

Извлечение знаний из текста и их обработка: состояние и перспективы

Ермаков А.Е.

ООО “ЭР СИ О” (www.rco.ru)

Информационные технологии. - 2009. – N 7. – С. 50-55.

Аннотация

Статья посвящена анализу достижений в области компьютерной обработки знаний, содержащихся в текстах на естественном языке. Формулируются актуальные направления прикладных исследований, связанные с извлечением и обработкой знаний в текстах Интернета. Описывается экспериментальная система для оценки потребительских свойств товаров на основании анализа отзывов их потребителей, размещенных в социальной сети Интернета.

Ключевые слова: извлечение знаний, обработка знаний, автоматизированные системы управления знаниями, компьютерный анализ текста, социальные сети Интернета.

Знания в человеко-машинных системах

Сегодня темой большого количества теоретических и практических исследований является тема построения автоматизированных человеко-машинных систем, которые реализуют комплекс функций, обозначаемых словами "извлечение/управление/обработка знаний". В большинстве случаев под знаниями понимается нечто, выражающееся на естественном языке и изначально содержащееся либо в тексте, либо в голове человека-эксперта. Настоящая статья является результатом критического анализа практических достижений в этой области, попытки определения актуальных направлений развития и собственных экспериментальных исследований в выбранном направлении.

Обзор множества публикаций, некоторое количество которых отмечено в ссылках [1-8], показывает, что в контексте рассматриваемой темы существуют два понимания термина "знание".

Первая точка зрения, акцентированная на прагматических аспектах и принятая в рамках направления knowledge management, представляет знания как данные, полученные в нужном месте и в нужное время для решения практической задачи, обычно для принятия решения, в том числе выполнения действия, человеком или технической системой. При этом по своей структуре и способу хранения знания могут ничем не отличаться от прочих данных – любой фрагмент базы данных или полнотекстового архива документов превращается в знание, как только на него обращается взгляд заинтересованного потребителя. Именно положение этого взгляда – фокуса утилитарного

интереса – определяет, какой фрагмент данных в настоящий момент интерпретируется как знание.

Вторая точка зрения, акцентированная на содержательных аспектах и принятая в рамках направления искусственного интеллекта, полагает, что знания отличаются от обычных данных прежде всего своей структурой. Именно к совокупности особым образом структурированных данных применимо понятие база знаний, подразумевающее:

- логическую упорядоченность данных на основании определенных критериев, устанавливаемых моделью предметной области (онтологией);
- представление данных в соответствии с определенной формальной моделью (семантической сетью, фреймами, набором продукций и др.);
- возможность получения новых данных из старых на основании определенного формального механизма;
- хранение данных в специальных структурах, обеспечивающих высокую эффективность типовых операций над ними (поиск на графах, анализ иерархий, логический вывод и др.).

Полный процесс управления знаниями (knowledge management) в общем случае содержит фазу их извлечения и фазу их обработки, которые реализуются в автоматизированной системе управления знаниями (АСУЗ). При этом технологические составляющие фазы извлечения определяют, каким образом данные превращаются в элементарные знания (аксиомы), а составляющие фазы обработки определяют то, как из элементарных знаний порождается новое знание, используемое для принятия решений. Два подхода к знанию нашли свое отражение в двух совершенно различных типах АСУЗ, автоматизирующих две различные фазы работы со знаниями, которые пока на практике не совмещаются.

Так, при прагматическом подходе к знаниям в центре внимания оказывается фаза их извлечения, поддержка которой обеспечивается информационно-поисковой составляющей АСУЗ. Фаза же обработки найденных первичных знаний реализуется за рамками АСУЗ, в голове аналитика или принимающего решения человека. В итоге, ключевым компонентом АСУЗ является поисковая машина, обеспечивающая оперативный отбор и доставку адекватной информации по запросам. Поисковый компонент либо является предметно-независимым, либо допускает простую настройку даже малоподготовленным пользователем, что определяет основное достоинство этого типа АСУЗ: применимость к широкому кругу заранее неизвестных задач.

