ПОИСК И ИЗВЛЕЧЕНИЕ ЗНАНИЙ: ПОРОЖДЕНИЕ НОВЫХ ЗНАНИЙ НА ОСНОВЕ АНАЛИЗА ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА*

Д.Е. Пальчунов

Для решения вопроса о том, может ли человеческое мышление быть реализовано на компьютере, мы предложили рассматривать не отдельный компьютер, а компьютерную сеть, например Интернет. Такой подход более соответствуют реальному процессу порождения знаний в человеческом обществе, где новые знания – продукт работы не одного полностью изолированного ученого, а научных коллективов и школ. При этом принципиальным является использование ранее накопленного знания, представленного в огромном количестве текстов, написанных на естественном языке.

Ключевые слова: мышление, компьютер, язык, знание

1. Введение

Статья посвящена исследованию методов поиска и извлечения информации из текстов естественного языка и порождения новых знаний, в этих текстах явно не содержащихся. Мы продолжаем разрабатывать подход к решению вопроса, сформулированного в предыдущей статье [6]: в каком объеме творческая, исследовательская деятельность человека может быть формализована, и насколько она может быть реализована на вычислительной машине?

В данной статье мы расширяем постановку этого вопроса. Во-первых, вместо одного вычислительного устройства мы рассматриваем сеть компьютеров – такую, например, как сеть Интернет. Во-вторых, мы предполагаем, что вычислительные устройства при сво-

* Работа выполнена при поддержке гранта Междисциплинарного интеграционного проекта фундаментальных исследований СО РАН № 47 «Логика, математический анализ выразительных возможностей языка в представлении знания: соотношение синтаксиса, семантики, прагматики в формализации научных теорий», а также при поддержке Федерального агентства по науке и инновациям РФ, государственный контракт № П-1008.

Публикуется в авторской редакции.

ей работе могут использовать накопленные людьми знания, представленные в текстах естественного языка (опять же, как это имеет место и в сети Интернет).

Такая постановка на наш взгляд является более адекватной исходному вопросу – может ли творческая и исследовательская деятельность человека быть реализована на компьютере? Действительно, для того, чтобы научиться мышлению и творчеству недостаточно родиться человеком – ребенок должен находиться в окружении других людей и изучать совокупный опыт, накопленный человечеством и представленный в первую очередь в текстах естественного языка. Так называемые «маугли» – дети, которые воспитывались вне человеческого общения, имеют, как известно, интеллектуальный уровень, не принципиально отличающийся от интеллектуального уровня животных. Отдельным вопросом является возможность полной формализации и моделирования на компьютере мышления животных. Рассмотрение этого вопроса выходит за рамки данной статьи; отметим только, что в настоящий момент времени имеются значительные продвижения в этом направлении (см., например, [16, 17]).

Таким образом, исходный вопрос о возможности реализации человеческого мышления на компьютере мы расширим и несколько ослабим. В измененной постановке он будет звучать так: возможно ли автоматически, без участия человека, при помощи анализа текстов естественного языка порождать принципиально новые знания, в этих текстах явно не содержащиеся?

Заметим, что на первый взгляд данная постановка вопроса кажется значительно более слабой, чем амбициозная исходная проблема «Может ли компьютер мыслить? (Как человек!)». Однако если мы проанализируем реальную творческую деятельность людей, то мы легко увидим, что именно так она и проистекает: сначала человек изучает достаточно большое количество различных текстов, написанных на естественном языке, а только затем, после их осмысления, порождает новые знания.

Естественно возникает следующий вопрос – насколько новыми являются эти полученные знания. Но, во-первых, и у разных исследователей степень новизны, революционности полученных знаний является разной – этим и отличаются великие ученые от посредственных. И, во-вторых, что является наиболее важным, – мы уходим от принципиальной постановки вопроса: «Может компьютер мыслить или не может?» и приходим к качественной (или количествен-

ной) – «Насколько новые знания может порождать компьютер?». Еще раз отметим, что в такой постановке сравнение результатов интеллектуальной деятельности человека и компьютера уже принципиально не отличается от сравнения серьезных научных результатов и халтурных.

Таким образом, данная работа посвящена решению вопроса о возможности автоматического порождения новых знаний на основе анализа текстов естественного языка – знаний, в этих текстах явно не содержащиеся. Мы предложим обоснование положительного ответа на данные вопрос и остановимся на исследовании методов поиска и извлечения знаний из текстов естественного языка, а также методов порождения новых знаний. В первую очередь нас будет интересовать работа с текстами естественного языка, представленными в сети Интернет.

