



Онтологический инжиниринг

**в системах извлечения знаний
из текста**

*опыт машинного анализа сообщений
блога “Живой Журнал”
(www.livejournal.ru)*

*Александр Ермаков, ООО «ЭР СИ О»
ermakov@rco.ru, ermakov@rco.ru, www.rco.ru*

Знания в текстах: откуда, что и зачем извлекать?

Извлечение из Интернета первичных элементов знания:

- а) утверждения (*лекарство Антипилин – полная ерунда; вероятная причина свиста под капотом автомобиля в сырую погоду – слабое натяжение ремня генератора*);
- б) факта (*после принятия Антипилина может подниматься давление; летом 2006 фирма Пежо отозвала 20000 автомобилей из-за возможного возгорания в системе электроусилителя руля*).

Порождение сложного знания из элементов знания:

- а) логический вывод, например: *продукт X некачественный* (утверждение), *X - продукт компании Y в 1997* (факт), *Z - технический директор компании Y с 1996 по 1998 годы* (факт), следовательно, *Z - плохой руководитель* (знание);
- б) обобщение, например, порождение выводов: *препарат Антипилин имеет меньше побочных эффектов, чем Глипирон* (на основании статистики отзывов больных) или *Типичная причина поломок автомобиля Форд Фокус – засорение бензонасоса* (на основании статистики сообщений автомобилистов).

Социальные сети в Интернет: главный источник знаний

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 состоянию на лето 2007 года русскоязычная часть блога содержит:

- более 75 тысяч тематических сообществ;
- более 1 миллиона 200 тысяч пользователей;
- в день добавляется около 100 тысяч постов и 400 тысяч комментариев.

Сообщество **auto_ru** (“Все об автомобилях”) – крупнейшее из автомобильных.

В целом за 2007 год:

- 500 тысяч сообщений, порожденных 19 тысячами постов;
- 3000 авторов постов и 6 тысяч авторов комментариев;
- объем русскоязычного текста около 60 Мбайт.

Знания об автомобилях из Интернет-сообщества (скриншот 1)

Знания | **Автомобили** | **Источники**

- ⊕ ДЕВЯТКА (32)
- ⊕ ФОРД (30)
- ⊕ **ВОЛГА (27)**
 - общая характеристика: хорошо (8)
 - **комфорт: хорошо (4)**
 - безопасность: хорошо (3)
 - внешний вид: хорошо (3)
 - общая характеристика: плохо (3)
 - участие в аварии (3)
 - надежность: плохо
 - надежность: хорошо
 - ходовые качества: хорошо
- ⊕ ЛОГАН (26)
- ⊕ МЕРСЕДЕС (24)
- ⊕ НИВА (24)
- ⊕ АКЦЕНТ (22)
- ⊕ МАЗДА (21)
- ⊕ МЕРС (18)
- ⊕ ТОЙОТА (18)
- ⊕ БЭХА (16)
- ⊕ ГОЛЬФ (16)
- ⊕ ЛЕКСУС (16)
- ⊕ НИССАН (16)
- ⊕ ЛАНСЕР (15)
- ⊕ МАТИЗ (15)
- ⊕ МОСКВИЧ (15)
- ⊕ ОПЕЛЬ (15)
- ⊕ АСТРА (14)
- ⊕ АУДИ (14)
- ⊕ ЗУБИЛО (14)
- ⊕ ПРИОРА (14)
- ⊕ СМАРТ (14)
- ⊕ ТАЗ (14)
- ⊕ ТАЗИК (14)

Объект	Знание	Комментарий	Кол-во
ВОЛГА	безопасность: хорошо: БЕЗОПАСНЫЙ		2 (8/99 +24)
ВОЛГА	безопасность: хорошо: УСТОЙЧИВЫЙ		1 (8/1003 +24)
ВОЛГА	внешний вид: хорошо: КРАСИВЫЙ		3 (1/514 +21)
ВОЛГА	комфорт: хорошо: КОМФОРТНО		
ВОЛГА	комфорт: хорошо: КОМФОРТНЫЙ		
ВОЛГА	комфорт: хорошо: МЯГКИЙ		
ВОЛГА	надежность: плохо: СТУК		
ВОЛГА	надежность: хорошо: КРЕПКИЙ		1 (0/1027 +44)
ВОЛГА	общая характеристика: плохо: НЕНАВИДЕТЬ		3 (5/407 +58)
ВОЛГА	общая характеристика: хорошо: ЗАМЕЧАТ...		1 (0/117 +33)
ВОЛГА	общая характеристика: хорошо: КЛЕВЫЙ		1 (1/240 +81)

26.12 14:37, jiuho 43/1115 +46 • http://community.livejournal.com/ru_auto/9548926.html?thread=227362686
комфорт: хорошо: волга очень мягкая

06.08 21:48, rocky_g 0/1027 +44 • http://community.livejournal.com/ru_auto/7384805.html?thread=169947621
комфорт: хорошо: а по пробкам один какого размера машина, всё равно стоять... я себя, например, на Волге вполне комфортно чувствую что в пробках, что на узких улицах, что на трассе...

