И. В. Смирнов, А. О. Шелманов, Е. С. Кузнецова, И. В. Храмоин, Семантико-синтаксический анализ естественных языков. Часть II. Метод семантико-синтаксического анализа текстов, *Искусственный интеллект и принятие решений*, 2014, выпуск 1, 11–24

И.В. Смирнов, А.О. Шелманов, Е.С. Кузнецова, И.В. Храмоин

Семантико-синтаксический анализ естественных языков
Часть II. Метод семантико-синтаксического анализа текстов[1]

Аннотация. Выполнен обзор методов семантико-синтаксического анализа текстов. Описан синтаксический анализатор текстов, созданный на основе MaltParser. Приведены экспериментальные исследования по выявлению признаков, влияющих на качество синтаксического анализа русскоязычных текстов. Описан семантический анализатор на основе метода реляционно-ситуационного анализа. Приведены результаты экспериментальных исследований семантического анализатора на русскоязычных текстах. Описан предложенный метод семантико-синтаксического анализа, в котором синтаксический и семантический анализ выполняются совместно. Приведены результаты его экспериментальных исследований на русскоязычных текстах, а также сравнение предложенного метода с подходом, в котором синтаксический и семантический анализ выполняются раздельно и последовательно.

Ключевые слова: семантико-синтаксический анализ, машинное обучение, синтаксический анализ, семантический анализ, semantic role labeling.

Введение

Настоящая статья является продолжением работы [1], в которой выполнен обзор методов синтаксического и семантического анализа текстов на естественном языке. Целью работы является исследование взаимодействия синтаксиса и семантики и решение задачи интеграции семантического и синтаксического анализа в рамках единого семантико-синтаксического анализатора.

Последнее время все больше проводится исследований, посвященных автоматическому семантическому анализу текстов. Многие разработчики информационно-аналитических и поисковых систем заявляют о применении методов семантического анализа в своих решениях. Однако, в ряде случаев происходит подмена понятий и за семантический анализ выдаются информационные измерения текста. Один из подходов такого рода состоит в учете статистических характеристик слов и их сочетаемости, учете семантических классов слов. При этом из лингвистической информации используется только морфологические признаки и леммы слов, в некоторых случаях выполняется синтаксический анализ. Эффективность использования этих подходов видится недостаточной для решения, таких задач как, например, извлечение информации или вопросно-ответный поиск. Поэтому весьма актуальным является 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. В нашей работе развивается подход к описанию семантики языка, основанный на понятиях значения и семантической роли [2]. Таким образом, решаемая в нашей работе задача семантического анализа близка к задаче semantic role labeling (SRL) [3].

Исследования естественного языка подвели к идее о неразрывной связи синтаксиса и семантики. На взаимосвязь синтаксиса и семантики указывают как лингвистические теории языка, например, коммуникативная грамматика [4], так и результаты экспериментальных исследований [3], показывающие, например, возможность выполнения семантического анализа

[1] Работа выполнена при поддержке РФФИ, проект №12-07-33068 «мол_а_вед»

главным образом на основе синтаксической информации. Исследователи стали использовать термин *синтактико-семантический* или *семантико-синтаксический анализ* (joint syntactic and semantic analysis), понимая под этим разные подходы и методы, интегрирующие два вида анализа с целью повышения качества каждого из них. С одной стороны, этапы синтаксического и семантического анализа могут выполняться различными методами, но при этом результаты синтаксического и семантического анализа оказывают взаимное влияние друг на друга. С другой стороны, если считать, что синтаксис и семантика тесно взаимосвязаны, то выполнять эти два этапа можно одним методом.

Результатом работы является основанный на машинном обучении семантико-синтаксический анализатор, выполняющий синтаксический и семантический анализ текстов на основе единого подхода, в одной процедуре с использованием единой структуры данных. Обучение семантико-синтаксического анализатора проводилось на корпусе синтаксических деревьев, в который автоматически была добавлена семантическая разметка с помощью анализатора, использующего набор заданных эвристик и семантический ресурс – словарь предикатных слов [5]. Чтобы определить признаки, важные для синтаксического анализа, мы отдельно провели обучение и оценку синтаксического анализатора. Полученный синтаксический анализатор мы использовали для построения системы, в которой синтаксический и семантический анализ выполняются отдельно и последовательно. Для оценки качества семантического анализа мы создали тестовый подкорпус синтаксических деревьев, в который вручную была добавлена семантическая разметка. На этом подкорпусе был отдельно оценен семантический анализатор на основе словаря предикатных слов, система, выполняющая синтаксический и семантический анализ последовательно, а также разработанный семантико-синтаксический анализатор.

В первом разделе приведен обзор методов семантико-синтаксического анализа текстов, во втором и третьем разделах описаны эксперименты отдельно с синтаксическим и семантическим анализом, в четвертом разделе описан метод семантико-синтаксического анализа, приведены результаты его экспериментальных исследований на русскоязычных текстах, а также сравнение предложенного метода с подходом, в котором синтаксический и семантический анализ выполняются раздельно и последовательно. В заключении приводятся результаты работы и предлагаются направления дальнейших исследований.

1. Обзор методов семантико-синтаксического анализа

Рассмотрим зарубежные и российские исследования в области семантико-синтаксического анализа.

1.1. Зарубежные исследования в области семантико-синтаксического анализа

Зарубежные исследователи уделяют большое внимание методам компьютерного семантико-синтаксического анализа текстов на естественном языке. Наибольший интерес для нас представляют подходы, в которых совместно решаются задачи синтаксического разбора и установления семантических ролей (SRL).

Работа [3] – первая работа, в которой был предложен подобный подход. В ней с помощью синтаксического анализатора, основанного на стохастической контекстно-свободной грамматике, для каждого предложения строилось большое количество вариантов синтаксического разбора. Для каждого из вариантов выбиралось наилучшее распределение ролей, а затем выбирался вариант с наибольшей общей вероятностью. Эксперименты показали небольшой прирост полноты установления ролей.

Большой вклад в развитие методов семантико-синтаксического анализа внесли семинары CoNLL Shared Task 2008 [6] и CoNLL Shared Task 2009 [7], в которых оценивалось качество и синтаксического (dependency parsing), и семантического анализа (SRL) совместно. В 2008 году задача ставилась только для одного языка – английского, в 2009 году участники должны были выполнить анализ для еще шести языков: каталанского, китайского, чешского, немецкого, японского и испанского. Шесть команд, участвовавших в семинаре в 2008 году, и четыре команды, участвовавших в семинаре 2009 году, реализовали методы, которые так или иначе совмещают синтаксический и семантический анализ.