2. Проблема новизны и достоверности полученного знания

Первый вопрос, который мы бы хотели обсудить – а является ли принципиально возможным порождение действительно нового знания, основываясь на уже известных текстах? Или, сформулируем по-другому, – возможно ли получать принципиально новое знание без обращения к реальному миру? Вторая сторона этого вопроса – а является ли полученное новое знание достоверным? То есть, является ли оно действительно знанием, а не просто некоторым гипотетическим предположением? Вывод, который мы предлагаем – принципиально новое знание принципиально не может быть гарантированно достоверным. Несмотря на парадоксальность этого утверждения, мы попытаемся его обосновать.

Сначала подробно остановимся на том, что мы вкладываем в слово «новый», когда говорим о новом знании. Первый, более слабый смысл – новое это то, что в явном виде еще не было известно человеку (человечеству). Второй, более сильный смысл: новое – это то, что не следует из уже известного. Рассмотрим, чем второй смысл отличается от первого.

Обозначим через $\boldsymbol{K}$ совокупность всех уже известных фактов, а через $\boldsymbol{N}$ – новое знание. Мы утверждаем, что $\boldsymbol{N}$ не является новым знанием во втором смысле тогда и только тогда, когда импликация $(\boldsymbol{K} \to \boldsymbol{N})$ является аналитически истинной по Р.Карнапу [4, 10]. Если

мы рассмотрим еще и онтологию ***O***, содержащую полный набор определений всех используемых в ***K*** и ***N*** понятий, то ***N*** не является новым знанием во втором смысле тогда и только тогда, когда импликация (***K & O → N***) является логически истинной, то есть логической тавтологией. Это соответствует точке зрения Л. Витгенштейна, по которой любое математическое знание является логической тавтологией [1, 15].

Предположим теперь, что импликация (***K → N***) не является аналитически истинной. Это означает, что ее истинность зависит от состояния реального мира. Но в таком случае, можем ли мы утверждать, что новое «знание» ***N*** является гарантированно достоверным? Очевидно, не можем! Утверждение ***N***, если оно осмысленно, является верифицируемым или фальсифицируемым, но в любом случае необходима проверка этого утверждения на практике, необходима проверка его соответствия реальному миру. (В данном случае мы, вслед за Р. Карнапом и К. Поппером [8], считаем научное утверждение осмысленным, только если оно является верифицируемым или фальсифицируемым.) Таким образом, по существу ***N*** является не полностью достоверным знанием, а лишь эмпирической гипотезой.

Таким образом, мы видим, что если утверждение ***N*** является действительно новым, то есть если импликация (***K → N***) не является аналитически истинной, то ***N*** не является гарантированно достоверным знанием, а является лишь эмпирической гипотезой. Конечно, даже в том случае если импликация (***K → N***) и является аналитически истинной, утверждение ***N*** может также не быть достоверным знанием – это зависит от того, насколько достоверным является исходное знание ***K***.

Подводя итог, отметим следующее. Обрабатывая тексты естественного языка (в частности, представленные в сети Интернет) мы можем рассчитывать на автоматическое получение новой информации следующих двух видов:

1. Знания, в явном виде не содержащиеся в рассматриваемых текстах, но вытекающие из них аналитически (или логически). Такие знания являются достоверными «по модулю» достоверности знаний, представленных в рассматриваемых текстах.

2. Эмпирические гипотезы, порожденные на основе рассматриваемых текстов естественного языка. Достоверность эмпирических гипотез должна будет в дальнейшем проверяться.

Знания первого вида являются более достоверными, второго вида – более информативными, и, соответственно, возможно, более ценными.

С определенной точки зрения только знания второго вида могут рассматриваться как действительно принципиально новые (особенно, если речь идет не о математических теоремах). Таким образом, мы получили следующий достаточно парадоксальный вывод – принципиально новое знание не может быть гарантированно достоверным.

3. Извлечение неявных знаний

Рассмотрим теперь вопрос о возможности извлечения неявных знаний из текстов естественного языка, представленных в сети Интернет. Речь идет о том, что разные документы, представленные в Интернете, могут содержать различные знания. Комбинируя эти знания, мы можем получить новое знание, в явном виде не содержащееся ни в одном документе.

