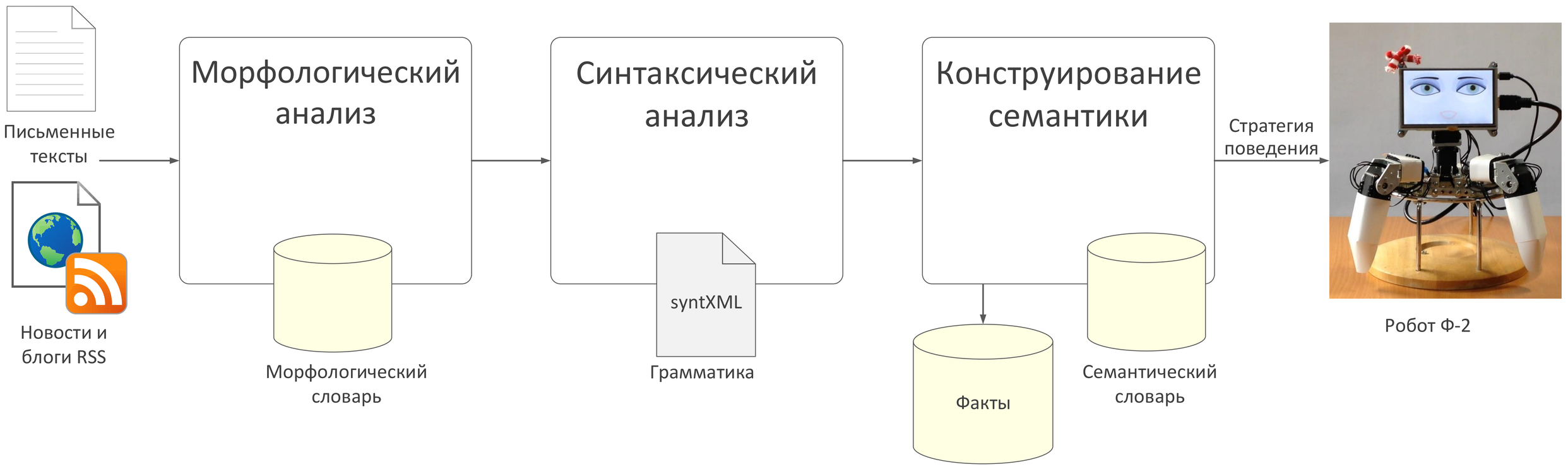
Система понимания текста для робота Ф-2: синтаксический анализ и извлечение смысла

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР «КУРЧАТОВСКИЙ ИНСТИТУТ»

Зайдельман Л. Я. (zaydelman_ly@nrcki.ru), Котов А. А. (kotov_aa@nrcki.ru),
Зинина А. А. (zinina_aa@nrcki.ru), Аринкин Н. А. (arinkin_na@nrcki.ru)

Курчатовский институт (Россия, Москва)



Робот Ф-2

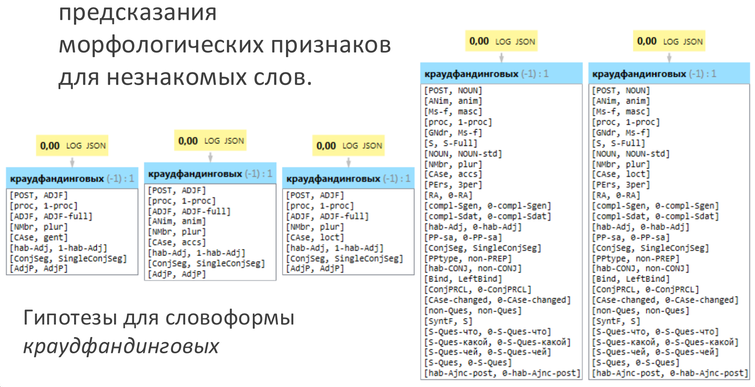
Морфологический анализ

Для каждого слова входящего текста строятся все варианты морфологического разбора.

Словарь словоформ для 100 тысяч лексем. На базе словаря OpenCorpora.

41 регулярное выражение для разбора чисел и буквенно-цифровых комплексов (*12-го, 12,4%*).