18.07 03:25, mcjabberwock 2/449 +57 • http://community.livejournal.com/ru_auto/7121196.html?thread=163644972
комфорт: хорошо: В волге себя чувствуешь комфортно.

05.01 17:40, k0stjan 1/514 +21 • http://community.livejournal.com/ru_auto/4496415438.html?thread=96415438
комфорт: хорошо: А, кстати, Волга тоже очень комфортна!

Знания по объекту Волга: оценки потребительских свойств автомобиля

Объекты оценки: марки автомобилей

Подкрепление знаний: цитаты из сообщений с ссылками в текст

Знания об автомобилях из Интернет-сообщества (скриншот 2)

Знания | **Автомобили** | Источники

- отзывает автомобили (12)
- открывает новый завод
- практичность: плохо (151)
- практичность: хорошо (206)
- проходимость: плохо (67)
- проходимость: хорошо (54)
- участие в аварии (60)**
 - ДЖИП (4)
 - ВОЛГА (3)**
 - ДЕВЯТКА (3)
 - МЕРСЕДЕС (3)
 - АВТОМОБИЛЬ "ЖИГУЛИ" (2)
 - БМВ (2)
 - ЛОГАН (2)
 - НИВА (2)
 - ТАЗ (2)
 - BMW
 - HYUNDAI
 - KIA
 - LEXUS
 - MERCEDES
 - PAJERO SPORT K90
 - АВТОМОБИЛЬ "SUBARU"
 - АВТОМОБИЛЬ "ПЕЖО"
 - АВТОМОБИЛЬ "ЛЕКСУС"
 - АВТОМОБИЛЬ «МИТСУБИШИ»
 - АВТОМОБИЛЬ «ГАЗ»
 - АВТОМОБИЛЬ ВАЗ
 - АВТОМОБИЛЬ ВОЛГА
 - АКЦЕНТ
 - АМУЛЕТ

АКЦЕНТ	участие в аварии: СТОЛКНОВЕНИЕ	1 (0/311 +46)
АМУЛЕТ	участие в аварии: АВАРИЯ	1 (0/96 +46)
БМВ	участие в аварии: СТОЛКНОВЕНИЕ	
БМВ	участие в аварии: СТОЛКНУТЬСЯ	
ВАЗ 2114	участие в аварии: ПЕРЕВЕРНУТЬСЯ	
ВАЗ 2115	участие в аварии: АВАРИЯ	
ВОЛГА	участие в аварии: АВАРИЯ	
ВОЛГА	участие в аварии: ВРЕЗАТЬСЯ	
ВОЛГА	участие в аварии: ПРОТАРАНИТЬ	
ДЕВЯТКА	участие в аварии: ВРЕЗАТЬСЯ	1 (0/117 +1...)
	участие в аварии: ПЕРЕВЕРНУТЬСЯ	1 (0/200 +27)
	участие в аварии: СТОЛКНУТЬСЯ	1 (0/429 +42)

kenjin 0/117 +100 • , [city fm](#)

ВОЛГА: а пять мин назад в блоке новостей сообщили, что в какой то деревне врезались волга и девятка

14.03 17:45, antonaz 0/172 0 • , http://community.livejournal.com/ru_auto/5197750.html?comment=117535158

ВОЛГА: В моего отца, пересекавшего шоссе Энтузиастов на ВАЗ-2104, на красный протаранила Волга военная прямо в водительскую дверь.

2 13:54, ge1 0/181 +14 • , [Ночные гонки с препятствиями](#)

ВОЛГА: Водитель автобуса видит перед собой аварию и начинает уходить влево, "приминая" всех гонщиков... я вижу, что могу пролезть между ограждением и попавшей в аварию волгой.

Типы извлеченных знаний: оценки свойств автомобилей и полезные факты

Подкрепление знаний: цитаты из сообщений с отсылками в текст

Полезные факты по объекту Волга: участие в авариях (к оценке безопасности: а что останется от автомобиля?)