Рассмотрим достаточно простой модельный пример. Допустим, мы хотим узнать дату рождения жены второго президента РФ. В качестве поисковой машины возьмем наиболее развитую на настоящий момент времени – Google, запрос введем на английском языке – на нем Google наиболее хорошо работает: «What is the date of birth of the wife of the second president of Russia». Получаем выдачу поисковой системы, которую трудно назвать релевантной данному вопросу:

Теперь проделаем следующее: разобьем вопрос на части и будем вводить запросы в Google по частям. По существу мы будем производить декомпозицию сложного поискового запроса на простые запросы. С помощью первого поискового запроса мы попробуем означить первый референтный индекс – установить имя второго президента России. Введем в Google запрос «second president of Russia» и получим следующую выдачу:

Теперь нам даже не нужно загружать какой-либо документ – уже во втором абзаце выдачи мы видим ответ на наш вопрос: «The **second President of Russia** was Vladimir Putin … ».

Далее постараемся означить второй референтный индекс – установить имя жены второго президента России, то есть имя жены Владимира Путина.

Для этого введем в Google запрос «wife of Vladimir Putin» и получим следующую выдачу:

Снова нам не нужно загружать какой-либо документ – в третьем абзаце выдачи мы видим ответ на наш второй вопрос: «… Ludmila Putina, **wife of Vladimir Putin**, President of the Russian Federation …».

Теперь мы можем попробовать получить ответ на наш исходный вопрос – узнать дату рождения жены второго президента РФ, то есть дату рождения Людмилы Путиной.

Введем в Google запрос «Ludmila Putin was born» и получим выдачу, в которой уже самая первая строчка содержит ответ на наш вопрос:

А именно: «**Ludmila Putin**(a) was **born** on January 6, 1958, in Kaliningrad». В дополнение мы узнали не только дату рождения жены второго президента РФ, но и место – Калининград.

Приведенный пример, конечно, является достаточно простым, но, несмотря на это, он показывает несколько моментов. Во-первых, мы видим, что информация, в явном виде не содержащаяся в одном документе, по частям может содержаться в нескольких документах, в данном случае, в трех.

Во-вторых, данный пример показывает возможный алгоритм извлечения неявной информации. Сначала необходимо разложить сложный вопрос в последовательность простых. Произвольный вопрос на естественном языке с помощью морфологического разбора превращается в комбинацию простых вопросов. Для дальнейшей работы алгоритма необходим набор шаблонов возможных ответов: для каждого простого вопроса – свой шаблон ответа (или свой набор шаблонов возможных ответов). После подбора шаблонов ответов происходит поиск информации в Интернете – по шаблонам возможных ответов порождаются поисковые запросы. Далее происходит обработка выдачи поисковой системы и построение ответов на простые вопросы. После этого – сборка ответов на простые вопросы для получения ответа на исходный сложный вопрос. Таким образом, в результате работы алгоритма порождается ответ на исходный вопрос, сформулированный на естественном языке.

4. Вопросно-ответные системы

В настоящее время внимание специалистов по информационному поиску привлекают две наиболее интересные проблемы:

1. Создание возможности формулировать поисковый запрос на естественном языке.
2. Удовлетворение поисковой потребности вместо выполнения поиска документов, обладающих лишь определенными синтаксическими свойствами.

Когда мы рассматриваем проблему извлечения знаний и порождения новых знаний, очень важным вопросом является следующий – знания какого вида нас интересуют. Этот вопрос можно сформулировать более точно и конкретно, когда рассматриваются вопросно-ответные системы – программные системы, предназна-

ченные для поиска ответов на вопросы, сформулированные на естественном языке.

Разработка технологий обработки вопросов на естественном языке и поиска ответов на такие вопросы вызывает большой интерес исследователей, работающих в области формализации естественного языка и инженерии знаний. Растет количество работ по этой теме, существует ряд уже разработанных программных систем. Каждая из таких систем предлагает свою классификацию видов вопросов и, соответственно, видов знания, которое можно получить с их помощью. Кратко остановимся на некоторых из них.

Среди вопросно-ответных систем можно выделить следующие:

– START
– CLEF
– Wolfram Alpha
– LTAG

START (аббревиатура от «SynTactic Analysis using Reversible Transformations») – вопросно-ответная система на естественном языке (Natural Language Question Answering System) [18]. В этой системе вопросы разбиты на четыре группы:

– География

Например:

- У какой южноамериканской страны население наибольшее?
- Какой город является наибольшим городом во Флориде?
- Перечислите государства, граничащие с Колорадо.
- Какие города находятся в пределах 250 миль от столицы Италии?
- Сколько людей живет в Израиле?