В работах [8, 9] описывается система, которая участвовала в семинаре в 2008 году, а в ра-

боте [10] представлена ее модификация, которая участвовала в семинаре в 2009 году. Семантико-синтаксический анализ в этих системах выполнялся путем построения дерева зависимости одновременно с установлением семантических меток синтаксическим связям. Сначала проводился предварительный синтаксический анализ, который предоставлял признаки для последующих этапов анализа, затем с помощью SVM классификатора определялись предикатные слова, после чего проводился основной совмещенный анализ. Предварительный и основной анализ выполнялись с помощью алгоритма синтаксического разбора Eisner'a [11], который относится к классу алгоритмов поиска максимального (минимального) остовного дерева (maximum (minimum) spanning tree, сокращенно MST). Для определения весов синтаксических связей использовалась нейронная сеть. Вес синтаксической связи с семантической пометкой в совмещенном анализе складывался из веса синтаксической связи и нормированного веса семантических меток (ролей) от разных предикатов. Строя синтаксическое дерево одновременно с назначением семантических меток, анализатор максимизирует совокупный вес семантико-синтаксической структуры. Экспериментальное тестирование этого метода показало, что совмещенный анализ не влияет на качество синтаксического анализа, но увеличивает качество установления семантических ролей по сравнению с системой, в которой синтаксический и семантический анализ выполняются последовательно.

Средний результат на семинаре в 2008 году и один из лучших результатов на семинаре в 2009 году показала система, описанная в работах [12-14]. Авторы применили анализатор на основе системы переходов с использованием метода перенос-свертка для параллельного построения отдельной синтаксической и семантической структуры предложения. Для предсказания действия анализатора была обучена модель на основе Incremental Sigmoid Belief Networks (ISBN) [15], которая представляет собой динамическую байесовскую сеть, последовательно изменяющую свою структуру на основе частично построенного вывода. В качестве признаков задействовались уже построенные в ходе вывода синтаксические и семантические зависимости. Таким образом, модель при обучении максимизирует совместную вероятность синтаксических и семантических зависимостей. Эксперименты, проведенные в работах, показали, что при отсутствии взаимодействия между синтаксической структурой и семантической структурой в процессе вывода существенно ухудшается качество семантического анализа. Авторы также показали, что совмещение семантического анализа и синтаксического анализа не оказывает существенного влияния на качество последнего.

В [16] описывается система, которая принимала участие в семинаре в 2008 году. В ней было применено смешивание результатов восьми синтаксических анализаторов, построенных на основе MaltParser с помощью варьирования алгоритмов и направлений разбора, в единое комбинированное синтаксическое дерево. Подобный шаг позволил значительно улучшить качество синтаксического анализа по сравнению с результатами лучшего единичного анализатора. Были построены две различные системы семантического анализа. Кроме того, на основе MaltParser была создана система семантико-синтаксического анализа, которая одновременно строит синтаксическое дерево зависимости и назначает связям семантические метки. Предикаты в предложении были предсказаны вспомогательной системой. Синтаксическое дерево, построенное системой семантико-синтаксического анализа, смешивалось с комбинированным синтаксическим деревом. Новое комбинированное дерево направлялось на вход двум системам семантического анализа. Окончательная семантико-синтаксическая структура строилась путем смешивания результатов двух систем семантического анализа с результатом системы семантико-синтаксического анализа.

В [17] описывается система, принимавшая участие в семинаре в 2009 году. Основная идея подхода, реализованного в этой системе, заключается в том, чтобы сначала последовательно провести синтаксический и семантический анализ, а затем их повторить, задействовав на новой итерации признаки, полученные из уже построенной семантико-синтаксической структуры. В работе проводилось до двух итераций цепочек анализа. Эксперименты показали, что этот подход может довольно значительно улучшить первоначальный результат. Однако улучшения наблюдались не для всех языков и не всегда лучший результат достигался на последней итерации. Авторы заключают, что эф-

фект от повторного анализа зависит от качества первоначального анализа.

В [18] описывается система, которая участвовала в семинаре в 2008 году и достигла лучших совместных оценок качества синтаксического и семантического анализа. В системе этапы анализа выполнялись последовательно. В результате строилось несколько вариантов синтаксических деревьев для каждого из которых строились предикатно-аргументные структуры. Затем выполнялась выбирающая окончательный вариант процедура переранжирования результатов, обученная максимизировать совместную оценку синтаксического и семантического анализа. Эксперименты показали, что процедура переранжирования улучшает результаты и синтаксического, и семантического анализа, и, соответственно, совместные оценки. Схожий подход использовался также в работе [19].

В [20] описывается система, принимавшая участие в семинаре в 2008 году. Авторы реализовали подход, в котором интегрированы этапы назначения меток синтаксическим связям и поиска семантических аргументов с использованием марковской модели максимальной энтропии.

В [21] описывается система, принимавшая участие в семинаре в 2009 году. В ней реализовано полностью совместное обучение установлению синтаксических и семантических связей. Для этого использовались memory-based классификаторы: один предсказывал наличие связи между токенами, другие два – наличие связи и ее метку (синтаксическую метку или семантическую роль). Результаты классификаторов комбинировались с помощью процедуры ранжирования.

Как видно из представленного обзора проблема семантико-синтаксического анализа хорошо изучена. Многие из исследователей сообщают о том, что системы, в которых реализовано взаимодействие между синтаксическим и семантическим анализом, превосходят по оценкам качества аналоги, в которых эти этапы выполняются последовательно. Лучший результат на семинаре CoNLL Shared Task 2008 показала система, реализующая процедуру переранжирования результатов на основе синтаксической и семантической информации [18]. Однако системы, выполняющие синтаксический и семантический анализ раздельно, в 2008 году были в тройке лучших, а в 2009 году заняли первые два места. На данный момент нельзя однозначно сказать имеют ли существенное превосходство методы совместного анализа над подходами, в которых синтаксический и семантический анализ выполняются последовательно.

1.2. Российские исследования в области семантико-синтаксического анализа

В российских исследованиях семантический анализ в большинстве случаев сводится к поиску семантических классов слов с использованием тезаурусов. В отдельных исследованиях решается задача установления семантических значений (ролей) (например [22]), однако нами не было найдено работ, в которых бы решалась задача количественной оценки качества установления семантических ролей на размеченном корпусе для русского языка.