Извлечение знаний из Интернета: оценка потребительских свойств товаров на основании анализа отзывов

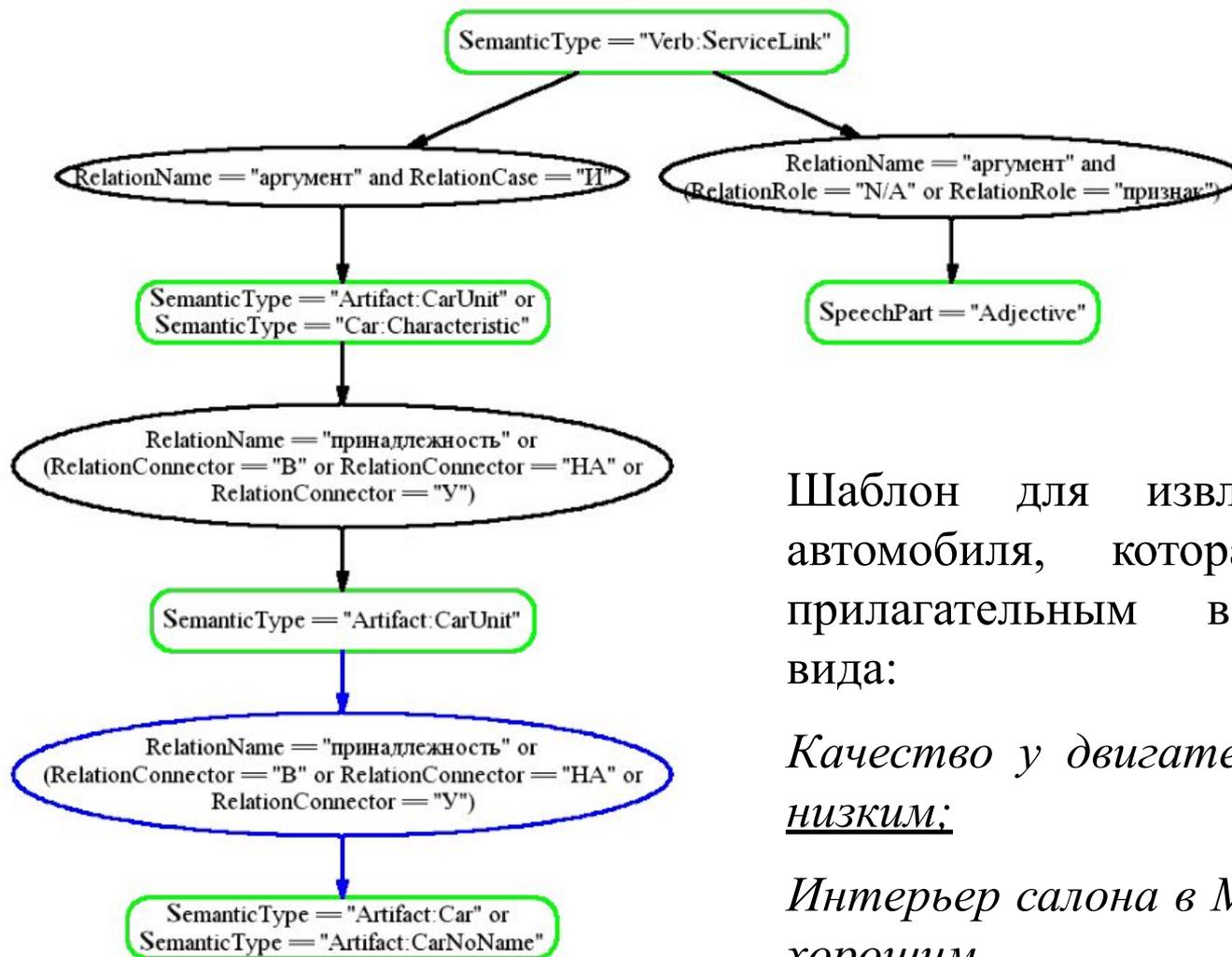
Задача: для каждой модели автомобиля "выловить" положительные и отрицательные отзывы и классифицировать их: за что хвалят/ругают?.

Экспериментальная онтология для оценки автомобилей с точки зрения характеристик (*положительная/отрицательная*) их потребительских свойств. Содержит более 1200 терминов (24 группы):

- 211 наименований узлов (*движок, коробка передач, ходовая часть*);
- 71 наименование свойств 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*);
- 37 эмоциональных характеристик (*любить, жалоба, плеваться*).

Синтаксические связи в предложении между 24 группами терминов из онтологии описываются около 100 семантических шаблонов.