– Наука и справочная информация

Например:

- Из чего сделана атмосфера Юпитера?
- Кто первым открыл радиоуглеродное датирование?
- Как далеко Нептун находится от солнца?

- Почему небо синее?
- У какой планеты площадь поверхности наименьшая?
- Переведите 100 долларов в Евро.
- Покажите карту метро Москвы.
- На скольких языках говорят в Афганистане?
- Что изобрел Маркони?

– Искусство и эстрада

Например:

- Кто поставил фильм «Унесенные ветром»?
- Покажите некоторые картины Клода Моне.
- Когда родился Бетховен?
- Чем известен Александр Пушкин?
- В каких кинофильмах снимался Дастин Хоффман?

– История и культура

Например:

- Какие страны говорят на испанском языке?
- Кто был президентом в 1881?
- Покажите некоторые стихи Роберта Фроста.
- Кто был пятым президентом Соединенных Штатов?
- На каких языках говорят в самой густонаселенной стране Африки?

Не обсуждая полноту данной классификации вопросов, отметим, что данные четыре группы вопросов, по существу, пересекаются. Например, вторая и четвертая: «На скольких языках говорят в Афганистане?» и «На каких языках говорят в самой густонаселенной стране в Африке?»; третья и четвертая: «Покажите некоторые картины Клода Моне», «Чем известен Александр Пушкин?» и «Покажите некоторые стихи Роберта Фроста» и т.п.

Следующая вопросно-ответная система – **CLEF** – аббревиатура от «Cross-Language Evaluation Forum» [19, 20]. Рассматриваются следующие три категории вопросов:

а) фактографические вопросы;

б) определения;
в) закрытые списки.

а) Фактографические вопросы – основанные на факте вопросы: имя человека, местоположение, сущность чего-то, день, в который что-то произошло, и т.д.

Рассматриваются 8 типов ответов для таких вопросов:

− ЧЕЛОВЕК, например:
Вопрос: Кого называли "Железным канцлером"?
Ответ: Отто фон Бисмарка.

− ВРЕМЯ, например:
Вопрос: В каком году был убит Мартин Лютер Кинг?
Ответ: В 1968.

− МЕСТОПОЛОЖЕНИЕ, например:
Вопрос: В каком городе родился Вольфганг Амадей Моцарт?
Ответ: В Зальцбурге.

− ОРГАНИЗАЦИЯ, например:
Вопрос: В какой партии состоит Тони Блэр?
Ответ: В лейбористской партии.

− МЕРА, например:
Вопрос: Какова высота Канченджанги?
Ответ: 8598 м.

− СЧЕТ, например:
Вопрос: Сколько людей умерло во время террора Пол Пота?
Ответ: 1 миллион.

− ОБЪЕКТ, например:
Вопрос: Из чего состоит магма?
Ответ: Из расплавленной породы.

− ДРУГИЕ, то есть все, что не вписывается в категории представленные выше.

Вопрос: Какое соглашение было подписано в 1979?
Ответ: Мирный договор между Египтом и Израилем.

б) Вопросы об определениях – такие вопросы, как "Каков / кем / чем является X?"; они разделены на следующие подтипы:

− ЧЕЛОВЕК, то есть вопросы, спрашивающие о роли / работе / важной информации о
ком-то, например:
Вопрос: Кто такой Роберт Альтман?
Ответ: Кинопроизводитель.

− ОРГАНИЗАЦИЯ, то есть вопросы, запрашивающие назначение / полное название / важную информацию об организации, например:
Вопрос: Что такое Кнессет?
Ответ: Парламент Израиля.

− ОБЪЕКТ, то есть вопросы, запрашивающие описание / функции объектов, например:
Вопрос: Что такое Атлантис?
Ответ: Шаттл.

− ДРУГИЕ, то есть вопросы об описании естественных явлений, технологий, юридических процедур и т.д., например:
Вопрос: Что такое Евровидение?
Ответ: Конкурс песни.

в) Закрытые вопросы о списках: то есть вопросы, которые требуют в одном единственном ответе указать требуемое число позиций, например:

Вопрос: Назовите имена всех аэропортов в Лондоне, Англия.
Ответ: Gatwick, Stansted, Heathrow, Luton, City.

