Извлечение фактов при помощи шаблонов

Бочаров Виктор

Этапы решения задачи

1) Отбор текстов про спорт
2) Графематика
   a) Разделение на предложения
   b) Разделение на токены
3) Морфология
4) **Извлечение фактов из одного документа**
5) *Связывание фактов между документами**
6) *Идентификация объектов**
7) Парсинг запроса
8) Выполнение запроса

Ожидаемый результат

«московский "Локомотив" на своем стадионе в Черкизово обыграл норвежский "Бранн" со счетом 3:2»

"Локомотив" с трудом прошел в групповой турнир Кубка УЕФА

/news/2005/09/29/lokomotiv/

- Team1 = Локомотив

  Team2 = Бранн

  Placecity = Черкизово

  Result1 = 3

  Result2 = 2

Способы решения

- Регулярные выражения
  - $text =~ /со счетом (\d+):(\d+)/
  - $text =~ /(Зенит|Спартак|ЦСКА|...)/
- Gazetteer + постобработка
  - Найти ключевые цепочки слов
  - Заполнить ими слоты
- Синтаксический анализ + постобработка
  - Построить синтаксическое дерево
  - Найти нужные ветки

Синтаксический анализ

- Состоит из
  - программы
  - набора правила
- Синтаксис у АОТ
  - Программа + правила на C++
  - Программа + правила в виде КС-грамматики

Синтаксический анализ

- Разделение на предложения и токены
- Морфология
- Сборка синтаксического дерева

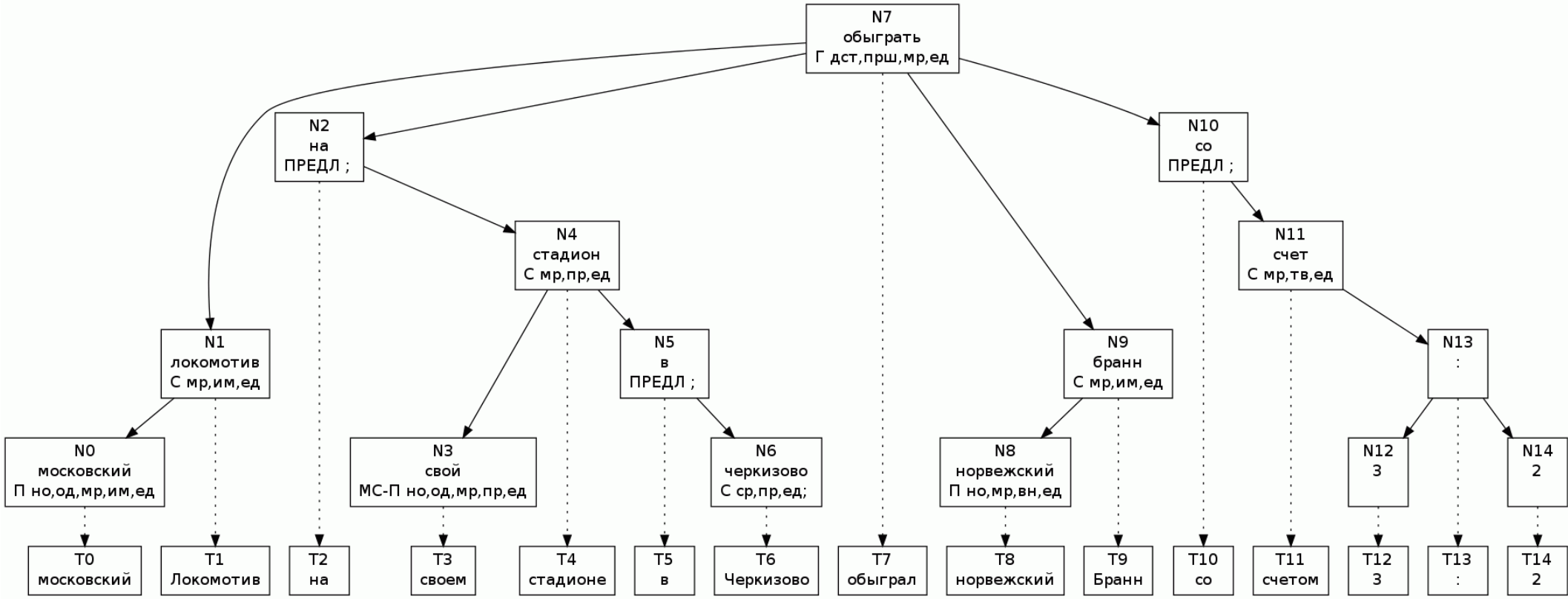
Синтаксический анализ

«московский "Локомотив" на своем стадионе в Черкизово обыграл норвежский "Бранн" со счетом 3:2»