При содержательном подходе к знаниям фокус внимания направлен на фазу обработки элементарных знаний, получения из его кирпичиков нового знания на основании обобщения, сопоставления, логического вывода и т.п. В зависимости от формальной модели представления первичных знаний, на этой фазе могут применяться различные математические методы (статистический, регрессионный, кластерный, факторный анализ, анализ графов, вывод в системах продукций и др.). Результатом являются обобщения, выявление скрытых зависимостей и корреляций, прогнозы. В этом случае за рамками АСУЗ реализуется фаза извлечения первичных знаний, их формализации и размещения в базе знаний, логическая структура и фактическое наполнение которой зависят от особенностей предметной области и должны разрабатываться в тесном взаимодействии экспертов и инженера по знаниям. Узкая специализация базы знаний и трудоемкость ее разработки являются недостатками АСУЗ, воплощающих содержательный подход. Достоинством же АСУЗ является возможность быстрого получения решения для тех типовых задач, на решение которых они ориентированы.

Прикладные АСУЗ сегодня

Основными потребителями знаний сегодня выступают следующие группы людей:

1. Руководящие работники, принимающие управленческие решения;
2. Аналитики, составляющие обзоры, прогнозы и рекомендации для (1), в том числе сотрудники спецслужб;
3. Узкие сообщества профессионалов в некоторых областях, для которых разрабатываются специализированные системы – экспертные системы в медицине, геологии; системы извлечения формул органических соединений из научных публикаций в химии и т.п.
4. Прочие работники научной и информационной сферы, нуждающиеся в своевременном и полном получении информации для производства интеллектуальной продукции (например, структурированных новостей по интересующим разделам науки и техники, общественно-политической жизни);
5. Непрофессиональные потребители - люди, желающие использовать знания для бытовых нужд с целью принятия решения, например, о выборе модели товара при покупке, выборе поставщика услуг или способа действия в определенной ситуации (юридические и медицинские вопросы, устранение неисправностей техники).

Как показывают результаты обзора литературы, в том числе интернет-материалов, представленных фирмами-производителями программных решений в области knowledge management, на сегодняшний день все внимание теоретиков и практиков направлено на удовлетворение потребностей групп 1 и 2, что, по-видимому, связано с наибольшей ожидаемой здесь отдачей от капиталовложений. Так, в автоматизированных информационных системах, позиционируемых на рынке как системы поддержки принятия решений, конкурентной разведки, business intelligence, уже присутствует множество подсистем, реализующих те или иные функции извлечения, накопления, поиска и генерации новых знаний. Также определенное внимание привлекают к себе группы (3), для которых, обычно за счет бюджетного финансирования, разрабатываются специализированные прикладные информационные системы, включающие в себя средства data mining и text mining, экспертные системы.

Группы (4) и особенно группа (5), к которой относится каждый человек, обделены вниманием разработчиков и, несмотря на свободный доступ к потенциальному источнику знаний - Интернету, ограничены в инструментарии извлечения знаний простейшими (с точки зрения потребительских функций) поисковыми машинами типа Google или Яндекс. Отчасти это мотивировано необозримой широтой интересов данных групп, отсутствием ограниченной предметной области.

Окончательно, автору не удалось обнаружить не только полноценной АСУЗ, совмещающей в себе фазу извлечения знаний из текста с фазой их обработки, но даже убедительного примера практически полезной работы такой системы. Прикладных программ, использующих методы искусственного интеллекта, способные нетривиально перерабатывать извлеченные из текста элементы знаний (интерпретировать, обобщать, выявлять зависимости, прогнозировать и т.п.), сегодня не существует даже для английского языка. Такая ситуация обусловлена, по-видимому, двумя причинами. Во-первых, слабым распространением систем лингвистического анализа текста, способных интерпретировать отношения между словами и потому действительно извлекать знания как некие элементы, обладающие внутренней структурой и пригодные для нетривиальной смысловой обработки искусственным мозгом – такие системы понимания текста на мировом и российском рынках только недавно начали появляться и еще не успели обрести приложениями: Net Owl (www.netowl.com), Attensity (www.attensity.com), RCO Fact Extractor (www.rco.ru). Во-вторых, потенциально низкой достоверностью автоматически извлекаемых из текста утверждений и фактов, что обусловлено как несовершенством алгоритмов интерпретации текста, так и низким качеством источников информации,

поскольку практически интересно извлечение знаний не из научной литературы, а из различного рода текстовых “помоек”, к каковым относятся социальные сети Интернет, современные СМИ, и даже архивы научно-технических отчетов.

