ОТЗЫВ

официального оппонента на кандидатскую диссертацию

Шелманова Артема Олеговича

«Исследование методов автоматического анализа текстов и разработка интегрированной системы семантико-синтаксического анализа»,

представленную на соискание ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики»

Кандидатская диссертация А.О. Шелманова посвящена созданию системы автоматического синтактико-семантического анализа русского текста высокого уровня, обеспечивающей выявление и фиксацию смысловой структуры текста как основы для решения комплекса задач дальнейшей его смысловой обработки.

Давно было понято, что успешное решение целого комплекса взаимосвязанных задач информационного обеспечения (работы с информацией, выраженной в текстах) – поиск информации по запросам различных типов, аннотирование, реферирование, группирование по смыслу и т.п. – может быть достигнуто на основе экспликации семантической (ролевой) структуры текстов, содержащих эту информацию. Проблема состояла в том, как эту экспликацию осуществить в автоматическом (автоматизированном) режиме, в частности для текстов, написанных на русском языке.

По мере развития средств анализа текстовой информации происходил переход ко все более высоким уровням анализа – от простейшего, на котором смысл текста представлялся неупорядоченным набором входящих в него слов или словосочетаний, к самому распространенному на сегодняшний день представлению его в виде цепочек слов, характерному, например, для наиболее типичных «поисковых машин» Интернета, таких как Гугл или Яндекс, и, наконец, к наиболее полному и точному представлению текста в виде сети, или графа, входящих в него понятий с отражением их ролевой структуры – выполняемых ими смысловых ролей. Последнее предполагает выявление как синтаксической, так и семантической структуры текста на уровне как отдельных предложений, так и текста в целом. Комплекс связанных с этим задач представляет собой основное направление исследовательских работ в данной области.

В нем можно выделить различные подходы – в частности, ориентированные на одноразовую последовательную реализацию этапов синтаксического и семантического анализа; ориентированные на совместное («вперемешку») использование операций синтаксического и семантического анализа; ориентированные на последовательную итеративную реализацию этапов синтаксического и семантического анализа с коррекцией возможных ошибок предыдущих этапов. Одним из самых интересных и актуальных направлений исследований и практических разработок этого последнего типа является направление, реализуемое за последние годы в Институте системного анализа РАН. Именно в рамках этого направления было проведено диссертационное исследование А.О. Шелманова, основным результатом которой явилась разработка интегрированной системы синтактико-семантического

анализа[1]. Поэтому тема представленной диссертационной работы является **актуальной**.

Диссертация состоит из введения, четырех глав, заключения, списка литературы и четырех приложений.

В первой главе последовательно рассматриваются и оцениваются существующие подходы к решению задач синтаксического и семантического анализа текста, различные используемые для этого модели синтаксиса и семантики текста, в том числе модель семантики текста на основе ролевой структуры предложения, и в заключение – реляционно-ситуационная модель, основанная на теории коммуникативной грамматики Г.А. Золотовой, в которой центральное место занимает понятие «синтаксемы» – минимальной словесной конструкции, имеющей обобщенное смысловое значение (из приводимых автором примеров: «в лесу» – место действия, «из леса» - исходный пункт действия, «в лес» – конечный пункт действия, «топором» – орудие действия и т.п.); использование этой теории позволяет выявлять и фиксировать ролевую структуру текста как основу для решения перечисленных выше информационных задач. Справедливо отмечается слишком малое число исследований, посвященных задаче автоматического определения ролевых структур высказываний в текстах на русском языке, и отчасти объясняющая этот факт недостаточность размеченных корпусов, которые могли бы играть роль обучающей выборки и «золотого стандарта» при тестировании систем анализа.

