического анализа текстов на втором этапе работы мы будем руководствоваться следующими соображениями:

1. Грамматика зависимостей является наилучшим, на наш взгляд, подходом к описанию как синтаксических, так и семантических феноменов в русском языке. В случае установления значений синтаксем, зависимость устанавливается между предикатным словом и синтаксемой, при этом последняя является зависимой, а тип зависимости совпадает со значением синтаксемы. В случае установления семантических связей, один из участников связи становится зависимым. Таким образом, процедуры синтаксического и семантического анализа могут выполняться в едином формализме.

2. Синтактико-семантический анализ в контексте реляционно-ситуационного анализа должен быть поверхностным, не глубоким, т.е. синтаксический анализ ограничивается только выделением именных и предложных групп, поверхностной сегментацией (выделением границ оборотов и простых предложений в составе сложных), семантический анализ ограничивается установлением значений синтаксем и связей между ними. Это обеспечит высокую скорость анализа и возможность обработки больших массивов текстовой информации.

3. Необходимо использовать закономерности языка, которые сильно упрощают поиск синтаксических зависимостей. Так, "...несмотря на свободный порядок слов в русском, некоторые синтаксические зависимости имеют обязательным критерием выделения жесткий линейный порядок: генитивное определение должно следовать за определяемым словом ('ножка стола', 'сын отца'); предлог предшествует существительному ('на столе', 'у отца'); в 90% случаев определение, выраженное прилагательным или местоименным прилагательным, стоит до существительного..." [29]. В анализаторе АОТ [78], например, эта закономерность приводит к регулярности грамматики именных групп [79].

4. Целесообразно по возможности использовать размеченные корпуса, как для обучения, так и для проверки разрабатываемого синтактико-семантического анализатора.

Во второй части работы предполагается разработка метода семантико-синтаксического анализа текстов, основанного на правилах, которые были бы отделены от алгоритмов обработки текста, и экспериментальная проверка метода. Пополнение множества правил предполагается как вручную, так и с помощью обучения по размеченным корпусам.

Литература

1. Осипов Г. С., Смирнов И. В., Тихомиров И. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. — 2008. — № 2. — С. 3–10.
2. Золотова Г. А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка // Институт русского языка РАН им. В. В. Виноградова. — 2004.
3. Tesnière L. Elements de syntaxe structurale. — Editions Klincksieck, 1959.
4. Chomsky N. Syntactic structures. — Mouton, The Hague, 1957. — P. 117.
5. Hudson R. Word grammar. — Blackwell Oxford, 1984.