Извлечение знаний: семантическая интерпретация текста (1)

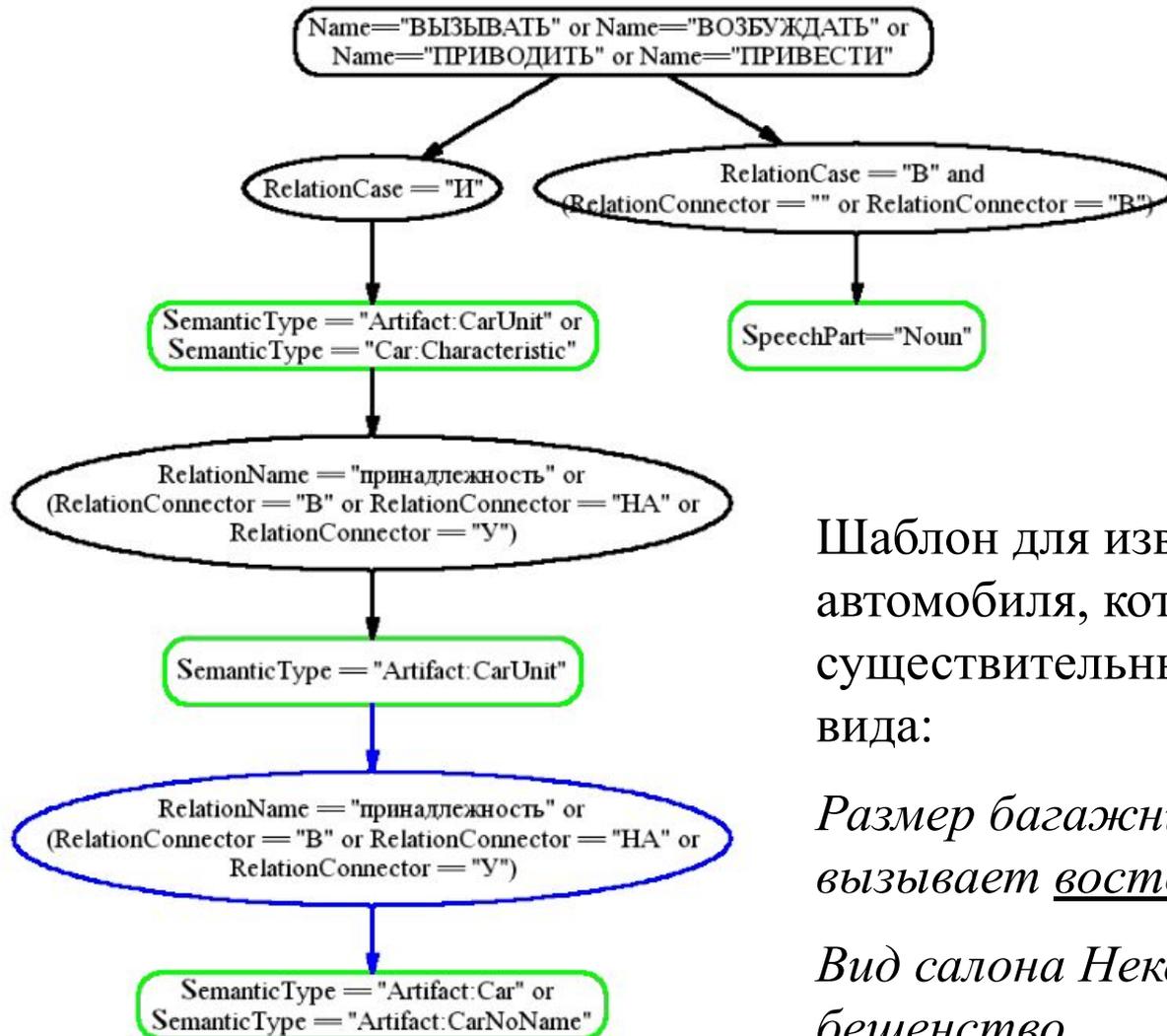


Шаблон для извлечения оценки автомобиля, которая выражается прилагательным в конструкциях вида:

Качество у двигателя Опеля стало низким;

Интерьер салона в Мазде считается хорошим.

Извлечение знаний: семантическая интерпретация текста (2)



Шаблон для извлечения оценки автомобиля, которая выражается существительным в конструкциях вида:

Размер багажника на Subaru вызывает восторг;

Вид салона Нексии приводит в бешенство.

Извлечение знаний из Интернета: результаты

Из 500 000 сообщений “ЖЖ” (60 Мбайт текста) извлечено:

- всего более 5000 оценок автомобилей, их узлов и характеристик;
- более 1000 (795 хороших и 328 плохих) оценок привязано к маркам автомобилей;
- более 4000 оценок узлов и характеристик не удалось привязать к конкретным маркам (связь с референтом анафорическая);

Достигнута точность: 84%

Оценка полноты: около 20%

Спасибо за внимание!

Александр Ермаков, ООО «ЭР СИ О»
ermakov@rco.ru, ermakov@rco.ru, www.rco.ru