Термин семантико-синтаксический или синтактико-семантический анализ текстов в российских работах используется довольно часто, но имеет различные интерпретации. Под семантико-синтаксическим или синтактико-семантическим анализом (далее будем отождествлять эти два варианта) часто подразумевается использование семантической информации при синтаксическом анализе, а именно использование словарей, тезаурусов и других ресурсов [23]. В ряде работ синтактико-семантический анализ означает поверхностный синтаксический анализ с последующим выделением целевой информации – адресов, номеров машин, организаций и т.п. Например, в работе [24] синтактико-семантический анализ в таком понимании выполняется с помощью правил. В лингвистическом процессоре Semantix правила выделяют из текста группы слов, описывающих какой-либо объект, и заменяют их на одно (абстрактное) слово, с которым связывается соответствующий фрагмент семантической сети и которому присваиваются определенные признаки, в том числе, указывающие тип объекта. Правила применяются в определенной последовательности, по мере применения таких правил строится семантическая сеть – содержательный портрет документа. Аналогичный подход применяется в технологии RCO [25], где синтактико-семантический анализ – это анализ семантической сети с помощью синтактико-семантических шаблонов, выполняемый с целью поиска целевых данных (ситуаций, их участников и т.п.). Таким образом, в указанных

подходах синтактико-семантический анализ состоит в извлечении информации.

Можно сделать вывод, что задача объединения синтаксического и семантического анализа в одной процедуре в российских исследованиях еще не решалась.

Основная проблема российских исследований, на наш взгляд, состоит в том, что явно задача установления семантических ролей для русского языка не поставлена, как следствие, отсутствуют корпуса с семантической разметкой и не проводятся семинары, аналогичные CoNLL Shared Task, посвященные данной задаче. Появление и развитие проекта Framebank [26] внушает надежды на решение указанных проблем.

2. Синтаксический анализ

Для экспериментов с семантическим и семантико-синтаксическим анализом в нашей работе была необходима система, выполняющая быстрый и качественный синтаксический анализ русскоязычных текстов. В этом разделе описывается построение и исследование системы синтаксического анализа на основе машинного обучения.

Эксперименты, в которых для синтаксического анализа текста на русском языке применялись методы машинного обучения, ранее уже проводились в работах [27, 28]. В них исследовался MaltParser [29, 30], – синтаксический анализатор, строящий синтаксическое дерево зависимостей, основанный на системе переходов, который свое каждое последующее действие (переход) определяет исходя не из заданной грамматики, а получает из предсказания предварительно обученного классификатора. Анализатор обучался на корпусе синтаксических деревьев зависимости СинТагРус [31, 32] (подкорпус НКРЯ). Проведенные в работах [27, 28] эксперименты показали, что этот подход позволяет добиться весьма высокого качества синтаксического анализа текстов на русском языке.

Для проведения синтаксического анализа мы также задействовали MaltParser (версия 1.7.2). Несмотря на то, что в работах [27, 28] проводится детальное исследование различных признаков, влияющих на качество синтаксического анализа, нас интересовал ряд случаев, которые в них не были освещены. Поэтому помимо создания синтаксического анализатора, в нашей работе была поставлена задача исследования влияния различных признаков и размера обучающего корпуса на качество синтаксического анализа текстов на русском языке. Нас также интересовал вопрос о возможности применения подобного подхода для решения практических задач и для проведения экспериментов более высокого уровня. Поэтому в работе внимание также уделялось производительности и ресурсоемкости процедуры синтаксического анализа.

В нашей работе в качестве обучающего корпуса также использовался синтаксический размеченный корпус русского языка СинТагРус, разработанный в ИППИ РАН. В СинТагРус тексты разбиты на предложения, а предложения на токены. Для токенов указаны форма слова, лемма, часть речи и другие морфологические характеристики. Синтаксические связи между токенами предложения имеют метки (теги), определяющие тип связи, и образуют связное синтаксическое дерево зависимости. Пунктуация в корпусе является «висящей», т.е. не соединяется связями с другими токенами. Версия корпуса, которая использовалась нами содержит 53 439 предложений и 774 373 токенов без учета пунктуации. СинТагРус был преобразован в формат CoNLL [33], а морфологические характеристики токенов были преобразованы в формат, соответствующий стандарту MULTEXT-East Morphosyntactic Specifications, Version 4 (MTE) [34, 35]. Между морфологическими характеристиками в СинТагРус и характеристиками, описанными в вышеуказанном стандарте, нельзя установить взаимно-однозначное соответствие. Часть характеристик, которые присутствуют в MTE, отсутствуют в СинТагРус и наоборот. Поэтому некоторые из морфологических тегов в преобразованном корпусе неполны и могут отличаться от тегов, которые используются в работах [27, 28].

Корпус был автоматически размечен категориально-семантическими классами существительных (КСК) (подробнее в разделе 3). Значения категориально-семантических классов необходимы для проведения семантического анализа, кроме того, в работе исследовалось их влияние на качество синтаксического анализа. Помимо КСК в корпус как отдельные признаки были добавлены морфологические характеристики, влияющие на согласование: число, падеж и род (а также их конкатенация в единую строку).

Предварительные эксперименты показали, что производительность и требования к ресурсу памяти, многократно различаются в случаях, когда анализатор обучен на корпусе с полным набором меток (типов) синтаксических связей и на корпусе без меток, в пользу последнего. Обучение на корпусе с полным набором меток позволяет анализатору не только строить синтаксическое дерево, но и помечать связи этими метками, определяя таким образом их тип, что может оказаться полезным при решении некоторых задач. Однако часто необходимо построить только синтаксическое дерево. В этих случаях большое преимущество в скорости анализатора, обученного на корпусе без меток, делает его применение выгодным. Поэтому в основном исследовалось два варианта корпуса: с полным набором меток синтаксических связей и вариант, в котором связи имеют только две пометки: «ROOT», если у токена нет родителя (он связан со вспомогательной вершиной – корнем дерева); и «DEP», если токен синтаксически зависит от другого токена.

В качестве обучающего корпуса использовался подкорпус СинТагРус, содержащий 33 000 предложений и 480 537 токенов без учета пунктуации, подкорпус для тестирования содержит 5 000 предложений и 71 413 токенов без учета пунктуации.

Для обучения модели, предсказывающей действия анализатора, MaltParser предоставляет два классификатора LIBSVM [36] (реализация метода опорных векторов, позволяющая использовать различные ядра) и LIBLINEAR [37] (включает в себя оптимизированную реализацию метода опорных векторов с линейным ядром). Как показано в работе [28] качество синтаксического анализа немного выше при использовании LIBSVM. Однако при этом обучение модели, а также сам процесс синтаксического анализа выполняются гораздо медленнее чем при использовании LIBLINEAR. Поскольку обучающий корпус весьма объемный разница между качеством анализа при использовании этих двух классификаторов небольшая. Учитывая весьма большие различия по скорости обучения и работы классификаторов, анализатор, использующий LIBLINEAR, лучше подходит для решения прикладных задач. Поэтому нами во всех экспериментах использовался LIBLINEAR.