В разделе 1.3.2 описываются методы, сочетающие синтаксический (в варианте деревьев зависимости) и семантический (определение ролевых структур высказываний) анализ для определения ролевой структуры предложения. В выводах главы (раздел 1.4) обосновывается выбор принятого автором оригинального варианта – сочетания синтаксического и семантического анализа, состоящего в последовательном применении сначала синтаксического, затем семантического анализа, за которым следует повторное применение синтаксического анализа, корректирующего (при необходимости) результаты первичного синтаксического анализа с последующей соответствующей коррекцией результатов первичного семантического анализа.

Вторая глава посвящена описанию разработанных методов семантического и (в терминологии автора) семантико-синтаксического анализа. В основу обоих методов положен семантический словарь, статьи которого содержат предикатные слова, задающие некоторую обобщенную ситуацию, в рамках которой действуют ее участники – аргументы, играющие свои специфические роли. Роли, задаваемые семантическим словарем могут быть обязательными (основными) или необязательными (периферийными); они обрабатываются по-разному, как и описания ситуаций, содержащие вопросительные слова (*кто*, *что*, *где*, *когда* и т.п.). При анализе предложения последовательно выявляются аргументы, роль которых обязательная, затем аргументы, включающие

---

[1] Автор использует термин «семантико-синтаксический анализ»; я предпочитаю вариант «синтактико-семантический», поскольку во всех вариантах такого анализа он начинается с построения первичного синтаксического дерева (ср. с. 87).

вопросительные слова и, наконец (если они имеют место) аргументы, роль которых необязательна. Роли аргументам назначаются в соответствии с их грамматическими и категориально-семантическими признаками. Снятие неоднозначности назначения ролей аргументам осуществляется с помощью решения оптимизационной задачи о назначениях. В разработанном автором методе семантико-синтаксического анализа стоит отметить оригинальную схему взаимодействия процедуры построения синтаксических деревьев зависимостей с процедурой определения ролевых структур высказываний. После первоначального синтаксического анализа предложения, в нем ищутся синтаксические конструкции, которым семантический анализатор способен назначить семантическую роль при заданном предикатном слове, но из-за потенциальных ошибок в первоначальном синтаксическом дереве они не удовлетворяют стандартным правилам выявления семантических аргументов. Если эти конструкции удовлетворяют ряду эвристик, то они помечаются как «дополнительные» аргументы предикатного слова и используются для корректировки синтаксического дерева предложения. Преобразования в синтаксическом дереве также проходят ряд проверок с помощью эвристик и классификаторов на основе машинного обучения. На откорректированном таким образом синтаксическом дереве заново проводится семантический анализ.

В третьей главе описываются проведенные автором исследования эффективности разработанной им системы анализа текста. При этом в качестве «золотого стандарта» (эталона) используется единственный существующий в настоящее время представительный корпус русских текстов с ручной синтаксической разметкой – СинТагРус, разработанный в ИППИ РАН группой исследователей под руководством Ю.Д. Апресяна. Проведенные эксперименты с достаточной надежностью показывают весьма высокое качество анализа русскоязычных текстов, обеспечиваемое разработанной автором системой, по сравнению с другими известными подходами.

С точки зрения перспективы дальнейших исследований наиболее существенным представляется вывод автора (на с. 123), что «повышение качества как синтаксического анализа, так и семантического анализа при использовании разработанного метода семантико-синтаксического анализа2 сильно связано с полнотой семантического словаря».

В четвертой главе рассматриваются и оцениваются методы решения прикладных задач обработки текстов на естественном языке, в которых используются результаты семантического и семантико-синтаксического анализа:

– построение семантической сети текста;

– реализация вопросно-ответного поиска, в частности в рамках метапоисковой системы Exactus, сводящей воедино результаты поиска, осуществляемого различными поисковыми системами Интернета, и ранжирующей их с использованием средств синтактико-семантического анализа; проведенные эксперименты показали, что результаты поиска, произведенного с

---

[2] А по моему мнению – не только данного метода.

