УДК 004.738.52

О.С. Волковский, Е.Р. Ковылин

АНАЛИЗ СОВРЕМЕННЫХ ПОДХОДОВ К ЗАДАЧЕ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Аннотация. Дана оценка существующим подходам к реализации систем автоматической генерации текстов на естественном языке. На основе проведенных исследований выбран наиболее перспективный из рассмотренных методов (с точки зрения осмысленности получаемых результатов).

Ключевые слова: автоматическая генерация текста, семантический анализ, генеративная грамматика, сематическая сеть, нейронная модель языка, теория «Смысл – Текст».

Введение. Направление обработки текстов на естественном языке включает в себя множество задач – от обычного синтаксического анализа до более сложных целей, вектор которых уходит глубоко в искусственный интеллект. Наиболее интересным, и, вместе с тем наименее реализованным на прикладном уровне (на данный момент) является процесс автоматической генерации текстов на естественном языке (ЕЯ). Сложности, которые возникают у разработчиков и исследователей систем, генерирующих текст, являются фундаментальными. Во-первых, ЕЯ не обладает постоянной структурой, необходимой для построения любой компьютерной системы. Во-вторых, возникает проблема описания смысловой составляющей текста. Понимание текста требует создания мощного компонента искусственного интеллекта, который мог бы работать с семантическими базами знаний, адаптировать систему для новых знаний и обучаться. Достаточно важной проблемой является неоднородность языка, как следствие большого разнообразия грамматических форм и правил, не позволяющих выделить некую единую структуру. Так, если мы имеем некий набор правил для языка с жестким порядком слов в предложении (как, например, английский язык), мы не сможем применить его для флективно богатого языка (языки славянской группы). Эти не-

сколько проблем прекрасно иллюстрируют основные вопросы в разработке системы понимания текста. Как описать алгоритм обработки текста, если нет конечного набора правил? Какими средствами описывать и представлять семантику текста? Как адаптировать систему под разные языковые группы? Ответов на эти вопросы мы не получили до сих пор – для подтверждения этого достаточно посмотреть на результаты работы систем автоматического перевода, явно далеких от идеала. Рынок систем синтеза текста представлен же по большей части западными англоязычными разработками, слабо применимыми для других языков. Поэтому исследования в области автоматической обработки текста не теряют свою актуальность до сих пор. В этой работе осуществлена попытка проанализировать состояние вопроса о генерации текстов в этой области на данный момент.

Постановка проблемы. Говоря именно об интеллектуальных системах автоматической обработки текста, мы сталкиваемся с несколькими научными вопросами. Алгоритмические проблемы связаны с необходимостью выбора аппарата описания и реализации ЕЯ на уровне и в форме, доступной машине. Лингвистические проблемы связаны напрямую со свойствами языка - его непостоянностью и неоднородностью правил его описания. Кроме того, говоря о языке как об объекте моделирования, встает вопрос об оценке полученных результатов - недостаточно просто получить текст, необходим глубокий оценочный анализ результатов работы системы. Ведь ЕЯ это не просто набор слов, связанный грамматическими правилами - приоритетной задачей является получение именно осмысленного текста, что, в свою очередь, приводит многих разработчиков к необходимости учета семантических связей не только между отдельными словами, но и между предложениями и даже между документами. С этой точки зрения, наиболее сложной и интересной является именно генерация новых текстов, реализация которой будет наиболее полно учитывать все важные смысловые связи в документе (или игнорировать не значимые).Менее очевидной сложностью является разрыв между лингвистическими описанием языка и его прикладной реализацией. Лингвистика ориентируется в первую очередь на описание природы языка как своего объекта, манипулируя, зачастую, понятиями из психологии, философии, антропологии и подобным, недостаточно формализированным наукам. При реализации систем автоматической

обработки текстов (АОТ) разработчики, используя инструменты компьютерной науки, вынуждены адаптировать их для работы сЕЯ, решая проблемы отнюдь не относящиеся к классической лингвистике. Это и породило такую гибридную область науки как компьютерную лингвистику, объектом которой является уже математическое моделирование ЕЯ.

Помимо всего вышесказанного, необходимо учитывать, что в понятие генерации текста могут входить весьма разнообразные системы. Так, к примеру, автоматическое составление некоторых шаблонных документов (типичный пример - автоматическая документизация программного кода) не может сравниться по сложности и семантическим свойствам с составлением аннотаций и квазирефератов к текстам на ЕЯ, хотя обе системы генерируют на выходе некоторый текст. Поэтому, в рамках этой работы будут рассмотрены, как подходы к созданию систем генерации осмысленных текстов, оперирующих флективно богатой текстовой информацией и некоторыми семантическими данными, так и прикладные разработки, реализующие этим модели.