В качестве базовой конфигурации анализатора и базового набора признаков использовалась конфигурация и набор признаков из работы [28], которая открыто доступна на электронном ресурсе [38]. В ней используется алгоритм разбора «nivreeager», который работает по принципу метода «перенос-свертка». В базовую конфигурацию была добавлена опция «covered_root = left», которая указывает анализатору провести предварительную обработку пунктуации в корпусе, задействовав механизм псевдо проективного анализа [39]. Поскольку пунктуация в СинТагРус не соединяется связями с другими токенами, в представлении MaltParser она привязывается к вспомогательной вершине – корню синтаксического дерева. Это приводит к появлению большого количества непроективных связей, соединяющих корень дерева и пунктуацию. Большое количество непроективных связей сильно ухудшает качество синтаксического анализа. Без добавления указанной опции оценки качества анализа до 7% ниже, чем результаты, представленные в [28]. Это может объясняться тем, что в работе [28] пунктуация дополнительным образом преобразовывалась. В дальнейшем базовой конфигурацией считается конфигурация из [28], в которую добавлена указанная опция.

Базовый набор признаков содержит информацию о части речи слов (POSTAG), форме слов в тексте (FORM), их лемме (LEMM) и морфологических характеристиках (FEATS), а также о метках уже установленных синтаксических связей (DEPREL). Заметим, что в базовой конфигурации весь набор морфологических характеристик рассматривается как атомарная строка.

Были проведены эксперименты со следующими наборами признаков:

- Base = Базовый = FORM + LEMMA + POSTAG + DEPREL + FEATS.
- NoForm = Базовый – FORM = LEMMA + POSTAG + DEPREL + FEATS.
- NoForm_NoLemma = Базовый – FORM – LEMMA = POSTAG + DEPREL + FEATS.
- NoForm_NoLemma_CSC = Базовый – FORM – LEMMA + КСК = POSTAG + DEPREL + FEATS + КСК.
- Base_CSC = Базовый + КСК = FORM + LEMMA + POSTAG + DEPREL + FEATS +КСК.

- NoFeats = Базовый – FEATS = FORM + LEMMA + POSTAG + DEPREL.
- NoFeats_SplFeats = Базовый – FEATS + SPLFEATS (число, падеж, род) = FORM + LEMMA + POSTAG + DEPREL + SPLFEATS. В этом наборе признаков из всех морфологических характеристик слов были оставлены лишь число, падеж и род, эти характеристики обозначены как SPLFEATS. Они использовались как отдельно, так и объединенные в одну строку.
- Base_SplFeats = Базовый + SPLFEATS = FORM + LEMMA + POSTAG + DEPREL + FEATS + SPLFEATS.

Качество синтаксического анализа измерялось с помощью двух оценок:

- UAS (unlabeled attachment score) – точность установления управляющего токена;
- LAS (labeled attachment score) – точность установления управляющего токена и метки синтаксической связи.

В Табл. 1 представлены значения UAS, рассчитанные при тестировании анализатора, обученного на разных наборах признаков на корпусе без пометок синтаксических связей. Все признаки за исключением КСК являются эталонными. Все оценки получены без учета пунктуации.

Отдельно проводился эксперимент с обучением анализатора на наборе признаков «Base» на корпусе, в котором полностью отсутствуют пометки синтаксических связей (все связи имеют метку «ROOT»). Значение UAS этого анализатора составило 86,7%.

В Табл. 2 представлены значения LAS и UAS, полученные при тестировании анализтаора, обученного на разных наборах признаков на корпусе с полным набором меток синтаксических связей. Все признаки за исключением КСК являются эталонными. Все оценки получены без учета пунктуации.

Лучшие результаты показывает анализатор, обученный на базовом наборе признаков (Base) и его модификациях. Однако такая модель сильно привязана к лексике и словоупотреблениям, присутствующим в обучающем корпусе. Анализатор, обученный на наборе признаков, в котором не учитываются формы слов (NoForm) показывает незначительное уменьшение UAS (на 0,2%), однако скорость обучения и анализа у него существенно выше, а требования к памяти ниже. Полное удаление признаков, учитывающих лексику (NoForm_NoLemma), сильно снижает точность анализа до 81,4%. При этом добавление КСК существительных (NoForm_NoLemma_CSC) в качестве признака хотя и дает небольшой прирост точности (на 0,7%), остается существенный разрыв между анализаторами с учетом лексики и без. Морфологические характеристики слов FEATS достаточно сильно влияют на качество анализа, если их не учитывать (NoFeats), UAS теряет 2,2%. Полная строка морфологических характеристик может быть заменена набором из трех характеристик: число, падеж, род (NoFeats_SplFeats), при незначительном снижении точности на 0,2%. Анализатор, который среди всех морфологических характеристик учитывает только три вышеперечисленных, во-первых, понижает требования к морфологическому анализатору, а, во-вторых, позволяет легче согласовывать морфологические характеристики разных стандартов. Это в свою очередь может упростить интеграцию в единую систему морфологического анализатора, выдающего характеристики, соответствующие одному стандарту, с синтаксическим анализатором, обученным на корпусе, разме-

Табл. 1. Точность анализатора, обученного на корпусе с двумя метками синтаксических связей на разных наборах признаков

Набор признаков	UAS,%
Base	87,3
NoForm	87,1
NoForm_NoLemma	81,4
NoForm_NoLemma_CSC	82,1
Base_CSC	87,4
NoFeats	85,1
NoFeats_SplFeats	87,2
Base_SplFeats	**87,5**

Табл. 2. Точность анализатора, обученного на корпусе с полным набором меток синтаксических связей на разных наборах признаков

Набор признаков	LAS,%	UAS,%
Base	82,9	88,0
Base_NoForm	82,8	87,8
Base_NoForm_NoLemma	70,6	81,5
Base_NoForm_NoLemma_CSC	71,6	82,4
Base_CSC	83,0	88,1
NoFeats	78,9	85,6
NoFeats_SplFeats	82,9	88,1
Base_SplFeats	**83,1**	**88,2**

ченным по другому стандарту, поскольку между этими стандартами необходимо согласовать только род, число и падеж (а также часть речи).

Заметим, что обучение на корпусе, содержащим две метки синтаксических связей («DEP» и «ROOT»), по сравнению с обучением на корпусе без синтаксических меток выгоднее, значения UAS анализатора в первом случае выше на 0,6%.