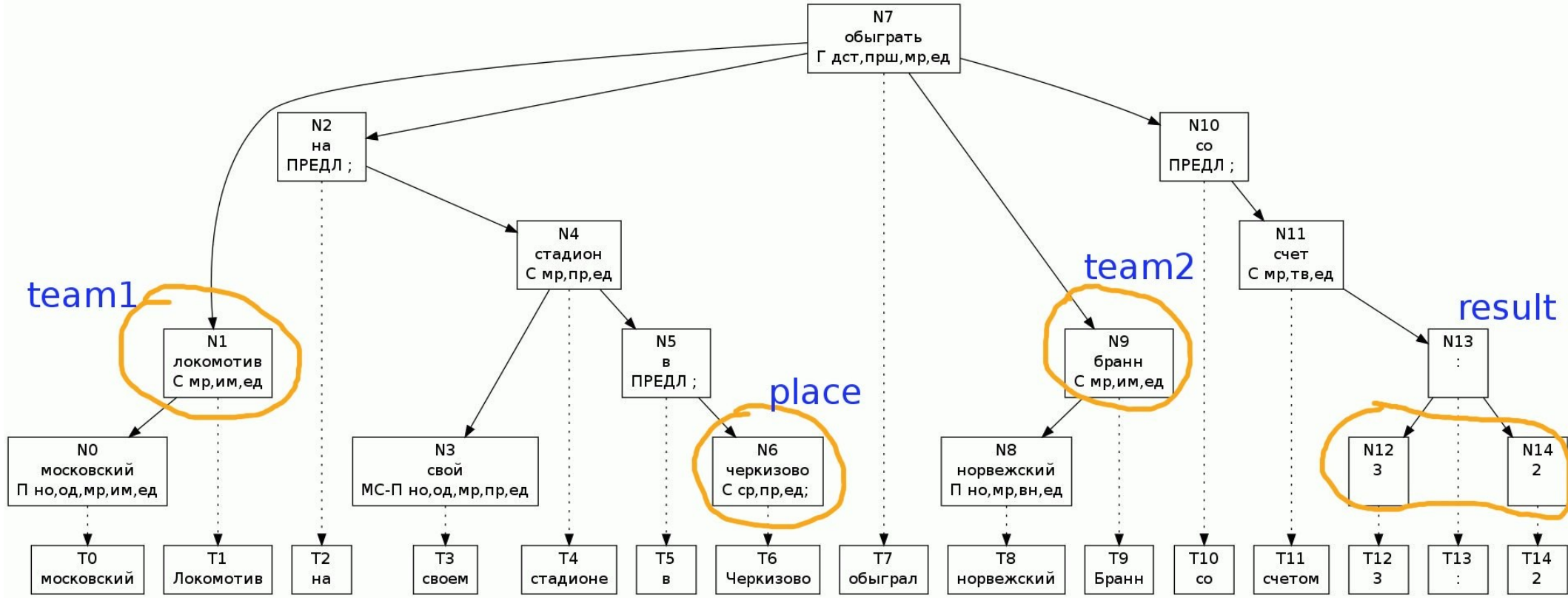
"Локомотив" с трудом прошел в групповой турнир Кубка УЕФА

/news/2005/09/29/lokomotiv/

Синтаксический анализ



Синтаксический анализ



КС-грамматика

- Plain text
- Кодировка — Windows-1251
- Компилируется в двоичное представление
- Привязана к морфологическому словарю АОТ

КС-грамматика

- Терминалы
  - [TOKEN grm="гр1,гр2,..."]
    - [TOKEN grm="С,им"]
    - [TOKEN grm="им,ед"]
  - [TOKEN register="Aa"]
  - Лемма в одинарных кавычках
    - 'СТАДИОН'
    - 'НА'
    - '1'
    - ':'

Граммемы

- http://www.aot.ru/docs/rusmorph.html
- Части речи:
  - С, П, Г, МС, ИНФИНИТИВ, ПРИЧАСТИЕ, КР_ПРИЧАСТИЕ, ПРЕДЛ, …
- Признаки:
  - мр, жр, ср - мужской, женский, средний род
  - од, но - одушевленность, неодушевленность
  - ед, мн - единственное, множественное число
  - им, рд, дт, вн, тв, пр, зв — падежи
  - ...