В итоге, несмотря на бум вокруг необходимости извлечения знаний из текста, их переработки и утилизации, поднятый сегодня разработчиками и продавцами АСУЗ, создается впечатление, что на практике такие системы пока бесполезны, во всяком случае, за пределами узко специализированных областей, что, однако, не верно.

Интернет как источник знаний

Современное состояние Интернета позволяет рассматривать его в качестве источника самых разнообразных знаний, которые скрываются в корпоративных интернет-порталах и домашних страничках экспертов, блогах и форумах, аналитических статьях. Сегодня в сети существуют тысячи профессиональных сообществ, объединенных интересами в общей сфере: обсуждение научных проблем и технологий (нанотехнологии, искусственный интеллект), потребление определенных классов товаров и услуг (автомобили, туристическое обслуживание), общественно-политические события и исторические проблемы (международные отношения, история церковного раскола).

Из разбросанных по Интернету знаний стоит выделить следующие классы, представляющие утилитарный интерес для обширных целевых аудиторий:

- Знания о технических и качественных характеристиках товаров и услуг, позволяющие произвести их сравнение и выбрать оптимальный вариант для покупки: электронные устройства и бытовая техника, автомобили; услуги по туризму, ремонту, лечению и т.д.;
- Знания о способах и особенностях использования технологий: ремонт и отделка жилья, устранение неполадок автомобилей и т.п.;
- Научные, технологические и общественно-политические события: открытия и находки, появление новых продуктов и технологий, происшествия и прогнозы;
- Полезные факты различной природы, характеризующие деятельность людей и организаций: историко-биографические факты, взаимоотношения и связи.

Как особый источник информации сегодня следует выделить социальные сети Интернета (блоги, форумы, конференции и прочие виды электронных сообществ), терабайты текстовых сообщений в которых содержат реальные элементы утилитарных знаний, полученные людьми в результате их профессиональной и бытовой деятельности.

В качестве примера рассмотрим известный в Интернете блог “Живой Журнал” (<http://www.livejournal.ru/>) – сеть электронных дневников пользователей, которые делают записи (посты) в своих дневниках и комментарии на записи других пользователей в своих и чужих дневниках. По нашим оценкам, летом 2007 года русскоязычная часть блога содержала более 75 тысяч тематических сообществ, записи в блоге оставили более 1 миллиона 200 тысяч пользователей, а в день в среднем добавлялось около 100 тысяч постов и 400 тысяч комментариев. Следует ожидать, что объем информации и количество пользователей будут увеличиваться в несколько раз ежегодно.

К актуальным задачам, которые сегодня должны и могут решаться АСУЗ в Интернете, по мнению автора, относятся следующие:

- Поиск и извлечение элементов знания, явно присутствующих в текстах в виде: а) утверждений (*лекарство Антипилин – полная ерунда; вероятная причина свиста под капотом автомобиля в сырую погоду – слабое натяжение ремня генератора*); б) фактов (*после принятия Антипилина может подниматься давление; фирма Пежо отозвала 20000 автомобилей из-за возможного возгорания в системе электроусилителя руля*).

- Порождение обобщенного знания, скрытого в совокупности частных утверждений и/или фактов, например, порождение выводов типа *препарат Антипилин имеет меньше побочных эффектов, чем Глипирон* (на основании анализа отзывов больных) или *Типичная причина поломок автомобиля Форд Фокус – засорение бензонасоса* (на основании анализа отзывов владельцев автомобилей).

Возможность практического решения подобных задач сегодня исследуется в компании “ЭР СИ О”, с использованием собственного лингвистического анализатора русского текста (<http://www.rco.ru>).

Опыт извлечения знаний из Интернета: оценка потребительских свойств товаров

Задача разработанной экспериментальной АСУЗ состояла в том, чтобы для каждой модели автомобиля “выловить” положительные и отрицательные отзывы и классифицировать их: “за что хвалят/ругают?”.

Для извлечения знаний была выбрано крупнейшее из нескольких десятков автомобильных сообществ “Живого журнала” сообщество AUTO_RU: “Все об автомобилях” (http://community.livejournal.com/avto_ru/). За 2007 год сообщество содержит записи 3000 авторов постов и 6 тысяч авторов комментариев, всего около 500 тысяч сообщений, порожденных 19 тысячами постов, с объемом русскоязычного текста около 60 Мбайт.