Вопрос: Назовите имена трех последних американских президентов.
Ответ: Джордж Х.У. Буш, Билл Клинтон, Джордж У. Буш.

Все типы вопросов могут содержать временное ограничение, то есть временную спецификацию – это предоставляет важную информацию для поиска правильного ответа.

Примеры:

Вопрос: Кто был Канцлером Германии с 1974 до 1982 год?
Ответ: Гельмут Шмидт.

Вопрос: Какая книга была издана Джорджем Оруэллом в 1945 году?
Ответ: Скотный двор.

Вопрос: Какую организацию возглавил Шимон Перес после смерти Ицхака Рабина?
Ответ: Центральный комитет Лейбористской партии.

Таким образом, системой **CLEF** представлена достаточно детальная классификация видов информации, интересующих пользователя вопросно-ответной системы.

Программная система **Wolfram Alpha** [21] является чем-то средним между вопросно-ответной системой и системой поддержки математических вычислений типа MATHLAB.

Вопросы разделены на следующие рубрики: Математика, Статистика и Анализ Данных, Физика, Химия, Материалы, Инженерные науки, Астрономия, Науки о Земле, Науки о жизни, Технологии, Транспорт, Вычислительные Науки, Сети и Компьютерные системы, Единицы и Меры, Деньги и Финансы, Даты и Время, Местность и География, Социально-экономические Данные, Погода, Здоровье и Медицина, Пищевые продукты и питание, Языки и Лингвистика, Культура и СМИ, Люди и История, Образование, Организации, Спортивные состязания и Игры, Музыка, Цвета.

Каждая рубрика имеет свои подрубрики. Например:

Математика

Элементарная математика · Числа · Построение графиков · Алгебра · Матрицы · Математический анализ · Геометрия · Тригоно-

метрия · Дискретная математика · Теория чисел · Прикладная математика · Логика · Функции ·...

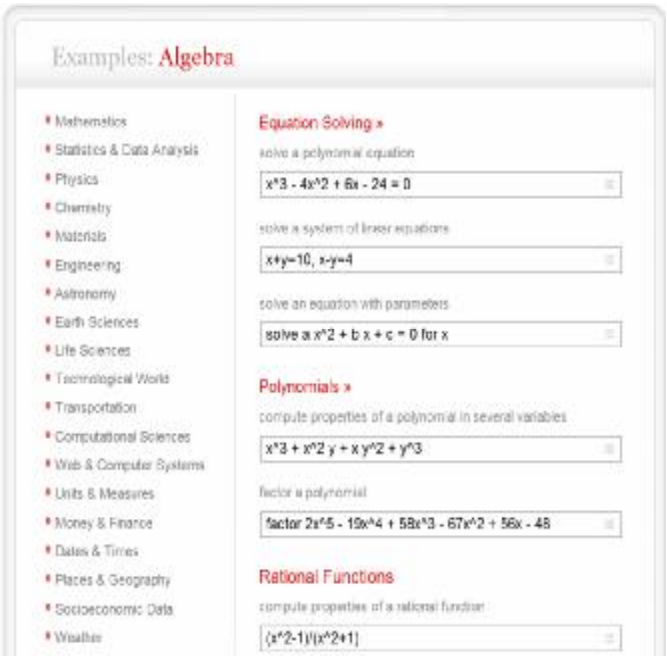
Статистика и Анализ Данных

Анализ данных · Регрессия · Статистические распределения · Вероятность ·...

Физика

Механика · Электричество и магнетизм · Оптика · Теория относительности · Ядерная физика · Квантовая физика · Физика элементарных частиц · Статистическая физика · Астрофизика · Физические константы ·...

Данная система позволяет проводить вычисления, находить решения уравнений и т.п. Примеры таких вычислений приведены на следующем рисунке:



Например, система **Wolfram Alpha** дает возможность решать уравнения. Если пользователь введет уравнение $x^3 - 4x^2 + 6x - 24 = 0$, система выдаст: решение в действительных числах $x = 4$ и два решения в комплексных числах $x = -i\sqrt{6}$ и $x = i\sqrt{6}$.

Как мы уже отмечали выше, программная система **Wolfram Alpha** больше похожа на систему поддержки математических вычислений MATHLAB, чем на вопросно-ответную систему.