Цель исследования: Провести анализ и оценку существующих подходов искусственного интеллекта и компьютерной лингвистики к созданию систем автоматической генерации текста с целью определения наилучшего из них по критериям степени реализуемости, адаптивности и интеллектуальности использующих их систем.

Основная часть: В качестве примера рассмотрим системы, подходами к созданию которых являлись три базовых концепции АОТ – генеративные грамматики Хомского, семантическая сеть и инструменты нейронных сетей. Основной проблемой при попытке проанализировать результаты работы и функциональные особенности выбранных методовявляется отсутствие возможности получить какие-либо их прикладные реализации для славянской группы языков – опираться мы можем только на описание систем их авторами и собственный теоретический анализ.

Первой мы рассмотрим систему синтеза учебных тестов на основе формальных грамматик Хомского, описанную в работе [1]. Несмотря на такие плюсы как возможность быстрой генерации текстов (тестов) и гибкость внедрения, система и подход, описанный Сорокиным С.И не решает проблем, поставленных нами в начале работы.

Для осознания этого, необходимо обратиться к теории и практическому применению генеративной грамматики в целом. Генеративная грамматика отталкивается от предположения о существовании явления языковой компетенции – врожденной способности человека к усвоению и пониманию человеческой речи, независимо от языка. Следуя этому, генеративная грамматика ставит перед собой цель смоделировать эту способность в рамках порождения правильных предложений, используя определенный конечный набор правил, алфавит и начальный символ предложения, из которого, с помощью правил, можно разворачивать бесконечное множество правильных схем построения предложения – непосредственные составляющие.

Границу применения генеративной грамматики подводит Мозговой М.М.: «….грамматики Хомского предназначены, прежде всего, для описания структуры предложения. Не менее важный вопрос описания смыслов отдельных слов остаётся за пределами их возможностей» [2]. Действительно, генеративная грамматика Хомского никогда не выходила за пределы уровня синтаксиса. Первоначальной целью является вывод грамматически правильных предложений из некоторого алфавита, используя цепочки глубинного уровня. И если теоретически мы можем выводить бесконечно большое количество таких цепочек, что, собственно и позволяет описать абсолютно любой ЕЯ, то на практике это не представляется возможным – даже если отбросить такое свойство языка как изменчивость, количество цепочек будет хоть и не бесконечно, но, безусловно, огромно. И чем более выражена флексия в языке – тем сложнее будет его описать.

Это еще одна проблема для грамматик Хомского. Флексия объясняется наличием определенного набора окончаний, которые морфологически и лексически меняют структуры слов, чаще всего – в зависимости от контекста. Чем флективно богаче язык, тем он сложнее, и тем свободенее порядок слов в предложении. Это нас приводит к проблеме языковой зависимости грамматик Хомского – реализация их для английского языка, языка с относительно бедной флексией и жестким порядком слов в предложении – не подходит для славяноязычных систем: «…в грамматиках Хомского порядок слов указывается непосредственно, поэтому проблема линеаризации вообще не возникает. Однако тем самым резко снижается выразительная мощь модели. Для английского языка с его строгим порядком слов в предложении

 ISSN 1562-9945

ограничения метода Хомского не являются критическими, однако при работе с русским языком их уже нельзя игнорировать» [2]. Наконец, взглянем на применение генеративных грамматик: «...задача анализа формального языка (возникающая, например, при компиляции программ на Паскале) не является тривиальной: она была полностью решена лишь в 60-70-е гг. после появления работ Н. Хомского...»[2]. Действительно, основная область использования порождающей грамматики – парсинг языка программирования, где глубинные структуры представлены цепочками формального языка, и грамматическая правильность стоит над смысловой семантикой. Тестовое множество, описанное в роботе, задается заранее неким конечным набором правил, и имеет определенное сходство с языком программирования в плане семантических связей. Однако, к таким задачам как автоматический перевод, рубрикация, неограниченная генерация текстов – подход малоприменим: «Остается неясным, как в рамках какой-либо из ветвей генеративной теории превратить некую семантическую сеть в последовательность деревьев зависимостей или составляющих, отвечающих отдельным предложениям»[3].

Следующим шагом развития (или же вспомогательным подходом) АОТ стало появление технологии семантических сетей, применение которых к задаче генерации текста мы оценим на основе системы автоматического консультирования, описанной в работе [4].Основной задачей, которую ставят перед собой авторы разработки, является генерация базы знаний (БЗ) конкретной предметной области для обеспечения диалога с пользователем по соответствующим ей вопросам. Семантическую сеть предлагается использовать для хранения извлекаемых знаний.