6. Melcuk I. Dependency syntax: theory and practice. — State University of New York Press, 1988.
7. Апресян Ю. Д. Интегральное Описание Языка и Системная Лексикография. — Школа «Языки русской культуры», 1995.
8. Debusmann R. An introduction to dependency grammar // Hausarbeit fur das Hauptseminar Dependenzgrammatik SoSe. — 2000. — Vol. 99.
9. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies / K. V. Anisimovich, K. Ju. Druzhkin, F. R. Minlos et al. // Papers from the Annual International Conference "Dialogue" (2012). — Vol. 2. — 2012. — P. 91–103.
10. Синтаксический анализатор системы Этап и его оценка с помощью глубоко размеченного корпуса русских текстов / И. М. Богуславский, Л. Л. Иомдин, Д. Р. Валеев, В. Г. Сизов // Труды международной конференции «Корпусная лингвистика – 2008». — 2008.
11. Синтаксически размеченный корпус русского языка: инструкция пользователя. — 2013. — фев. — URL: http://www.ruscorpora.ru/instruction-syntax.html.
12. Протасов С. Преимущества грамматики связей для русского языка // Труды международной конференции «Диалог 2005». — 2005.
13. Протасов С. Обучение с нуля грамматики связей русского языка // X Национальная конференция по искусственному интеллекту с международным участием «КИИ-06». — 2006. — С. 515–524.
14. Sleator D. D., Temperley D. Parsing english with a link grammar. — 1991.
15. Link grammar. — 2013. — feb. — URL: http://www.abisource.com/projects/link-grammar/.
16. Abney S. P. Parsing by chunks // Principle-Based Parsing. — Kluwer Academic Publishers, 1991. — P. 257–278.
17. Federici S., Montemagni S., Pirrelli V. Shallow parsing and text chunking: a view on underspecification in syntax // Cognitive science research paper-university of Sussex CSRP. — 1996. — P. 35–44.
18. Wide-coverage deep statistical parsing using automatic dependency structure annotation / A. Cahill, M. Burke, R. O'Donovan et al. // Computational Linguistics. — 2008. — Vol. 34, no. 1. — P. 81–124.
19. Синтаксический анализатор системы этап: современное состояние. / Л.Л. Иомдин, В. В. Петроченков, В. Г. Сизов, Л. Л. Цинман // Papers from the Annual International Conference "Dialogue" (2012). — 2012.
20. Semantic services in freeling 2.1: Wordnet and ukb / Lluis Padro, Samuel Reese, Eneko Agirre, Aitor Soroa // Principles, Construction, and Application of Multilingual Wordnets / Ed. by Pushpak Bhattacharyya, Christiane Fellbaum, Piek Vossen ; Global Wordnet Conference 2010. — Mumbai, India : Narosa Publishing House, 2010. — February. — P. 99–105.
21. An integrated architecture for shallow and deep processing / Berthold Crysmann, Anette Frank, Bernd Kiefer et al. // Proceedings of ACL-2002, Association for Computational Linguistics 40th Anniversary Meeting, July 7-12. — 2002. — P. 441–448.
22. Integrated shallow and deep parsing: Topp meets hpsg / Anette Frank, Markus Becker, Berthold Crysmann et al. // In Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL 2003. — 2003. — P. 104–111.
23. Антонова А., Мисюрев А. Анализатор русского языка syntautom для соревнования синтаксических парсеров (Диалог-2012) // Papers from the Annual International Conference "Dialogue" (2012). — 2012.
24. Argamon-engelson S., Dagan I., Krymolowski Y. A memory-based approach to learning shallow natural language patterns // Journal of Experimental and Theoretical AI. — 1999. — Vol. 11. — P. 369–390.
25. Punyakanok V., Roth D. The use of classifiers in sequential inference // NIPS. — MIT Press, 2001. — P. 995–1001.
26. A learning approach to shallow parsing / M. Mufioz, V. Punyakanok, D. Roth, D. Zimak // Proc. of EMNLpo VLC'99. — 1999.
27. Экспериментальная реализация сегментационного анализа русского предложения / А.М. Баталина, М.Е. Епифанов, Т.Ю. Кобзарева и др. // Труды международной конференции «Диалог 2007». — 2007.
28. Кобзарева Т. Принципы сегментационного анализа русского предложения. // Московский лингвистический журнал. — 2004. — Т. 8. — С. 31–80.
29. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы) : Дисс кандидата наук / И.М. Ножов. — 2003.
30. Kay M. Readings in natural language processing / Ed. by Barbara J. Grosz, Karen Sparck-Jones, Bonnie Lynn Webber. — San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1986. — P. 35–70.
31. Earley J. An efficient context-free parsing algorithm // Commun. ACM. — 1970. — feb. — Vol. 13, no. 2. — P. 94–102.
32. Jurafsky D., Martin J. Speech And Language Processing: An Introduction to Natural Language Processing , Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence. — Pearson Prentice Hall, 2009.
33. Marcus M. P., Marcinkiewicz M. A., Santorini B. Building a large annotated corpus of english the penn treebank // Comput. Linguist. — 1993. — Vol. 19, no. 2. — P. 313–330.
34. Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing russian without any linguistic knowledge // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). — No. 10. — 2011. — P. 17.
35. Казенников А. О. Сравнительный анализ статистических алгоритмов синтаксического анализа на основе деревьев зависимостей // Труды международной конференции «Диалог 2010». — 2010.
36. Nivre J., Boguslavsky I. M., Iomdin L. L. Parsing the SynTagRus treebank of russian // Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). — Manchester, UK : Coling 2008 Organizing Committee, 2008. — August. — P. 641–648.
37. Collins M. Head-driven statistical models for natural language parsing // Comput. Linguist. — 2003. — Vol. 29, no. 4. — P. 589–637.
38. Charniak E. A maximum-entropy-inspired parser // Proceedings of the 1st North American chapter of the Asso-