Web-интерфейс пользователя АСУЗ, обеспечивающий просмотр извлеченных из текста знаний, построен на базе трех взаимосвязанных окон (Рис. 1).

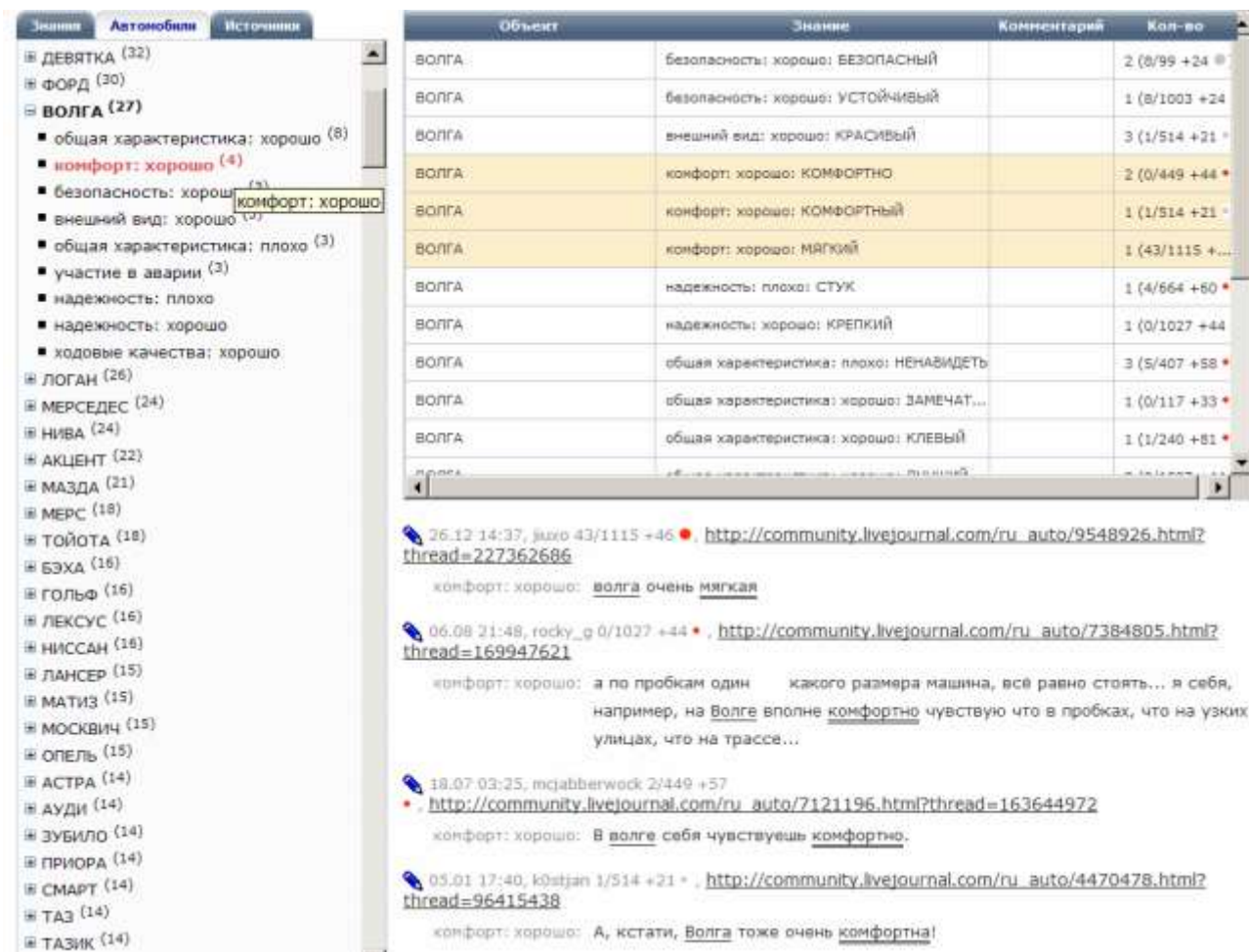


Рис. 1. Вход в базу знаний от выбранного объекта (*Волга*) с просмотром знаний выбранного типа (*комфорт: хорошо*).

