Лексический анализ: от шаблонов к семантике

Даниил Скатов

ООО «Диктум»
г. Нижний Новгород

Поисковые технологии 2010
Яхрома, 26 февраля 2010 г.

Даниил Скатов

ООО «Диктум»

г. Нижний Новгород

26 февраля 2010 г.

Объекты…

Даниил Скатов

ООО «Диктум»

г. Нижний Новгород

26 февраля 2010 г.

Объекты…

Персона

Даниил Скатов

Имя	"Даниил"
Фамилия	"Скатов"
Отчество	∅

Организация

ООО «Диктум»

Название	"Диктум"
Тип	"ООО"

Населенный пункт

г. Нижний Новгород

Имя	Нижний
Тип	Город

Дата

26 февраля 2010 г.

День	26
Месяц	02
Год	2010

Объекты…

Персона

Даниил Скатов

Имя	"Даниил"
Фамилия	"Скатов"
Отчество	∅

Скатов Даниил ; Скатов Д.

Даниил Сергеевич Скатов; Скатов Д.С.

Даниил Сергеевич; Скатов

Организация

ООО «Диктум»

Название	"Диктум"
Тип	"ООО"

Общество с огр. отв-ю «Диктум»

компания «Диктум» ; Dictum Ltd

Диктум

Населенный пункт

г. Нижний Новгород

Имя	Нижний Новгород
Тип	Город

Н. Новгород; г. Н. Новгород

НН; столица Поволжья; город Горький

Горький; НН; Нижний

Дата

26 февраля 2010 г.

День	26
Месяц	02
Год	2010

26.02.2010 ; Feb 26, 2010

Двадцать шестое февраля

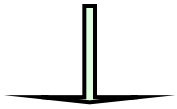
Последняя пятница февраля 2010 года

Объекты, факты …

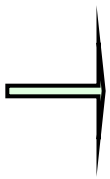
Сотрудник компании «Диктум»
Скатов Даниил (г. Н. Новгород)
26.02.2010 посетил конференцию
«Поисковые технологии»

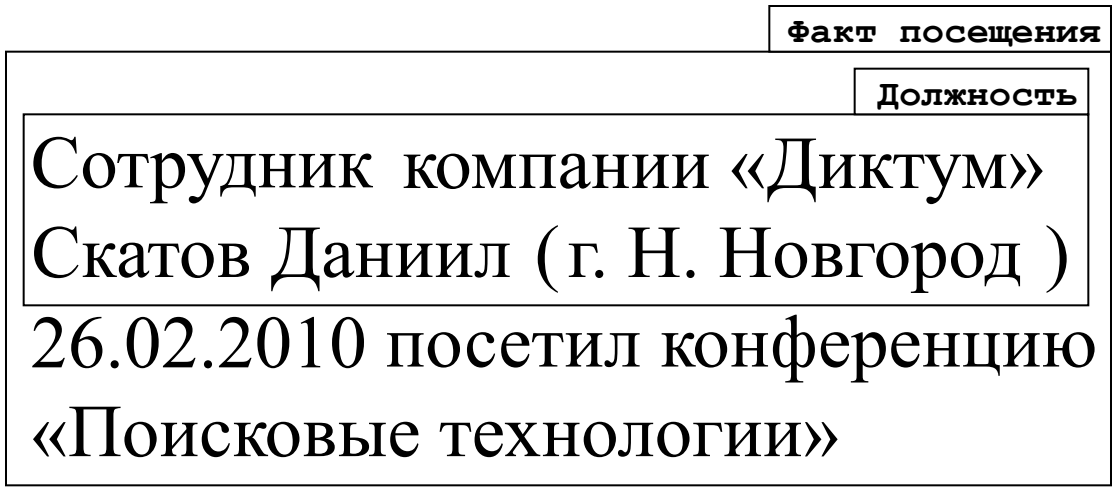
Объекты, факты …

- **Даты:** 20/03/06, 7 февраля 2007 г., 1991-2006 гг.
- **Персоны:** Петров И.С., Иван Петров, Иван Сергеевич, Петров И.
- **Адреса Интернет и e-mail:** http://www.dictum.ru
- **Географические адреса:** Россия, г. Н.Новгород, пр-т Гагарина, 23, корп. 7
- **Названия организаций:** Университет им. Н.И.Лобачевского, КБ «Квазар», Школа № 7
- **Спортивные события:** Зимняя олимпиада, Кубок УЕФА, Чемпионат мира по хоккею
- **Числа прописью:** две тысячи восемьсот единиц техники
- **Результаты измерений:** 8 кг., не более 50 км/ч
- **Денежные единицы:** 2 000 р., 80 454,2 USD
- **Порядковые числительные:** 1-ый, 18-ого
- **Номера телефонов:** (831) 278-67-57, +79200459731
- Номера кредитных карт, ИНН
- …



Факты — отношения между объектами





Объекты, факты и не только

- Фразы-определения авторских терминов, их синонимов и связанных атрибутов: ***«Лексический анализ — это …»***
- Нормализация слабоструктурированных источников данных: автоматизированное формирование и коррекция номенклатурных списков (имущества, оборудования и т.д.): ***«Квартира 2-х комнатная 80 кв. м. …»***
- Прошивка законодательства: извлечение инструкций (связанных с обновлением текстов во времени) для их последующего применения: ***«Часть первую статьи 41 дополнить словами "или его заместителем"»***
- Графематический анализ: выявление в тексте простых лексических конструкций (ФИО с инициалами, электронные адреса, имена файлов), а также предложений, абзацев, заголовков, примечаний
- Выявление составных слов — напр.: ***для того чтобы***