5. Сильные и слабые стороны существующих вопросно-ответных систем

Несмотря на свои несомненные достоинства, существующие вопросно-ответные системы имеют ряд принципиальных недостатков. Часть из таких систем осуществляет поиск информации, только содержащейся в своей базе данных. Это приводит к отсутствию новой, свежей информации и является совершенно неприемлемым для целого ряда областей, где знания постоянно обновляются – от технологий информационной безопасности до политологии и Интернет-маркетинга.

У вопросно-ответных систем, которые пытаются осуществлять информационный поиск по всему Интернету, поиск, как правило, идет только по определенным видам шаблонов вопросов. Это, во-первых, крайне неудобно для пользователя – по существу, исчезает наиболее важное преимущество, связанное с возможностью формулирования поискового запроса на естественном языке. Во-вторых, происходит потеря универсальности запросов (по всем возможным предметным областям и видам информационных потребностей), свойственной таким поисковым системам как Google и Яндекс.

Заметим также, что описанные вопросно-ответные системы – START, CLEF и Wolfram Alpha не решают наиболее важной для нашего рассмотрения проблемы порождения нового знания.

Тем не менее, такие системы, во-первых, дают возможность реализации одного из этапов алгоритма извлечения неявной информации, описанного в конце предыдущего параграфа. А именно – построения ответов на простые вопросы. Важность полной реализации этого этапа трудно переоценить.

Во-вторых, при разработке данных систем предложена довольно подробная и детальная классификация вопросов – т.е., по суще-

ству, видов «атомарных» знаний, которые можно комбинировать и порождать из них более сложные, в том числе и новые знания – ответы на сложные вопросы.

В разных вопросно-ответных системах классификация вопросов разная и, в определенном смысле, «перпендикулярная», порождающая двумерное или даже многомерное пространство вопросов или «атомарных» знаний. У системы CLEF классификация вопросов построена по типу получаемой информации, а у систем START и Wolfram Alpha – по предметным областям, к которым относится данная информация. Еще одно измерение в классификации вопросов, о котором мы говорили, описывая систему CLEF – временное ограничение или временная спецификация.

Таким образом, разработка описанных вопросно-ответных систем является очень важным первым шагом к решению проблемы автоматического порождения новых знаний на основе обработки текстов естественного языка, представленных в сети Интернет. Для реализации дальнейших шагов необходима разработка как методов декомпозиции сложных вопросов в последовательность простых, так и методов «сборки» ответов на сложные вопросы на основе уже полученных ответов на простые вопросы.

6. Порождение новых знаний

Рассмотрим теперь следующий вопрос: каким образом можно получать (порождать) новые знания, в явном виде ни в одном документе не содержащиеся? Пример решения этого вопроса – выяснение даты рождения жены второго президента РФ, приведенное в предыдущем параграфе. Прежде всего, несколько уточним данный вопрос – он будет звучать следующим образом: «Каким образом можно получать (порождать) новые знания, в явном виде не содержащиеся ни в одном документе из данного набора документов?». Разница состоит в том, что теперь нам не нужно нести ответственность за наличие ситуации, когда ответ на указанный вопрос все-таки содержится в явном виде в некотором (неизвестном нам) документе, представленном в сети Интернет.

Один из путей решения этой проблемы – обработка вопросов, написанных на естественном языке. При таком подходе мы одновременно решаем сразу две задачи. Во-первых, мы порождаем но-

вую информацию, которой в явном виде не было ранее. Во-вторых, мы предоставляем пользователю понятный и удобный интерфейс, предназначенный для точной формулировки его информационной потребности.

При данном подходе мы разбиваем задачу получения новой информации на несколько подзадач, которые должны решаться последовательно. Первое, это определение контекста вопроса; здесь мы идем от синтаксиса вопроса к прагматике – способам употребления, использования требуемой информации. Речь идет, во-первых, о том, в каком типе документов мы ищем информацию: это могут быть научные статьи, материалы Интернет-энциклопедий типа Википедии, тексты форумов и т.п. Во-вторых, по сложному вопросу значительно легче, чем по простому, понять, к какой предметной области и к какому периоду времени относится требуемая информация.

Далее необходимо разложение сложного вопроса в набор – последовательность простых, «атомарных» вопросов. Такое преобразование исходного вопроса необходимо по двум причинам. Во-первых, чем информация более просто, кратко сформулирована, тем больше вероятность того, что она в явном виде содержится в некотором Интернет-документе. Во-вторых, для набора всевозможных простых вопросов можно построить набор шаблонов возможных ответов. Например, для поиска ответа на вопрос, кто являлся вторым президентом России, можно использовать следующие шаблоны возможных ответов: «***X*** – второй президент России», «***X*** был вторым президентом России», «***X***, второй президент России», «второй президент России ***X***» и т.д.