Окно в левой части экрана содержит перечень марок автомобилей, по которым в базе содержатся выделенные знания. Тривиальная операция отождествления синонимичных названий (*БМВ = бумер = бэха*, *ВАЗ 2109 = Девятка*) и их группировки по производителям ($VAZ = \{ VAZ\ 2105, \dots, жигули, жига, таз, тазик, \dots \}$) пока не проводилась. Вход в базу от выбранной марки или модели позволяет просмотреть все извлеченные знания, относящиеся к ней, которые сгруппированы по типу: *общая характеристика: хорошо; комфорт: хорошо; надежность: плохо* и т.п. На каждый тип знаний указано количество записей, содержащихся в базе.

Окно в верхней части экрана справа содержит таблицу с записями, относящимися к выбранному типу знания по выбранной марке. Первый столбец содержит название марки; второй - тип извлеченного знания + конкретное ключевое слово, извлеченное из текста,

например, *безопасность: хорошо: УСТОЙЧИВЫЙ*; третий - дополнительную извлеченную информацию (дату, место); а четвертый - характеристики достоверности: количество упоминаний знания в текстах и характеристики наиболее достоверного из источников, из которого были извлечены знания: его относительный вес в сообществе (индекс цитирования) и характер (средняя позитивность или негативность высказываемых им оценок).

Окно в нижней части экрана справа отображает цитаты из источника, в которых упоминается извлеченное знание - предложения исходного текста, выделенные анализатором текста. Для каждой цитаты указывается источник с его характеристиками, а также ссылка на полный текст сообщения для просмотра контекста цитаты.

Для оценки высказываний об автомобилях с точки зрения характеристик их потребительских свойств (*положительная/отрицательная*) была разработана экспериментальная онтология, содержащая:

а) более 700 различных наименований марок автомобилей (*Легенда, Сивик, Приора*) и фирм-производителей (*BMW, БМВ, ВАЗ*). Первоначальный список был получен с сайта <http://www.auto.ru/>, после чего все наименования были расширены синонимами с учетом вариантов написания и известных “народных” названий (*Mitsubishi = Мицубиси = Митсубиши = Мицу; BMW = БМВ = бумер = бэха*). Конкретные символно-цифровые обозначения моделей, упоминаемые в тексте (*BMW 325i, ВАЗ 21053*), не включались в словарь, а распознавались по формальным правилам.

б) более 1200 терминов в 24 группах, среди которых:

- 211 наименований узлов автомобиля (*движок, коробка передач, ходовая часть*);
- 71 наименование свойств, которые классифицированы на 8 оцениваемых групп (*ходовые качества, комфорт, безопасность, надежность, ...*);

- 882 наименования оценок характеристик узлов и свойств, включающие прилагательные, существительные, глаголы и наречия (*крутой, поломка, глючить, отстойно*), в том числе и нелегитимную лексику;

- 37 эмоциональных характеристик (*любить, жалоба, плевать*).

в) около 100 семантических шаблонов, описывающих возможные синтаксические связи в предложении между 24 группами терминов (б) из онтологии. На рисунке 2 приведен пример одного из шаблонов, предназначенного для семантической интерпретации фраз, построенных по схеме типа: *Размер багажника на Outlander XL вызывает восторг; Вид салона Некси приводит в бешенство*. Описание семантического

интерпретатора с соответствующим лингвистическим анализатором текста приведено в работе [9].