Анализаторы, обученные на корпусе с полным набором синтаксических меток показывают схожее соотношение точности по разным наборам признаков. Качество анализа несколько выше: UAS на 0,1 – 0,9% выше чем у анализаторов, обученных на корпусе с двумя метками. Однако анализатор, обученный на корпусе с полным набором меток на много отстает по производительности от анализатора, обученного на корпусе с двумя метками, и предъявляет более высокие требования к объему оперативной памяти. К примеру, при обучении на корпусе размером 30 000 предложений (434 814 токенов) требуется около 14 Гб ОЗУ для корпуса с полным набором меток, в то время как для корпуса с двумя метками требуется всего около 2Гб. В первом случае анализ 5 000 предложений (72 390 токенов) выполняется за 241 сек, тогда как во втором случае выполняется всего за 13 сек.

Была построена зависимость оценки точности синтаксического анализа от размера корпуса. При этом использовалась конфигурация «Base». Были отдельно построены зависимости для корпуса с полным набором меток синтаксических связей и для корпуса с двумя метками. В первом случае рассчитывались и LAS, и UAS во втором случае только UAS.

Корпус, в котором синтаксические связи помечены двумя метками, был поделен на тестовую и обучающую части следующим образом: тестовая содержит 5 439 предложений и 78 396 токенов (без учета пунктуации), часть для обучения, содержит максимум 48 000 предложений. Корпус, содержащий полный набор синтаксических связей был поделен на тестовую и обучающую части следующим образом: тестовая часть содержит 5 439 предложений и 78 396 токенов (без учета пунктуации), часть для обучения содержит максимум 35 000 предложений. Обучающий корпус наращивался с 5 000 предложений с шагом по 5 000 предложений. На Рис. 1 представлена зависимость UAS анализатора от размера обучающего корпуса с двумя метками синтаксических связей. На Рис. 2 представлены зависимости UAS и LAS анализатора от размера обучающего корпуса с полным набором синтаксических меток.

Разница между UAS анализаторов, обученных на корпусе с двумя метками синтаксических связей размером 5 000 предложений и размером 48 000 предложений, составляет в среднем 2,7%. Разница между UAS и LAS анализаторов, обученных на корпусе с полным набором меток размером 5 000 предложений и размером 35 000 предложений, составляет в среднем 2,6% и 4,1% соответственно. Интересно, что даже сравнительно небольшие корпуса

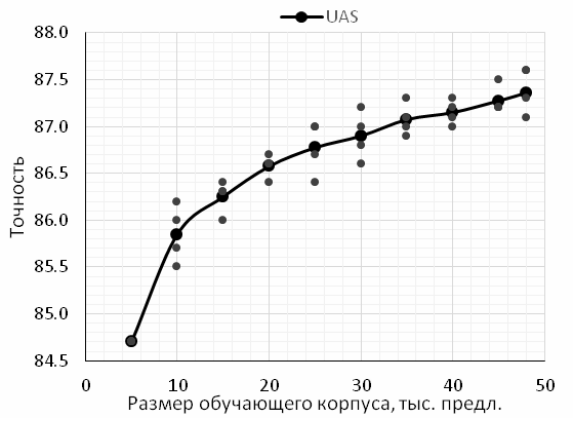

Рис. 1. Зависимость точности синтаксического анализа от размера обучающего корпуса с двумя метками синтаксических связей

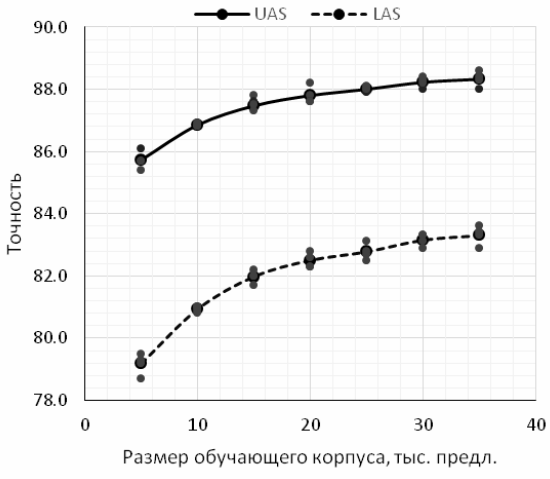

Рис. 2. Зависимость точности синтаксического анализа от размера обучающего корпуса с полным набором меток синтаксических связей

от 10 000 предложений позволяют обучать анализаторы, показывающие приемлемое качество анализа для решения многих прикладных задач.

3. Семантический анализ

Под семантическим анализом текста мы понимаем установление обобщенного категориального смысла синтаксем и семантических связей между ними. Синтаксема представляет собой минимальную синтактико-семантическую единицу языка [4], несущую обобщенный, категориальный смысл. В дальнейшем обобщенный категориальный смысл синтаксемы будем называть её значением. В этой статье внимание уделяется только вопросам автоматического установления значений синтаксем. Вопросы автоматическом установлении семантических связей между синтаксемами не рассматриваются. Задача семантического анализа в этом представлении близка к задаче semantic role labeling [3]. Для ее решения использовался метод реляционно-ситуационного анализа [2, 5, 40].

Основным источником информации для семантического анализа в этом методе является словарь предикатных слов. Он содержит словарные статьи, каждая из которых соответствует некоторой ситуации. В словарной статье содержатся предикатные слова (глаголы, девербативы), которые указывают какую ситуацию описывает данное предложение. Помимо этого, словарная статья содержит семантические значения, описывающие ролевую структуру ситуации (которые могут иметь синтаксемы в заданной ситуации), а также признаки, которыми должны обладать синтаксемы, чтобы им были сопоставлены те или иные значения. В качестве основных признаков используются падеж, предлог, а также категориально-семантический класс существительного (КСК). КСК – это обобщенное значение, которое характеризует слова, принадлежащие к одной из следующих категорий [41]:

- «предметное» – материальные сущности;
- «признаковое» – нематериальные сущности;
- «личное» – лицо, способное целенаправленно действовать, мыслить;
- «пространственное» – место в пространстве;
- «темпоративное» – момент времени;
- «единица_измерения» – имена со значением единиц измерения;
- «параметр_измерения» – имена со значением параметра измерения;
- «количественное» – имена со значением количества.

В словаре предикатных слов для русского языка содержится 2 856 словарных статей и 3 585 глаголов.

Перед анализом предложения разбиваются на клаузы, которые задают основную область поиска аргументов (синтаксем, которым могут быть назначены семантические значения) при предикатном слове. Для разбиения предложения на клаузы в наших экспериментах использовался программный пакет АОТ [42].

Процедуру семантического анализа предложения можно условно разделить на четыре этапа: поиск предикатных слов и соответствующих им словарных статей, поиск потенциальных аргументов, установление аргументам семантических значений, выбранных из каждой словарной статьи, выбор наилучшей словарной статьи и семантических значений с помощью решения оптимизационной задачи о назначениях.