КС-грамматика

- Не терминалы
  - В прямоугольных скобках
    - [NP]
    - [V_3RD_PAST]
    - [VERY_COMPLEX_IDENTIFIER]
  - Назначаются разработчиком

КС-грамматика

- Правила
  - Левая часть — нетерминал
  - Стрелочка (- >)
  - Правая часть — список терминалов и нетерминалов
  - *Двоеточие*
  - *Ограничения на согласование*
  - Точка с запятой
- [NP] - > [TOKEN grm="П"] [TOKEN grm="С"];

КС-грамматика

- [NP] - >

```
[TOKEN grm="П"]
[TOKEN grm="С"]
: $0.grm
  := case_number_gender(
     $1.grm,
     $2.grm
     );
```

КС-грамматика

- [NP] -> [TOKEN grm="С" root];
- [NP] -> [TOKEN grm="П"] [NP root] : $0.grm := case_number_gender($1.grm, $2.grm);
- [KW] -> 'ОБЫГРАТЬ';
- [PLACE_KW] -> 'СТАДИОН' | 'АРЕНА';
- [PLACE] -> [PLACE_KW] 'в' [TOKEN grm="С,пр"];
- [PLACE_KW] -> 'СТАДИОН' | 'АРЕНА';
- [WHERE] -> 'на' [PLACE grm="пр"];
- [VP] -> [KW root] [NP grm="вн"];
- [S] -> [NP grm="им"] [VP root];

КС-грамматика

- Подключение тезауруса:

  [TEAM] - > 'Зенит' | 'Спартак' | 'Реал' | … ;
  [CITY] - > 'Петербург' | 'Москва' | …;

  [TEAM] - > 'ФК' [TEAM root];

КС-грамматика

- Предложения
  - Клаузы
    - Синтаксические варианты клауз
      - Группы
        - Слова

КС-грамматика

<group name='NP' fw='7' lw='8' main_fw='8' main_lw='8' root='Локомотив'>московский Локомотив </group>

<group name='PLACE' fw='11' lw='13' main_fw='11' main_lw='11' root='стадионе'>стадионе в Черкизово </group>

<group name='PLACE' fw='10' lw='13' main_fw='11' main_lw='13' root='стадионе'>своем стадионе в Черкизово </group>

<group name='WHERE' fw='9' lw='13' main_fw='9' main_lw='9' root='на'>на своем стадионе в Черкизово </group>

<group name='NP' fw='15' lw='16' main_fw='16' main_lw='16' root='Бранн'>норвежский Бранн </group>

<group name='VP' fw='14' lw='16' main_fw='14' main_lw='14' root='обыграл'>обыграл норвежский Бранн </group>

КС-грамматика

<word no='0' lemma='МОСКОВСКИЙ' pos='П' grm='но,од,мр,им,ед,; '
form='московский' main_no='1'/>
<word no='1' lemma='ЛОКОМОТИВ' pos='С' grm='мр,им,ед,; '
form='Локомотив' main_no='7'/>
<word no='2' lemma='НА' pos='ПРЕДЛ' grm='; ' form='на' main_no='7'/>
<word no='3' lemma='СВОЙ' pos='МС-П' grm='но,од,мр,пр,ед,; '
form='своем' main_no='4'/>
<word no='4' lemma='СТАДИОН' pos='С' grm='мр,пр,ед,; ' form='стадионе' main_no='2'/>
<word no='5' lemma='В' pos='ПРЕДЛ' grm='; ' form='в' main_no='4'/>
<word no='6' lemma='ЧЕРКИЗОВО' pos='С' grm='ср,пр,ед; '
form='Черкизово' main_no='5'/>
<word no='7' lemma='ОБЫГРАТЬ' pos='Г' grm='дст,прш,мр,ед,; '
form='обыграл' main_no='-1'/>
<word no='8' lemma='НОРВЕЖСКИЙ' pos='П' grm='но,мр,вн,ед,; '
form='норвежский' main_no='9'/>
<word no='9' lemma='БРАНН' pos='С' grm='мр,им,ед,; ' form='Бранн' main_no='7'/>

Gazetteer

- Выделение именованных сущностей из текста по спискам.

- Реализован в GATE (нет русского языка)
- Можно написать свой

Gazetteer

- Действующий чемпион России по футболу **московский** ***Локомотив*** в матче **4-го тура первенства страны** обыграл **на своем поле казанский** **Рубин** со счетом **1:0** , сообщает РИА Новости .

Gazetteer

- Разделение на предложения и токены
- Морфология
- Выделение слов и словосочетаний по спискам

В чём разница?

- Синтаксис
- 2 инструмента:
  - Грамматика
  - Обход дерева
- Учитываются связи слов
- Лучше на коротких предложениях

- Gazetteer
- 1 инструмент:
  - Gazetteer
- Не учитываются связи слов
- Безразличен к длине предложений

Вопросы?