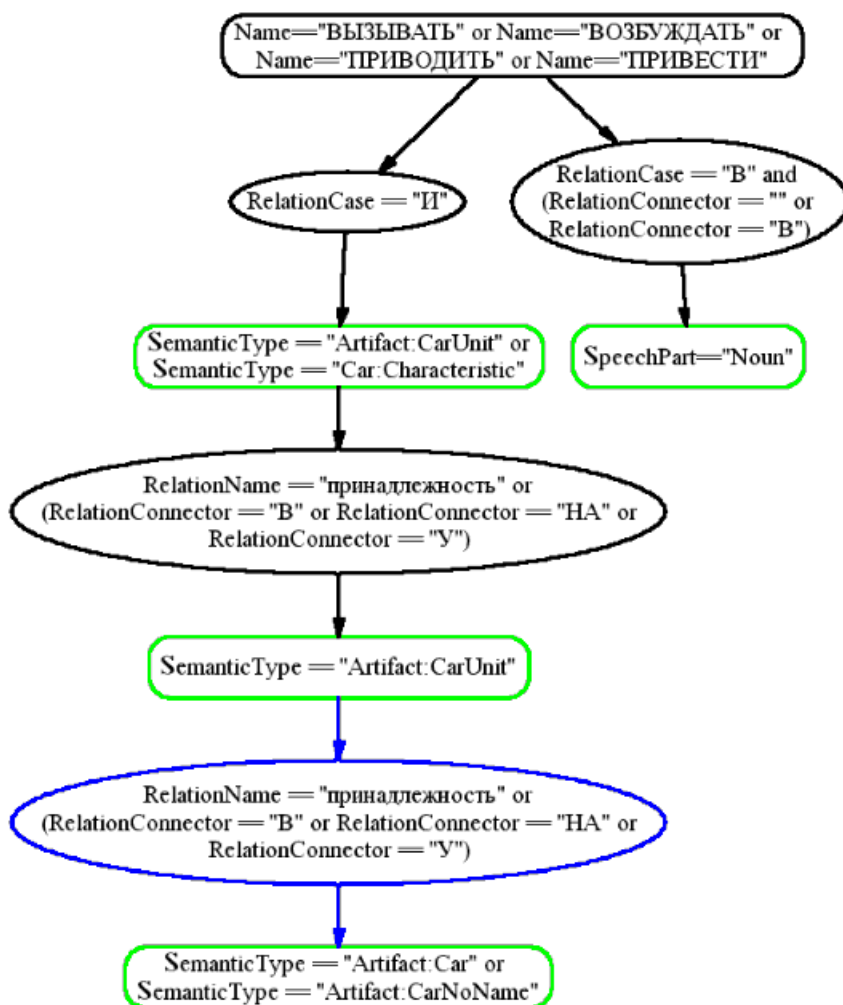


Рис. 2. Семантический шаблон для извлечения оценки автомобиля, которая выражается существительным в конструкциях вида: *Размер багажника на Outlander XL вызывает восторг*; *Вид салона Нексии приводит в бешенство*. Шаблон задает лексико-грамматические ограничения на искомую конфигурацию связей между словами в тексте, которые определяются синтаксическим анализатором. В вершинах (прямоугольниках со скругленными краями) указываются ограничения на конкретные слова ($Name="ВЫЗЫВАТЬ"$), части речи ($SpeechPart="Noun"$ - существительное) или семантические разряды слов ($SemanticType="Artifact:CarUnit"$ – узел автомобиля). В эллипсах описываются ограничения на синтактико-семантические связи между словами: тип связи ($RelationName="принадлежность"$), предлог ($RelationConnector="НА"$), семантический падеж ($RelationCase="И"$ – именительный, субъект). Окончательно, такой шаблон параметризуется множеством конкретных слов из онтологии – названиями

эмоций типа *восторг*, *бешенство* параметризуется узел с ограничениями *SpeechPart="Noun"*, названиями узлов автомобиля (*багажник*, *салон*) и их характеристик (*размер*, *вид*) параметризуются узлы с ограничениями *SemanticType="Artifact:CarUnit"* и *SemanticType="Artifact:CarCharacteristic"*.

Автоматизированная разработка онтологии проводилась на базе анализа языкового материала сообщества AUTO_RU "Живого журнала" при помощи средств компьютерного анализа текста [10], на что было затрачено около 5 чел/дней работы эксперта-автомобилиста (формирование словаря терминов, их классификация по разделам онтологии) и около 50 чел/дней работы лингвиста (отбор и систематизация типовых способов выражения, создание соответствующих им семантических шаблонов, составление соответствующих словарей оценочных слов, общая настройка и тестирование созданного лингвистического обеспечения).

В итоге, из 500 000 сообщений "Живого Журнала" (60 Мбайт текста) было извлечено всего более 5000 оценок автомобилей, их узлов и характеристик, из которых более 1000 (795 хороших и 328 плохих) оценок привязано к маркам автомобилей, а более 4000 оценок узлов и характеристик программе не удалось привязать к конкретным маркам (в предложениях с синтаксически невыраженным референтом, наподобие: *А движок – просто зверь*). В результате достигнута точность 84%, а полнота извлечения около 20%.