Несмотря на простоту реализации, такой подход имеет массу недостатков, наиболее значимыми из которых является слабая адаптивность системы и тяжелый процесс переопределения БЗ при смене предметной области – в этих случаях, заранее заданные шаблоны могут сработать во вред системы. Кроме того, отсутствует схема принятия решения при поступлении информации, не совпадающей с шаблоном. И это оправдано, поскольку подходы, базирующиеся на синтаксически-лексическом анализе вполне пригодны для относительно простых лингвистических задач, на подобии квазиреферирования или рубрикации с заранее известными рубриками. Однако для нашей за-

дачи автоматической генерации текста, (или ряда других, таких как автоматический перевод) методы остаются неприменимы. Очевидна необходимость наличия в системе некоторого интеллектуального компонента отвечающего за автоматический анализ семантической составляющей текста.

Выходом из этого тупика частично стают инструменты искусственного интеллекта, а именно – реализация АОТ при помощи нейронных сетей. Искусственные нейронные сети сегодня повсеместно используются для решения различных прикладных задач, в том числе и задач автоматической обработки языка. Чаще всего используются 3 типа сетей: простая рекуррентная сеть для обработки последовательностей, таких как предложения; рекурсивная авто-ассоциативная память для обработки лингвистических структур, представленных в виде деревьев и самоорганизующаяся карта Кохонена для кластеризации таких представлений.Для оценки качества методов нейронных сетей в задаче автоматической генерации текстов обратимся к работе [5], где искусственная реккурентная сеть используется для создания описаний о товарах некоторого интернет магазина. Результаты работы системы приведены в таблице:

Таблица 1

Результаты работы реккурентной нейронной сети

Качество звука, удобный интерфейс, очень долго держит заряд. Отзывчивый экран, громкий звонок, крупный шрифт, рабочий день. Приятно лежит в руках, 2 симкарты выручают. Качество сборки, батарея, удобное меню, устойчив к воздействию воды. Явно лидируют, сочный дисплей, качество связи
Аккумулятор, скорость красивая. Дизайн, звук, функционал, масса разных дней хватает. Красив, несколько назад, процессор отзывчивый сенсор. Красивый экран, цветопередача. Дизайн, батарея, не тормозят, практичный.

Как видно, результаты получаются довольно неоднозначные. С одной стороны– безусловными плюсами такого подхода является полная автоматизация процесса генерации текста, высокая степень адаптивности системы, низкие затраты на ее настройку и внедрение. С другой – очевидны некоторые проблемы появления «смыслового мусора», на подобии «скорость красивая, отзывчивый экран, рабочий день». Загвоздка тут состоит в том, что несмотря на кажущееся наличие интеллектуальной обработки, система не понимает смыла того

что генерирует, отталкиваясь лишь от заранее заданных шаблонов. Преодолеть это ограничение можно двумя путями – это усложнение сети, что неизменно приведет нас к потерям в адаптивности и переносимости, или же внедрение дополнительных технологий представления текстовых данных – создание «совмещенных» систем.

Выводы и перспективы дальнейших исследований. Рассмотрев основные подходы к созданию систем АОТ в задаче генерации текста можно сказать, что наиболее перспективным направлением является именно искусственные интеллектуальные алгоритмы (нейронные сети и подобные им). Шаблонные семантические сети и генеративные грамматики применимы к определенному кругу задач (парсинг языка программирования, квазиреферирование), но для интеллектуальной генерации текста они слабо применимы. Однако на нейронные сети накладывается ограничение при попытке осознать и оценить обрабатываемую информацию, связанное с их ориентированностью на грамматическую структуру предложения или слова. Выходом из этой ситуации может стать соединение возможностей нейронных сетей (или иных алгоритмов, способных к автоматическому обучению и развитию) с инструментами альтернативной теории моделирования языка - моделью Мельчука – «Теория Смысл ↔ Текст» (ТСТ)[6]. ТСТ отделяет семантику от синтаксиса, настаивая при этом на ее научном описании. По разным причинам, (от политических до экономических) модель не получила широкого распространения на западе, оставшись в «информационном вакууме» постсоветского пространства. Тем не менее, многие ученые характеризуют ее как опередившую свое время. Все чаще звучат утверждения, что грамматикам составляющих только предстоит эволюционировать в некоторые парадигмы ТСТ[3]. Чтобы разобраться в проблематике ТСТ, нам необходимо оценить ее структуру, разработанную для автоматизации переводов с точки зрения задачи генерации текста, что даст нам основные плюсы и минусы модели в прикладной реализации.