Для разложения сложного вопроса в последовательность простых необходим его морфологический разбор. В частности, полезна нормализация слов вопроса – приведение их к каноническому виду: например, существительных к именительному падежу и единственному числу. Морфологический разбор сложного вопроса и нормализация слов сейчас не является очень сложной проблемой – в настоящий момент времени имеется достаточно большое количество программных систем, как предназначенных для морфологического разбора предложений естественного языка, так и осуществляющих нормализацию слов. Многие программные системы, предназначенные для нормализации слов русского языка, основаны на словаре А.А. Зализняка [2].

После этого происходит обработка простых вопросов, сформулированных на естественном языке. Для этого, с помощью набора шаблонов возможных ответов производится поиск информации в Интернете. Для каждого шаблона возможного ответа на простой вопрос, порождается один или несколько поисковых запросов. При этом, что очень важно, по существу используется определенный ранее контекст исходного сложного вопроса. В частности, знание о предметной области, в которой ищется требуемая информация, позволяет значительно более точно специфицировать поисковый запрос.

Кроме того, очень важным является то, что ответы на простые вопросы ищутся в разных документах, представленных в сети Интернет. Таким образом, мы получаем возможность получать знание, не содержащееся в явном виде ни в одном из документов.

В результате обработки выдачи, полученной в ответ на поисковые запросы, порождаются ответы на имеющийся набор простых вопросов. После этого происходит сборка полученных ответов на простые вопросы для порождения ответа на исходный сложный вопрос. Этот процесс является достаточно сложным. В частности, встает проблема неоднозначности: при помощи шаблонов ответов и порожденных на их основе поисковых запросов могут быть получены разные ответы на исходные вопросы. Как с их помощью порождать релевантный ответ на исходный сложный вопрос – тема отдельного исследования. Здесь мы только отметим, что для начала можно ограничиться более простой задачей – выдавать пользователю несколько ответов на исходный вопрос вместе со ссылками на источники информации. При этом окончательное решение, какой из ответов является релевантным, остается за человеком.

Предложенный подход к обработке вопросов, сформулированных на естественном языке, дает принципиально новую возможность для реализации поисковой потребности пользователя – вместо поиска документов, обладающих определенными синтаксическими свойствами, он получает явный ответ на явно сформулированный вопрос.

Однако даже сложные вопросы, сформулированные на естественном языке, не исчерпывают все возможные поисковые потребности пользователей. Для более общей постановки вопроса формализации поисковых потребностей различных пользователей мы можем применить теорию речевых действий [3, 9, 11–14].

Наиболее простое описание поисковой потребности – это тип поисковой задачи + спецификация предметной области. Например: «[купить книгу] по [логике описаний]» или «[найти статьи] по [онтологиям в медицине]». Здесь первая часть речевого действия описывает вид действия, которое нужно совершить, а вторая часть описывает содержание действия. Для реализации такой поисковой потребности сначала необходимо информационную потребность пользователя превратить в набор запросов к поисковым системам, затем провести анализ выдачи поисковой системы (поисковых систем) и проделать вторичную обработку документов. В результате будет получена итоговая выдача пользователю. При выполнении данной процедуры наиболее остро встает проблема релевантности отработки поисковых запросов [5, 7]. Для ее достижения необходимо для каждого используемого поискового запроса строить набор эвристик, которые будут обеспечивать его релевантность.

7. Заключение

В данной работе для решения вопроса – может ли человеческое мышление быть реализовано на компьютере – мы предложили рассматривать не отдельный компьютер, а компьютерную сеть, например, Интернет. Такая постановка вопроса более соответствуют реальному процессу порождения знаний в человеческом обществе, где новые знания являются продуктом работы не одного полностью изолированного ученого, а научных коллективов и школ. При этом принципиальным является использование ранее накопленного знания, представленного в огромном количестве текстов, написанных на естественном языке.

Таким образом, одной из центральных становится проблема автоматического порождения новых знаний исходя из уже имеющихся, представленных в текстах естественного языка. Вопрос о возможности моделирования человеческого мышления на компьютере мы рассмотрели в уточненной и несколько ослабленной постановке – возможно ли автоматически на основе анализа текстов естественного языка порождать новые знания?