Анализ ошибок показывает, что как точность, так и полнота могут быть еще повышены за счет дальнейшей доработки онтологии, но незначительно. В частности, принципиальную проблему представляет интерпретация фраз, содержащих отрицание, выраженное сложным образом, например:

- логически: *Я нигде, никогда, не писал, что Mercedes, BMW и Ягуар имеют высокую надежность; Расскажите про слабую подвеску CR-V кому-то другому; А вообще чтоб в такую погоду на 60 Лексус занесло - это бред!*

- метафорически: *За 7 лет от надежности Сивика остались одни воспоминания; Вольво c40 за милую душу хлебает водички и радуют владельца счётом на 2000 дензнаков ненашенских.*

Заключение

Автоматизированные системы извлечения и обработки знаний, не нашедшие пока практического применения за пределами узкоспециализированных областей, имеют

реальную перспективу войти в повседневную жизнь в ближайшем будущем, используя социальные сети Интернета в качестве источника знаний.

В проведенном эксперименте удалось показать практическую возможность извлечения из социальных сетей Интернета утилитарных знаний, полезных каждому человеку. Учитывая, помимо большого количества грамматических и орфографических ошибок, особый аграмматичный и аорфографичный стиль этого "жанра" текстов (http://ru.wikipedia.org/wiki/Жаргон_падонков), большое количество сленговой лексики, следует признать результаты эксперимента более чем удовлетворительными. Окончательно ожидаемая точность в районе 90% вселяет надежду на возможность извлечения знаний из "интернет-помоек" с приемлемым качеством, поскольку недостаточная полнота (20%) легко компенсируется избытком информации. Следует также учесть, что задача решалась в предельно сложной постановке – автоматически определить "хвалят или ругают?". Подобная задача уже решалась автором для текстов СМИ [11]. И там и здесь, похоже, оказывается не столь важно понять – хвалят или ругают, важнее понять – за что? В более узкой постановке – поиск оценочных высказываний, их отнесение к оцениваемым объектам (что хвалят/ругают?) и классификация (за что? – за двигатель, подвеску, проходимость, разгон, надежность и т.п.) – задача решается с точностью, близкой к 100%.

Область дальнейших исследований, определяющая окончательную эффективность утилизации извлекаемых знаний, - это разработка концепции пользовательского интерфейса АСУЗ, позволяющего минимизировать время на изучение отзывов по интересующим объектам и выработку их сравнительной оценки. В этой области могут помочь методы автоматической оценки достоверности извлеченных знаний, оценки компетентности их источников (авторов сообщений) и степени доверия к ним. Поиск экспертов, мнению которых можно доверять – это один из путей к скорейшему принятию решения.

На взгляд автора, решение рассмотренных задач имеет приоритетное народно-хозяйственное значение в сфере мирных приложений технологий обработки знаний, которые могут быть внедрены в массовые интеллектуальные системы поддержки принятия решений, в частности, помогающие в выборе товаров и услуг на основании анализа отзывов их потребителей.

Литература

1. Гаврилова Т., Хорошевский В. Базы знаний интеллектуальных систем: Учебник для вузов. - СПб.: Питер, 2000. - 384 с.

2. Девятков В.В. Системы искусственного интеллекта: Учеб. пособие для вузов. – М.: Изд. МГТУ им Н.Э.Баумана, 2001. – 352 с.
3. Букович У., Уильямс Р. Управление знаниями: руководство к действию: Пер. с англ. – М.: ИНФРА-М, 2002. - 504 с.
4. Джанетто К., Уиллер Э. Управление знаниями: Руководство по разработке и внедрению корпоративной системы управления знаниями: пер. с англ. - Добрая книга, 2005. - 192с.
5. Колесов А. А управлять – так знаниями! // Byte. - N.2 - М., 2002.
6. Петелин Д. Свалки данных и системы управления знаниями // PC Week (RE) - N.19 - 2006
7. Ландэ Д.В. Поиск знаний в Internet - М.: Диалектика-Вильямс, 2005. -272с.
8. Ашманов И. Информация и знания: невидимая грань - <http://newasp.omskreg.ru/intellect/f5.htm>
9. Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. – Москва, Наука, 2004.
10. Ермаков А.Е. Автоматизация онтологического инжиниринга в системах извлечения знаний из текста // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции "Диалог" (Бекасово, 4-8 июня 2008 г.). Вып. 7 (14). – М.: РГГУ, 2008.
11. Ермаков А.Е., Киселев С.Л. Лингвистическая модель для компьютерного анализа тональности публикаций СМИ // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2005. – Москва, Наука, 2005.