При поиске предикатных слов в предложении ищутся те слова, которые присутствуют в словаре предикатных слов и для которых выполняется ряд заданных условий. В основном условия накладываются на часть речи и на положение в синтаксическом дереве, также проверяется, является ли предикатное слово неполнознаменательным глаголом. Затем из словаря выбираются словарные статьи, в которых присутствует найденное в предложении предикатное слово. Для каждого найденного в предложении предикатного слова ищутся потенциальные аргументы. Поиск ведется в соответствии с заданным набором правил, которые учитывают характеристики предикатного слова, характеристики слов в предложении, синтаксические связи. С каждым найденным потенциальным аргументом связывается его вес – число от 0 до 1. Вес отражает уверенность в том, что найденная синтаксема действительно является аргументом ситуации. Он определяется не статистически, а на основе заданного набора эвристик. Для каждой словарной статьи и для каждого потенциального аргумента на основе характеристик синтаксемы и информации, содержащейся в словарной статье, определяет-

ся набор семантических значений, которые могут быть назначены данному аргументу независимо от остальных аргументов. В большинстве случаев проверяется, что падеж, КСК и связанный с синтаксемой предлог, совпадают с записью в словарной статье, соответствующей некоторому семантическому значению. В случае, если предикатное слово является причастием, используется особая процедура проверки. Затем выбирается такое распределение ролей, при котором каждое значение для некоторой статьи используется не более одного раза и которое дает наилучшее покрытие семантических значений, содержащихся в словарной статье. Для этого используется алгоритм оптимизации, который основан на венгерском методе. При этом учитываются веса, установленные потенциальным аргументам. Подход, в котором распределение семантических значений по аргументам происходит с помощью решения задачи оптимизации применяется и в других системах семантического анализа, например, [43].

После того как оптимальное распределение семантических значений между потенциальными аргументами найдено для каждой статьи, выбирается наилучшая словарная статья (или ситуация). В качестве оценки статьи используется значение целевой функции при оптимальном решении соответствующей задачи о назначениях. В итоге синтаксемы получают семантические значения, которые составляют наилучшее покрытие значений, содержащихся в словаре предикатных слов.

Семантический анализатор, использующий словарь предикатных слов будем называть словарным. Для оценки качества семантического анализа был вручную создан семантически размеченный подкорпус СинТагРус. В подкорпусе размечены синтаксемы с семантическими значениями для тех случаев, которые присутствуют в словаре предикатных слов, категориально-семантические классы синтаксем, предикатные слова, а также связи между предикатными словами и синтаксемами. Подкорпус содержит около 1 500 предложений (около 27 000 токенов) и 3 300 токенов с семантическими значениями, среди которых 61 уникальное значение.

На тестовом семантическом подкорпусе была определена точность установления категориально-семантических классов существительных, которая определяется как доля синтаксем с семантическим значением, которым правильно назначен КСК, среди всех синтаксем с семантическим значением. Точность установления КСК составила 93,7%.

Качество семантического анализа оценивалось для четырех случаев:

- GoldSynt + GoldCSC: на вход семантическому анализатору подаются предложения, в которых синтаксические связи и КСК существительных являются эталонными.
- GoldSynt + CSC: на вход семантическому анализатору подаются предложения, в которых синтаксические связи – эталонные, а КСК существительных установлены автоматически.
- Synt + GoldCSC: на вход семантическому анализатору подаются предложения, в которых КСК существительных являются эталонными, а синтаксические связи установлены автоматически синтаксическим анализатором, созданном на основе MaltParser. Для построения деревьев синтаксического анализа использовалась конфигурация «Base_CSC + SplFeats», анализатор обучался на корпусе без меток синтаксических связей размером 48 096 предложениях (74 665 токенов). Значение UAS для него на семантически размеченном корпусе составляет 87,8%.
- Synt + CSC: на вход семантическому анализатору подаются предложения, в которых и синтаксические связи, и КСК существительных установлены автоматически. Результаты, полученные в этом случае, можно рассматривать как результаты системы, в которой синтаксический и семантический анализ выполняются отдельно и последовательно.

Во всех случаях использовались эталонные морфологические характеристики токенов. В качестве оценок качества семантического анализа использовались точность полнота и F_1-мера. При сопоставлении результатов анализа с тестовым подкорпусом учитывались только те семантические значения, которые присутствуют в тестовом корпусе. В этом случае полнота – это доля синтаксем, значение которых установлено верно, среди всех синтаксем, значение которых указано в тестовом корпусе; точность – это доля синтаксем, значение которых установлено верно, среди всех синтаксем, которые выделил анализатор и значение которых указано в тестовом корпусе; F_1-мера – это среднее гармоническое точности и полноты. Оценки качества семантического анализа приведены в Табл. 3.

Табл. 3. Оценки качества семантического анализа, выполненного словарным анализатором

Случай	Полнота %	Точность %	F_1-мера %
GoldSynt + GoldCSC	71,6	93,1	81,0
GoldSynt + CSC	61,4	87,2	72,1
Synt + GoldCSC	68,5	93,0	78,9
Synt + CSC	58,7	86,9	70,1

Как видно из представленной таблицы, анализатор, основанный на словаре предикатных слов, обладает низкой полнотой, но достаточно высокой точностью. Жесткое сравнение признаков синтаксем с признаками, содержащимися в словарной статье, не позволяет эффективно проводить обобщение и разрешать случаи неоднозначности в словаре. Ошибки также возникают из-за неполноты словаря предикатных слов. Кроме того, негативное влияние на качество анализа оказывают ошибки разбиения предложений на клаузы и ошибки при обработке некоторых случаев с причастными и деепричастными оборотами.

4. Семантико-синтаксический анализ

Для проведения семантико-синтаксического анализа было предложено использовать возможность MaltParser назначать метки синтаксическим связям, которые можно интерпретировать как семантические значения. Если синтаксическим связям анализатор в качестве метки назначает семантическое значение, то в результате работы анализатора одновременно будет построено и синтаксическое дерево предложения, и его семантическая структура. Стоит отметить, что не все синтаксемы синтаксически связаны с предикатными словами, поэтому по метке синтаксической связи не всегда можно однозначно восстановить семантическую связь между синтаксемой и предикатным словом.

Чтобы обучить анализатор выполнять и синтаксический, и семантический анализ необходим корпус, в котором размечены и синтаксические связи, и семантические значения из необходимого нам набора. Поскольку для русского языка отсутствуют доступные крупные корпуса, содержащие и синтаксическую, и семантическую разметку, было предложено провести обучение на корпусе СинТагРус с семантической разметкой, полученной автоматически.