ciation for Computational Linguistics conference / Morgan Kaufmann Publishers Inc. — 2000. — P. 132–139.
39. Spitkovsky V. I., Alshawi H., Jurafsky D. Baby Steps: How “Less is More” in unsupervised dependency parsing // NIPS 2009 Workshop on Grammar Induction, Representation of Language and Language Learning (GRLL 2009). — Whistler, Canada, 2009. — December.
40. Spitkovsky V. I., Alshawi H., Jurafsky D. Bootstrapping dependency grammar inducers from incomplete sentence fragments via austere models // Proceedings of the 11th International Conference on Grammatical Inference. — 2012.
41. Speed and accuracy in shallow and deep stochastic parsing / Ronald M. Kaplan, Stefan Riezler, Tracy H. King et al. // In proceedings of HLT-NAACL’04. — 2004.
42. Briscoe T., Carroll J. Robust accurate statistical annotation of general text. — 2002.
43. Miyao Y., Tsujii J. Feature forest models for probabilistic hpsg parsing // Comput. Linguist. — 2008. — Vol. 34, no. 1. — P. 35–80.
44. Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка / С. Ю. Толдова, Е. Г. Соколова, И. Астафьева и др. // Papers from the Annual International Conference “Dialogue” (2012). — 2012.
45. Каневский Е. А., Боярский К. К. Семантико-синтаксический анализатор SemSin // Papers from the Annual International Conference "Dialogue" (2012). — 2012.
46. Nivre J., Nilsson J. Memory-based dependency parsing // In Proceedings of CoNLL. — 2004.
47. Maltparser: A language-independent system for data-driven dependency parsing / Joakim Nivre, Johan Hall, Jens Nilsson et al. // Natural Language Engineering. — 2007. — Vol. 13, no. 2. — P. 95–135.
48. Vapnik V. N. The nature of statistical learning theory. — New York, USA : Springer-Verlag New York, Inc., 1995.
49. Апресян Ю. Д. Лексическая семантика. — М., 1974.
50. Смирнов И. В. Метод автоматического установления значений минимальных синтаксических единиц текста // Информационные технологии и вычислительные системы. — 2008. — № 3. — С. 30–45.
51. Charles J. F. The case for case. In Universals in Linguistic Theory. — 1968.
52. Филлмор Ч. Дело о падеже. // Новое в зарубежной лингвистике. — 1981. — № 10. — С. 400–444.
53. Gildea D., Jurafsky D. Automatic labeling of semantic roles // Comput. Linguist. — 2002. — Vol. 28, no. 3. — P. 245–288.
54. Welcome to FrameNet. — 2013. — feb. — URL: https://framenet.icsi.berkeley.edu/fndrupal/.
55. Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming / Vincent Claveau, Pascale Sébillot, Cécile Fabre, Pierrette Bouillon // J. Mach. Learn. Res. — 2003. — Vol. 4. — P. 493–525.
56. Claveau V., Sebillot P. From efficiency to portability: acquisition of semantic relations by semi-supervised machine learning // Proceedings of COLING’04, 20th International Conference on Computational Linguistics. — Geneva, Switzerland, 2004. — P. 61–267.
57. Ichiro Yamada T. B. Automatic discovery of telic and agentive roles from corpus data // Proceeding of the 18th Pacific Asia Conference on Language, Information and Computation. — Tokyo, Japan, 2004. — P. 115–126.
58. Mooney R. J. Learning for semantic parsing // Computational Linguistics and Intelligent Text Processing: Proceedings of the 8th International Conference (CICLing 2007) / Ed. by A. Gelbukh. — Mexico City, Mexico : Springer: Berlin, Germany, 2007. — February. — P. 311–324. — Invited paper.
59. Kate R. J., Mooney R. J. Semi-supervised learning for semantic parsing using support vector machines // Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Short Papers (NAACL/HLT-2007). — Rochester, NY, 2007. — April. — P. 81–84.
60. Bharati A., Venkatapathy S., Reddy P. Inferring semantic roles using sub-categorization frames and maximum entropy model // Proceedings of the Ninth Conference on Computational Natural Language Learning. — CONLL ’05. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2005. — P. 165–168.
61. Jiang Z. P., Ng H. T. Semantic role labeling of nombank: A maximum entropy approach // Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. — Sydney, Australia : Association for Computational Linguistics, 2006. — July. — P. 138–145.
62. Semantic role labeling system using maximum entropy classifier / Ting Liu, Wanxiang Che, Sheng Li et al. // Proceedings of the Ninth Conference on Computational Natural Language Learning. — CONLL ’05. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2005. — P. 189–192.
63. Berger A. L., Pietra V. J. D., Pietra S. A. D. A maximum entropy approach to natural language processing // Comput. Linguist. — 1996. — Vol. 22, no. 1. — P. 39–71.
64. De Busser R., Moens M.-F. Learning Generic Semantic Roles. Unpublished report. Leuven: Interdisciplinary Centre for Law & IT. — 2003. — K.U.Leuven.
65. Blunsom P. Maximum entropy markov models for semantic role labelling // Proceedings of the Australasian Language Technology Workshop 2004 / Macquarie University. — Sydney, 2004. — December.
66. Toutanova K., Haghighi A., Manning C. Joint learning improves semantic role labeling // Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics / Association for Computational Linguistics. — 2005. — P. 589–596.
67. Padó S., Lapata M. Cross-lingual annotation projection for semantic roles // Journal of Artificial Intelligence Research. — 2009. — Vol. 36. — P. 307–340.
68. Johansson R., Nugues P. A framenet-based semantic role labeler for swedish // Proceedings of the COLING/ACL on Main conference poster sessions. — COLING-ACL ’06. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2006. — P. 436–443.
69. Fung P., Chen B. Biframenet: bilingual frame semantics resource construction by cross-lingual induction // Proceedings of the 20th international conference on Computational Linguistics. — COLING ’04. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2004.
70. Miller G. A. Wordnet: A lexical database for english // Communications of the ACM. — 1995. — Vol. 38, no. 1. — P. 39–41.