Первым, безусловным полюсом, является высокая степень ориентированности на синтез текста. И.А. Большаков отмечает, что «...насколько нам известно, синтез текста по произвольно заданной семантической сети серьезно продумывался именно в рамках модели ТСТ...»[3]. Задача синтеза текста, в отличии от задачи его анализа, требует описание глубинных семантических отношений (сферхфразо-

вые отношения), которые, как отмечалось выше, в грамматиках Хомского просто не учитываются. Лексико-семантические правила расширенных генеративных теорий направлены на решения грамматических неоднозначностей, и не могут решать задачу синтеза. Особенно чувствуется несовершенство грамматики составляющих при попытке синтезировать язык со свободным порядком слов, где грамматика неоднозначна, а в качестве опоры используются смысловые единицы. Это приводит нас ко второму плюсу модели ТСТ – независимость от порядка следования слов в предложении. Действительно, зная семантику предложения, нам не обязательно опускаться до синтаксического уровня, что дает нам возможность обойти свободный порядок слов: «основное преимущество грамматик зависимостей усматривается в том, что именно связи между (полнозначными) словами сохраняются на семантическом уровне, а для грамматик составляющих их обычно приходится выявлять на семантическом уровне отдельным механизмом»[3].Такой подход дает возможность реализовать формальную модель языка для всей индоевропейской группы, а не только для английского, где порядок следования отыгрывает ключевую роль. Эти факты говорят о преимуществе грамматик зависимостей перед грамматиками составляющих в задачах синтеза текста.

Главным минусом ТСТ является толково-комбинаторный словарь, базирующийся на механизме сем. Это подразумевает, что для каждой семы будет описан полный набор ее свойств (в специальной статье) – от лексических и морфологических до таких высоких семантических уровней как связанные идиомы. Необходимость составления полного словаря (в оригинальном виде) возможна только вручную – в случае анализа неизвестного слова мы не сможем выбрать сценарий поведения системы,основанной на ТСТ. Это связанно с тем, что структура статьи является слишком сложной: при возможности автоматического выбора морфологических признаков с некоторой, допустимой ошибкой, остальные пункты, как например идиомы, заполнить автоматически без специальных семантических знаний невозможно. И даже если говорить о выявлении идиом как коллокативных сочетаний или N-грамм, то для описания всего ЕЯ придется использовать настолько большой корпус, механизмов навигации по которому еще просто не разработано. Эта проблемя являетсяследствием из свойства изменчивости ЕЯ – даже если подобный корпус со-

ставлен, то нам нужна система его непрерывного пополнения. Если же действовать на неразмеченном информационном пространстве, как например интернет, мы столкнёмся с ростом знаний в геометрической прогрессии, причем по большей части – с дублированной семантикой данных (из-за свойства вариативности такого пространства).Доверить же составление статей оператору-человеку является довольно сложной массовой задачей. «Технология составления толково-комбинаторых словарей осталась не разработанной. Их составление оказывается до сих пор под силу только тем, кто осваивал модель много лет, а по существу создавал и совершенствовал ее. Вероятно, именно отсутствие ясной и массовой технологии разработки словарей явилось одной из основных исторических причин отставания модели от западных «конкурентов»...»[3]. Однако, даже несмотря на вышеуказанные минусы, ТСТ является более удобной для разработки системы синтеза текста, чем грамматически-ориентированные модели. Используя механизмы искусственного компьютерного интеллекта, имеющие возможности оперировать не грамматическими единицами, а семантическими элементами,возможно добиться высоких результатов, как например, в системах Н.Н. Леонтьевой [7], ориентированных на компьютерный перевод. С точки зрения генерации текста, пока ТСТ не рассматривалась, хотя, как видно из вышесказанного, ее применение в этой области более чем оправдано.

ЛИТЕРАТУРА

1. Швецов А.Н., Сорокин С.И., Мамадкулов Ю.О. -Система синтеза учебных тестов на основе формальных грамматик //НИИ «Центрпрограммсистем» - журнал«Программные продукты и системы», №2(102), 2013, с 181-185.
2. Мозговой Максим Владимирович //Машинный семантический анализ русского языка и его применения//Санкт-Петербург – 2006г.
3. I.A. Bolshakov, A.F. Gelbukh. // The Meaning ↔ Text Model: Thirty Years After. J. // International Forum on Information and Documentation, FID 519, ISSN 0304-9701, N 1, 2000.
4. Гурин Н.И., Жук Я.А. - Семантическая сеть электронного учебника для диалога с виртуальным преподавателем // Материалы международной научно-технической интернет конференции "Информационные технологии в образовании, науке и производстве"// Белорусский государственный технологический университет, Минск, 2015 г.
5. Тарасов Д.С. - Генерацияестественного языка, парафраз и автоматическое обобщение отзывов пользователей с помощью рекуррентных нейронных сетей // «Компьютерная лингвистика и интеллектуальные технологии», №14(том 1), 2015, с 607-614//Материалымеждународной конференции«Диалог», 2015 г.
6. Мельчук И.А. Опыт теории лингвистических моделей «Смысл ↔ Текст». – 2-е издание, доп. – М. 1999.
7. Н.Н. Леонтьева // Автоматическое понимание текстов: Системы, модели, ресурсы. // Москва – 2006 г.