использованием семантического анализа, заметно превосходят результаты, полученные без его использования, а результаты поиска, произведенного с использованием – в терминологии автора – семантико-синтаксического анализа превосходят результаты поиска, проведенного с использованием только семантического анализа, хоть и не столь существенно (до 4%), но устойчиво;

– автоматического извлечения определений и авторских терминов из текстов научных публикаций; результаты проведенных экспериментов показывают увеличение полноты извлечения на 10 % по сравнению со способами, использующими более слабые семантико-синтаксические средства.

К числу недостатков рецензируемой работы можно отнести разве что нечеткость или спорность некоторых формулировок. Так, на с. 5 диссертации говорится:

«Подчинение обуславливается набором общих принципов, которые в целом сводятся к тому, что зависимое слово в предложении является уточняющим, необязательным, менее важным для передачи смысла высказывания, чем главное».

С этим вряд ли можно безусловно согласиться. Достаточно типичным является случай, когда главное слово является достаточно общим, в какой-то мере «пустым»; например, в словосочетаниях «синтаксический анализ» или «семантический анализ» главное слово «анализ» является чрезвычайно общим (особенно в случае научного текста), и только зависимые слова «синтаксический» и «семантический» придают ему определенное значение, в отличие, например, от «химического» или «ситуационного» анализа.

Неудачным представляется определение синтаксического анализа на с. 15 как «сопоставление лексем некоторого языка (естественного или формального) с его формальной грамматикой»: лексемы и грамматика – объекты разной природы, они несопоставимы.

Приведенные замечания не снижают общей высокой оценки диссертационной работы А.О. Шелманова. Она является значимым вкладом в разработку средств автоматизации смыслового анализа текстов на естественном (в данном случае русском) языке, применимых к решению ряда актуальных информационных задач; она намечает направления дальнейших исследований, среди которых – повышение эффективности средств анализа и переход от анализа ролевой структуры отдельных предложений к анализу ролевой структуры текста.

Научную новизну диссертации составляют разработанные автором метод определения ролевых структур высказываний в русскоязычных текстах, опирающийся на теорию коммуникативной грамматики русского языка Г.А. Золотовой, и метод семантико-синтаксического анализа, в котором реализовано взаимодействие между процедурой определения ролевых структур высказываний и процедурой построения синтаксического дерева зависимостей.

Достоверность и обоснованность выводов, научных положений и рекомендаций, сформулированных в диссертации, подтверждена экспериментальными исследованиями. Результаты работы **обоснованы и**

достаточно полно отражены в публикациях автора. А.О. Шелмановым по теме диссертации опубликовано 7 работ: 4 из них в рецензируемых изданиях из списка ВАК РФ и приравненных к ним, 2 статьи опубликовано в трудах российских и международных конференций, получено 1 свидетельство о государственной регистрации программы для ЭВМ.

Автореферат в полной мере отражает содержание диссертации.

Диссертационная работа полностью соответствует критериям, установленным Положением о присуждении ученых степеней: в ней содержится решение задачи, имеющей существенное значение для развития методов поиска и анализа текстовой информации. Диссертация написана автором самостоятельно, обладает внутренним единством и содержит новые научные результаты, полученные автором лично. Работа А.О. Шелманова представляет собой научное исследование, удовлетворяющее всем требованиям ВАК РФ, предъявляемым к кандидатским диссертациям, а ее автор заслуживает присуждения ему ученой степени кандидата технических наук по специальности 05.13.17 – Теоретические основы информатики.

Официальный оппонент,
руководитель учебно-научного центра
программного и лингвистического
обеспечения интеллектуальных систем
Российского государственного гуманитарного
университета (РГГУ),
д.т.н., с.н.с. Д.Г. Лахути

«15» сентября 2015 г.

Делир Гасемович Лахути
125993, ГСП-3, Москва, Миусская пл., д. 6
Тел. 8(495)2506329
E-mail: delir1@yandex.ru

Подпись Д.Г. Лахути заверяю
Ученый секретарь РГГУ Л.В.Тропкина