Нейросетевой алгоритм предсказания морфологических признаков для незнакомых слов.



Гипотезы для словоформы *краудфандинговых*

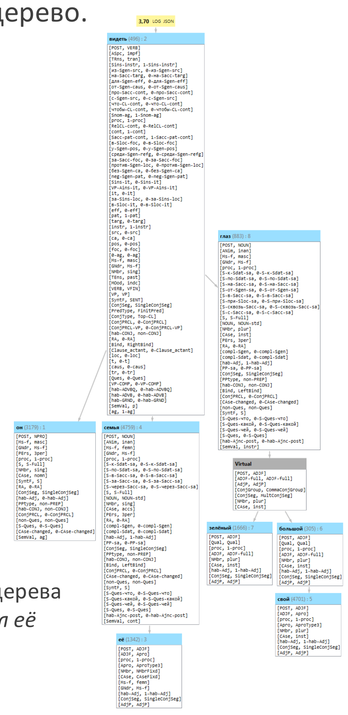
Синтаксический анализ

Слова последовательно помещаются в стек разбора и сворачиваются грамматическими правилами с целью построить синтаксическое дерево.

545 грамматических правил на языке SyntXML.

Синтаксическая структура комбинирует в себе принципы грамматики зависимостей и грамматики непосредственных составляющих.

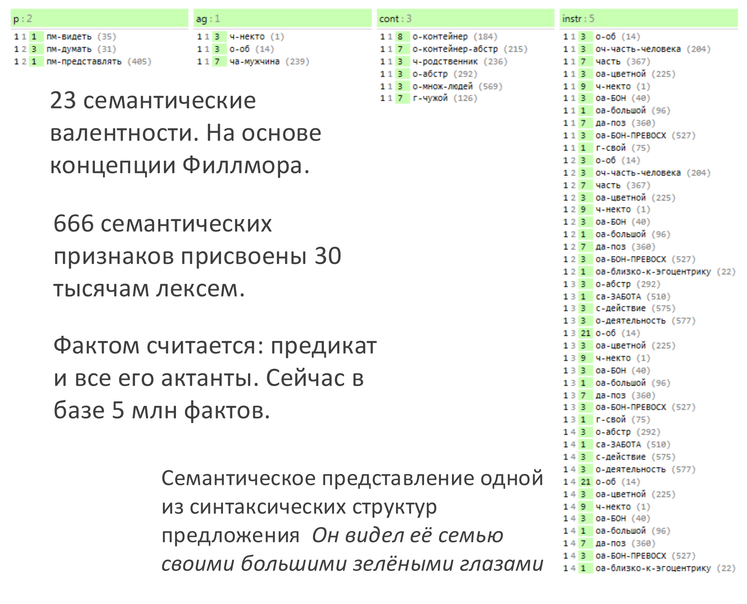
Элементам структуры присваиваются семантические валентности.



Вариант синтаксического дерева для предложения *Он видел её семью своими большими зелёными глазами*

Конструирование семантики

Каждой представленной в предложении семантической валентности ставятся в соответствие семантические признаки слова, занимающего эту валентность, а также признаки зависимых от него слов.



23 семантические валентности. На основе концепции Филлмора.

666 семантических признаков присвоены 30 тысячам лексем.

Фактом считается: предикат и все его актанты. Сейчас в базе 5 млн фактов.

Семантическое представление одной из синтаксических структур предложения *Он видел её семью своими большими зелёными глазами*

Работа с фактами

Семантический портрет лексемы

Выполняется поиск фактов, в структуру которых входит конкретная лексема, определяется занимаемая лексемой валентность.

Семантический портрет лексемы *снег*.

Лексема *снег* встречается в 1294 фактах. И занимает в них 20 разных валентностей.