Одним из путей порождения новых знаний при помощи текстов естественного языка является сравнение и интеграция знаний, содержащихся в разных текстах. Разработку вопросно-ответных сис-

тем для поиска информации в сети Интернет можно рассматривать как первый шаг на этом пути. Следующими шагами являются разработка методов разложения сложных вопросов в набор простых и методов порождения ответов на сложные вопросы на основе ответов на простые вопросы.

Также необходима разработка методов оценки достоверности знаний, полученных автоматически, и методов порождения эмпирических гипотез. Математической основой таких методов могут стать исследования по нечетким алгебраическим системам, в частности, разработка алгоритмов оценки степени достоверности произвольных предложений исходя из известной нечеткой оценки некоторого заранее заданного множества предложений.

Примечания

1. *Витгенштейн Л.* Философские работы. Ч. 1, 2. – М.: Гнозис, 1994.

2. *Зализняк А.А.* Грамматический словарь русского языка. Серия: Фундаментальные словари. АСТ-Пресс Книга, 2009, 720 стр.

3. *Пальчунов Д.Е.* Алгебраическое описание смысла высказываний естественного языка. Модели когнитивных процессов. – Новосибирск, 1997 – Вып. 158: Вычислительные системы. – С. 127–148.

4. *Пальчунов Д.Е.* Моделирование мышления и формализация рефлексии I: Теоретико-модельная формализация онтологии и рефлексии. Философия науки, N 4 (31), 2006. – С. 86–114.

5. *Пальчунов Д.Е., Сидорова Е.С.* Виртуальный каталог. Труды Всероссийской конференции «Знания-Онтологии-Теории», Новосибирск, 2007. – С. 166–175.

6. *Пальчунов Д.Е.* Моделирование мышления и формализация рефлексии. Ч.2. Онтологии и формализации понятий // Философия науки – № 2 (37) – 2008. – С. 62–99.

7. *Пальчунов Д.Е.* Решение задачи поиска информации на основе онтологий. Бизнес-информатика, № 1, 2008, стр. 3–13.

8. *Поппер, К.* Логика и рост научного знания. М.: Прогресс, 1983.

9. *Austin, J. L.* How to do Things with Words. Oxford: Oxford University Press. 1962.

10. *Carnap, R.* (1956): Meaning and Necessity. A Study in Semantics and Modal Logic. Chicago.

11. *Pal'chunov, D. E.* Algebraische Beschreibung der Bedeutung von ДuЯerungen der natьrlichen Sprache. In: Zelger, Josef/Maier, Martin (Hrsg.): GABEK. Verarbeitung und Darstellung von Wissen. Innsbruck-Wien: STUDIENVerlag, 1999, 310–326.

12. *Searle, John R.* Speech Acts. An Essay in the Philosophy of Language. Cambridge: Cambridge University Press. 1969.

13. *Searle, John R., Vanderveken, Daniel.* Foundations of Illocutionary Logic. Cambridge University Press, 1985.

14. *Vanderveken, Daniel*. Meaning and Speech Acts vol. I Principle of Language Use, vol. II Formal Semantics of Success and Satisfaction, Cambridge University Press, Cambridge, 1991.

15. *Wittgenstein L.* Tractatus Logico-Philosophicus. – London, Routledge & Kegan Paul, 1966.

16. http://www.prorobot.ru/04/robot-dog-aibo.php

17. http://1000news.org/robot/robot-sobaka-video

18. http://start.csail.mit.edu

19. http://clef-campaign.org

20. http://clef-qa.itc.it/2008/download/QA@CLEF08_Guidelines-for-Participants_new.pdf

21. http://www.wolframalpha.com/

Институт математики СО РАН, г. Новосибирск
Новосибирский государственный университет
palch@math.nsc.ru

Pal'chunov, D.E. Knowledge search and production: creation of new knowledge on the basis of natural language text analysis

Can human thinking be computer-realized? To solve this problem, we suggest to deal not with an individual computer but with a computer network, e.g. Internet. Such an approach more conforms to the real process of knowledge production in the human society where a new knowledge results from the work of scientific groups and schools but not of an entirely isolated scientist. In addition, it is essential to use the knowledge accumulated earlier and presented in lots and lots of texts written in a natural language.

Keywords: thinking, computer, language, knowledge