71. Burchardt A., Erk K., Frank A. A WordNet Detour to FrameNet. — 2005.
72. Johansson R., Nugues P. Using WordNet to Extend FrameNet Coverage. — 2007.
73. Automatic induction of framenet lexical units / Marco Pennacchiotti, Diego De Cao, Roberto Basili et al. // In proceedings OF EMNLP-08. — 2008.
74. Fürstenau H., Lapata M. Semi-supervised semantic role labeling via structural alignment // Comput. Linguist. — 2012. — Vol. 38, no. 1. — P. 135–171.
75. Lang J., Lapata M. Unsupervised semantic role induction via split-merge clustering // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — Vol. 1 of HLT '11. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2011. — P. 1117–1126.
76. Titov I., Klementiev A. Crosslingual induction of semantic roles // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. — Jeju Island, South Korea : Association for Computational Linguistics, 2012. — July.
77. McDonald R., Lerman K., Pereira F. Multilingual dependency analysis with a two-stage discriminative parser // Proceedings of the Tenth Conference on Computational Natural Language Learning. — Association for Computational Linguistics, 2006. — P. 216–220.
78. Автоматическая обработка текста. — 2013. — фев. — URL: http://www.aot.ru/.
79. Ножов И. Морфологическая и синтаксическая обработка текста (модели и программы). — 2003. — С. 81–84.

Смирнов Иван Валентинович. Старший научный сотрудник Института системного анализа РАН. Окончил Российский университет дружбы народов в 2003 году. Кандидат физико-математических наук. Автор 29 печатных работ. Область научных интересов: искусственный интеллект, обработка естественного языка, машинное обучение, интеллектуальные поисковые машины. E-mail: ivs@isa.ru

Шелманов Артем Олегович. Инженер-исследователь лаборатории Института системного анализа РАН, аспирант. Окончил Национальный исследовательский ядерный университет «МИФИ» в 2011 году. Автор 3 печатных работ. Область научных интересов: искусственный интеллект, компьютерная лингвистика, информационно-аналитические системы, методы оптимизации. E-mail: shelmanov@isa.ru.