валентность	n	пример
агенс	390	сухой снег припорошил коричневые борозды
атрибут субстантива	337	писатель сравнивает запах первого снега с запахом арбуза!
пациенс	126	резкие порывы ветра подымали снег с земли.
место	104	люди в одном белье остались босыми на снегу.
целевая точка	74	солдат глухо вскрикнул и упал в снег уже мертвый.
инструмент	52	руки у него, конечно, замерзли, но он тут же растер их снегом.
интерпретация	37	а тут еще и мелкий дождь превратился в первый снег.
содержание	33	тащу и вижу: черный снег такой.
траектория	30	на подъемах ямщик соскакивал с саней и шел по снегу рядом.
причина	22	кругом все белело от снега.
время	21	после снега во сне он твердо знал, что наутро все сладится.
параметр	19	в пятницу в москве наблюдается пасмурная, ветреная погода со снегом.
источник	18	из снега раздался заносчивый возглас:
контрагент	16	город совершенно не чистят от снега.
редуцированный агенс	10	такое похолодание будет сопровождаться усиленным ветром, снегом и ярким солнцем.

Поиск по семантическому шаблону

Выполняется поиск фактов, соответствующих семантическому шаблону: набору пар <валентность – семантический признак>.

2909 фактов соответствует шаблону *некий человек (ag) причиняет вред (p) некоему человеку (pat).*

Самый популярный агенс-вредитель – *он*, самый популярный пациенс-жертва – *она*, пациенса либо *ударяют*, либо уже *убили*.

шаблон	пример
убить(p) я(pat) кто(ag)	кто меня убил?
обидеть(p) что-то(cont) я(pat) ты(ag)	почему тебе кажется, что ты меня чем-то обидела?
убить(p) он(ag) поединок(loc) человек(pat)	всего он убил на поединках десять человек.
побить(p) человек(ag) жена(pat)	так вот эту самую жену тюменские люди побили и отобрали у нее соболью шубу
убить(p) она(pat) ты(ag)	это ты ее убил и закопал в орешнике...
обидеть(p) аллах(ag) избранник(pat) полководец(instr)	или аллах обидел своего избранника полководцами?
беспокоить(p) ты(ag) хозяин(pat)	ты же, кстати, беспокоишь моего замечательного хозяина.
терзать(p) демон(ag) я(pat) коготь(instr)	демоны страха терзали меня железными когтями.
ударять(p) он(ag) она(pat) кулак(instr)	он ударяет ее кулаком по плечу, выскакивает из постели...
мучить(p) изверг(ag) ребёнок(pat)	какой изверг мучил этого ребенка?
уничтожить(p) вы(ag) человек(pat) право(foc)	сколько людей вы уничтожили в соответствии с германским правом?
заразить(p) он(pat) она(ag)	может быть, она его просто заразила?
перебить(p) студент(pat) новиков(ag)	горячо перебил студента новиков.
убить(p) никто(pat) ты(ag)	ты никого не убил?

Автоматическое формирование прототипической категории

Выполняется поиск фактов, соответствующих семантическому шаблону; объединяются результаты по одной из валентностей.

Прототипическая категория «напитки».

В 3970 фактах *пьют (p) что-то (pat)* . Всего – 294 разных пациенса. Только в 707 фактах у пациенса есть признак *еда,* такие пациенсы – центр прототипической категории. Чаще всего пьют *чай* (527). На периферии категории – нетипичные напитки, ёмкости и др (см таблицу).

пациенс	n	пример
вода	172	в полудреме он глотнул воды из чайника и сел перед окном на табуретку.
кровь	172	они пытались вцепиться в горло и выпить кровь!
стакан	58	раздели, выпил стакан спирта, укрыли всеми шубами.
бокал	35	получается – в месяц среднестатистический россиянин выпил 10 бокалов вина.
рюмка	33	и сергей в доказательство навета выпил рюмку, а больше не стал.
лекарство	31	я постоянно пила лекарства, уже начал болеть от этого желудок
что-нибудь	27	хоть бы она предложила чего-нибудь выпить.
таблетка	23	сколько именно таблеток она выпила, не сообщается.
*висок	30	полина заявила, что на нём лица не было и он пил виски.
жидкость	19	он знал, что ему сейчас надо пить побольше жидкости.
литр	19	наши граждане за пять месяцев этого года выпили 1,6 литра на душу населения.
чашка	18	я пью три чашки зелёного чая в день.
ничто	16	дружок, кажется, ты забыл, что я давно не пью ничего, кроме минеральной воды.

Работа поддержана грантом РФФИ 16-29-09601
«Система автоматического выявления эмоциональных и экстремистских суждений в текстах на естественном языке»