Для семантической разметки корпуса использовался словарный семантический анализатор, описанный в разделе 3. На вход анализатору подавались эталонные синтаксические деревья, эталонные морфологические характеристики токенов, но при этом использовались КСК существительных, полученные автоматически (случай GoldSynt + CSC). Метки синтаксических связей, которые изначально присутствовали в СинТагРус были удалены.

При обучении анализатора семантико-синтаксическому анализу использовалась та же конфигурация, что и при обучении анализатора синтаксическому анализу (раздел 2). Эксперименты проводились с четырьмя наборами признаков:

- P1 = Base (из раздела 2);
- P2 = Base (из раздела 2) + КСК;
- P3 = P2 + SPLFEATS (отдельно число, падеж и род);
- P4 = P3 + идентификаторы словарных статей предикатных слов, предсказанные словарным анализатором. Эти идентификаторы несут информацию о том, какую ситуацию выражает предложение.

КСК были получены автоматически, все остальные признаки (лемма, морфологические характеристики, часть речи) были получены из корпуса. Обучение проводилось на корпусе, содержащим 48 096 предложений (699 708 токенов). В качестве оценок качества семантического анализа использовались точность, полнота и F_1-мера, рассчитанные аналогично оценкам из раздела 3. Они представлены в Табл. 4.

Точность синтаксического анализа на разных наборах признаков практически не меняется, UAS принимает значения от 87,6 – 87,8%, что соответствует лучшим оценкам точности синтаксических анализаторов, обученных на корпусе с двумя метками синтаксических связей, описанных в разделе 2.

Табл. 4. Оценки качества семантического анализа, выполненного семантико-синтаксическим анализатором на основе MaltParser

Набор признаков	Полнота %	Точность %	F_1-мера %
P1	54,6	84,4	66,3
P2	58,8	86,2	69,9
P3	**60,1**	**86,8**	**71,0**
P4	58,4	85,9	69,5

Наилучший результат показал анализатор, обученный на наборе признаков «P3»: FORM + LEMMA + POSTAG + DEPREL + FEATS + КСК + SPLFEATS (отдельно число, падеж и род). Значение F_1-меры составило 71,0%, что на 0,9% выше оценки для словарного семантического анализатора, в случае, когда и КСК, и синтаксические деревья не являются эталонными. Семантико-синтаксический анализатор обладает большей полнотой (выше на 1,4%) при несколько меньшей точности (меньше на 0,1%).

Качество семантико-синтаксического анализатора зависит от качества обучающего корпуса, созданного словарным анализатором. Таким образом, семантико-синтаксический анализатор подвержен схожим ошибкам, что и словарный анализатор. Поэтому необходимо повышать качество обучающего корпуса. С одной стороны, можно улучшать качество словарного анализатора. С другой стороны, можно попытаться исключить из обучающего корпуса «выбросы» – предложения с большим количеством ошибок семантического анализа. Отсеяв некоторую долю предложений обучающего корпуса, на которых семантико-синтаксический анализатор делает ошибки, можно обучиться на очищенной части корпуса и получить более точный анализатор.

Разработанный семантический-синтаксический анализатор, в отличие от словарного анализатора, может быть использован в тех случаях, когда словарная статья для предикатного слова из предложения текста отсутствует. Словарный анализатор в этом случае не может установить семантические значения для неизвестного ему предикатного слова. Таким образом, семантико-синтаксический анализатор может использоваться для увеличения полноты анализа. Кроме того, применение методов машинного обучения позволяет выявить признаки, которые влияют на синтаксический и семантический анализ.

Заключение

В работе исследованы подходы к семантико-синтаксическому анализу текстов. Показана возможность совместного семантического и синтаксического анализа текстов на русском языке, выполняемого одним методом, с качеством, соответствующим качеству разделенных последовательных этапов анализа. Выявлены признаки, оказывающие наибольшее влияние на качество семантико-синтаксического анализа.

Дальнейшие исследования в области семантико-синтаксического анализа планируется вести по трем направлениям:

- Пополнение тестового семантического корпуса, поиск ошибок в нем и его верификация. Для автоматизации этого процесса предлагается использовать созданные анализаторы.
- Модификация и разработка новых методов семантико-синтаксического анализа. Планируется создать интегрированную систему семантико-синтаксического анализа, которая сочетает в себе словарный анализатор и анализатор, построенный на основе машинного обучения. Результаты семинаров CoNLL Shared Task показывают, что комбинация анализаторов может существенно увеличить качество семантического анализа. Планируется продолжить выявление признаков, важных для семантико-синтаксического анализа, в частности будет оценено влияние типов синтаксических связей.
- Применение разработанных методов для решения конкретных прикладных задач, таких как вопросно-ответный поиск и извлечение информации из текстов.

Литература

1. Смирнов И.В., Шелманов А.О. Семантико-синтаксический анализ естественных языков. Часть I. Обзор методов синтаксического и семантического анализа текстов // Искусственный интеллект и принятие решений. — 2013. — № 1. — С. 41–54.
2. Osipov G. Methods for extracting semantic types of natural language statements from texts // 10th IEEE International Symposium on Intelligent Control. — Monterey, California, USA, 1995. — aug.
3. Gildea D., Jurafsky D. Automatic labeling of semantic roles // Computational Linguistics. — 2002. — Vol. 28, no. 3. — P. 245–288.
4. Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка // Институт русского языка РАН им. В. В. Виноградова. — 2004.
5. 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.
6. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies / Mihai Surdeanu, Richard Johansson, Adam Meyers et al. // Proceedings of the Twelfth Conference on Computational Natural Language Learning / Association for Computational Linguistics. — 2008. — P. 159–177.

7. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages / Jan Hajic, Massimiliano Ciaramita, Richard Johansson et al. // Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task / Association for Computational Linguistics. — 2009. — P. 1–18.
8. Lluìs X., Màrquez L. A joint model for parsing syntactic and semantic dependencies // Proceedings of the Twelfth Conference on Computational Natural Language Learning / Association for Computational Linguistics. — 2008. — P. 188–192.
9. Lluìs X. Joint Learning of Syntactic and Semantic Dependencies : Ph.D. thesis / Xavier Lluìs ; Master Thesis, Universitat Politecnica de Catalunya (Artificial Intelligence Program), Barcelona. — 2008.
10. Lluìs X., Bott S., Màrquez L. A second-order joint eisner model for syntactic and semantic dependency parsing // Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task / Association for Computational Linguistics. — 2009. — P. 79–84.
11. Eisner J. M. Three new probabilistic models for dependency parsing: An exploration // Proceedings of the 16th conference on Computational linguistics. — Vol. 1. — 1996. — P. 340–345.
12. A latent variable model of synchronous parsing for syntactic and semantic dependencies / James Henderson, Paola Merlo, Gabriele Musillo, Ivan Titov // Proceedings of the Twelfth Conference on Computational Natural Language Learning / Association for Computational Linguistics. — 2008. — P. 178–182.
13. A latent variable model of synchronous syntactic-semantic parsing for multiple languages / Andrea Gesmundo, James Henderson, Paola Merlo, Ivan Titov // Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task / Association for Computational Linguistics. — 2009. — P. 37–42.
14. Multi-lingual joint parsing of syntactic and semantic dependencies with a latent variable model / James Henderson, Paola Merlo, Ivan Titov, Gabriele Musillo // Computational Linguistics. — 2013.
15. Titov I., Henderson J. A latent variable model for generative dependency parsing // Proceedings of the 10th International Conference on Parsing Technologies. — IWPT '07. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2007. — P. 144–155.
16. Mixing and blending syntactic and semantic dependencies / Yvonne Samuelsson, Oscar Täckström, Sumithra Velupillai et al. // Proceedings of the Twelfth Conference on Computational Natural Language Learning / Association for Computational Linguistics. — 2008. — P. 248–252.
17. Dai Q., Chen E., Shi L. An iterative approach for joint dependency parsing and semantic role labeling // Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task / Association for Computational Linguistics. — 2009. — P. 19–24.
18. Johansson R., Nugues P. Dependency-based syntactic-semantic analysis with PropBank and NomBank // Proceedings of the Twelfth Conference on Computational Natural Language Learning / Association for Computational Linguistics. — 2008. — P. 183–187.
19. Chen E., Shi L., Hu D. Probabilistic model for syntactic and semantic dependency parsing // Proceedings of the Twelfth Conference on Computational Natural Language Learning / Association for Computational Linguistics. — 2008. — P. 263–267.
20. Sun W., Li H., Sui Z. The integration of dependency relation classification and semantic role labeling using bilayer maximum entropy markov models // Proceedings of the Twelfth Conference on Computational Natural Language Learning / Association for Computational Linguistics. — 2008. — P. 243–247.
21. Morante R., Van Asch V., Van den Bosch A. Joint memory-based learning of syntactic and semantic dependencies in multiple languages // Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task / Association for Computational Linguistics. — 2009. — P. 25–30.
22. 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.
23. Каневский Е. А., Боярский К. К. Семантико-синтаксический анализатор SemSin // Международная конференция «Диалог 2012». Доклады, принятые к публикации на сайте. — URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Kanevsky.pdf.
24. Кузнецов И. П. Методики выявления объектов и связей, заданных в неявном виде // Международная конференция «Диалог 2012». Доклады, принятые к публикации на сайте. — 2012. — URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Кузнецов_И_П.pdf.
25. Ермаков А. Е., Плешко В. В. Семантическая интерпретация в системах компьютерного анализа текста // Информационные технологии. — Т. 6. — С. 2–7.
26. Кашкин Е., Ляшевская О. Н. Семантические роли и сеть конструкций в системе FrameBank // Труды международной конференции «Диалог 2013». — 2013. — С. 325–343.
27. 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.
28. Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Papers from the Annual International Conference "Dialogue" (2011). — No. 10. — 2011. — P. 17.
29. 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.
30. MaltParser. — 2013. — дек. — URL: http://maltparser.org/.
31. Синтаксически размеченный корпус русского языка: инструкция пользователя. — 2013. — дек. — URL: http://www.ruscorpora.ru/instruction-syntax.html.
32. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы / Ю. Д. Апресян, И. М. Богуславский,

Б. Л. Иомдин и др. // Национальный корпус русского языка: 2003–2005. — 2005. — С. 193–214.
33. Buchholz S., Marsi E. CoNLL-X shared task on multilingual dependency parsing // Proceedings of the Tenth Conference on Computational Natural Language Learning / Association for Computational Linguistics. — 2006. — P. 149–164.
34. Designing and evaluating a Russian tagset / Sharoff Serge, Kopotev Mikhail, Erjavec Tomaz et al. // Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). — Marrakech, Morocco : European Language Resources Association (ELRA), 2008. — may.
35. MULTEXT-East morphosyntactic specifications, version 4. — 2013. — дек. — URL: http://nl.ijs.si/ME/V4/msd/html/msd-ru.html.
36. Chang C.-C., Lin C.-J. LIBSVM: A library for support vector machines // ACM Transactions on Intelligent Systems and Technology. — 2011. — Vol. 2. — P. 27.
37. LIBLINEAR: A library for large linear classification / Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh et al. // The Journal of Machine Learning Research. — 2008. — Vol. 9. — P. 1871–1874.
38. Russian statistical taggers and parsers. — 2013. — дек. — URL: http://corpus.leeds.ac.uk/mocky/.
39. Nivre J., Nilsson J. Pseudo-projective dependency parsing // Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics / Association for Computational Linguistics. — 2005. — P. 99–106.
40. Relational–situational method for intelligent search and analysis of scientific publications / Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Artem Shelmanov // Proceedings of the Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13). — Vol. 968. — Moscow, Russia : CEUR Workshop Proceedings, 2013.
41. Осипов Г. С. Методы искусственного интеллекта. — ФИЗМАТЛИТ, 2011.
42. Автоматическая обработка текста. — 2013. — дек. — URL: http://www.aot.ru/.
43. Punyakanok V., Roth D., Yih W.-t. The importance of syntactic parsing and inference in semantic role labeling // Computational Linguistics. — 2008. — Vol. 34, no. 2. — P. 257–287.

Смирнов Иван Валентинович. Старший научный сотрудник Института системного анализа РАН. Окончил Российский университет дружбы народов в 2003 году. Кандидат физико-математических наук. Автор 36 печатных работ. Область научных интересов: 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 5 печатных работ. Область научных интересов: искусственный интеллект, компьютерная лингвистика, информационно-аналитические системы, машинное обучение. E-mail: shelmanov@isa.ru

Кузнецова Екатерина Сергеевна. Онтоинженер компании ABBYY. Окончила Московский государственный университет им. М.В. Ломоносова в 2012 году. Автор одной печатной работы. Область научных интересов: анализ тональности, информационный поиск, формальные модели языка. E-mail: knnika@yandex.ru

Храмоин Иван Валерьевич. Инженер-исследователь Института системного анализа РАН. Окончил Российский университет дружбы народов в 2012 году. Автор одной печатной работы. Область интересов: интеллектуальные системы управления, интеллектуальное планирование траектории, методы интеллектуального поиска, методы выявления мнений. E-mail: hramoin@isa.ru