Лексический анализ

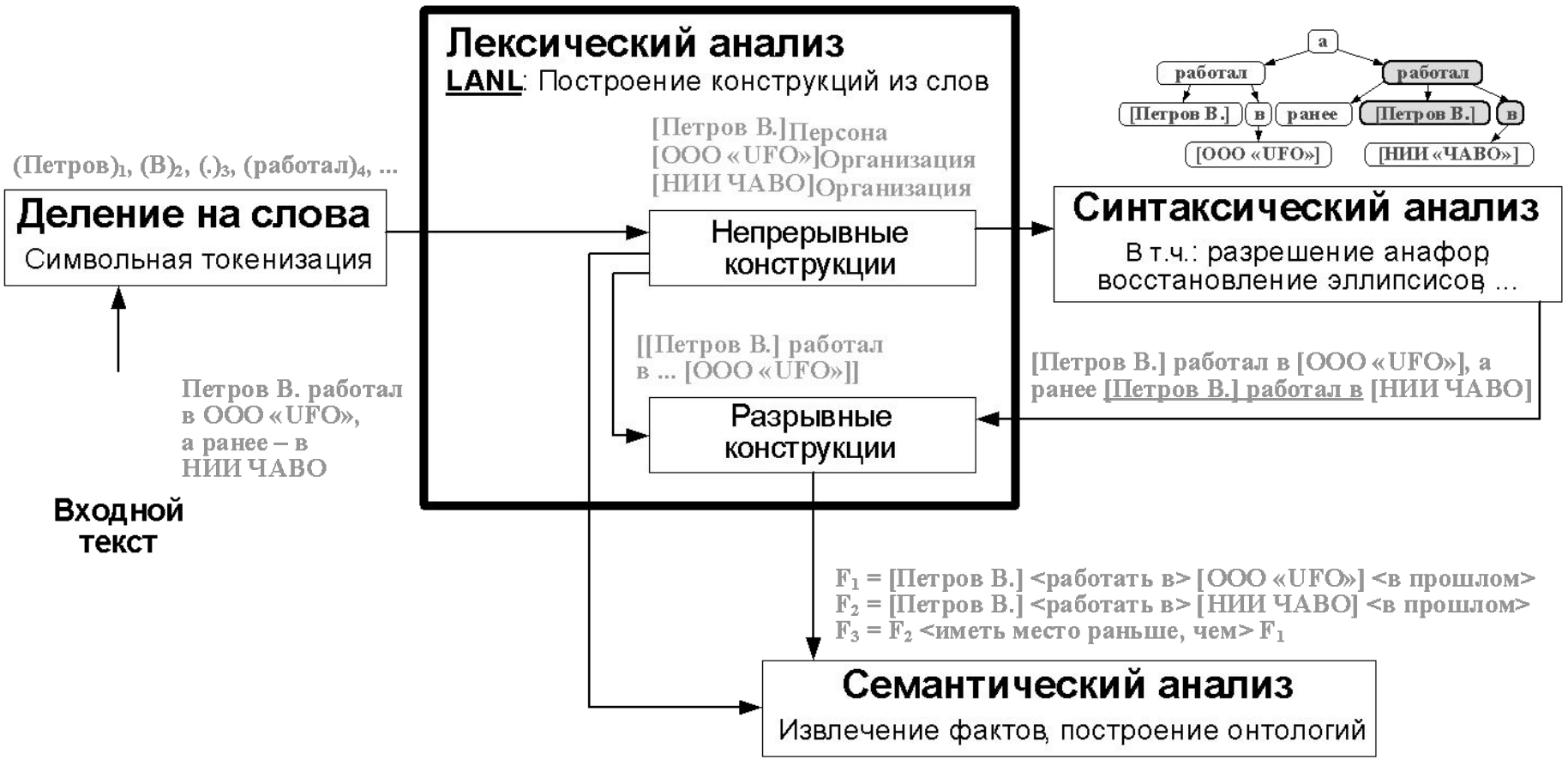
Задача: выявить в *неразмеченном* ЕЯ-тексте ***лексические конструкции*** — цепочки *слов* входного текста (*возможно, разрывные*), каждая из которых снабжается *набором данных* определенной структуры:

- ***имя класса***, которому принадлежит конструкция (**`Дата`**);
- ***нормальная форма конструкции***, которая состоит из *нормализованного текстового представления* (удобного для прочтения человеком) и набора именованных полей с присвоенными значениями (**`День = 26, Месяц = 2, Год = 2010`**)

Это **<u>лексический анализ естественного языка</u> (LANL):**

- Базовый механизм для выявления объектов (именованные сущности, как правило, являются непрерывными конструкциями)
- Вспомогательный механизм для выявления фактов (выявление утверждений — разрывных конструкций: «***Василий Петров***, мечтая о научной карьере, долгое время успешно ***трудился в НИИ ЧАВО***», м. быть установление кореференции объектов, но не логический вывод фактов)
- Вспомогательный механизм для деления текста на слова (поиск составных слов типа союзов, но не полноценная символьная токенизация — японский, арабский, *«первыйвторой»*)

Лексический анализ

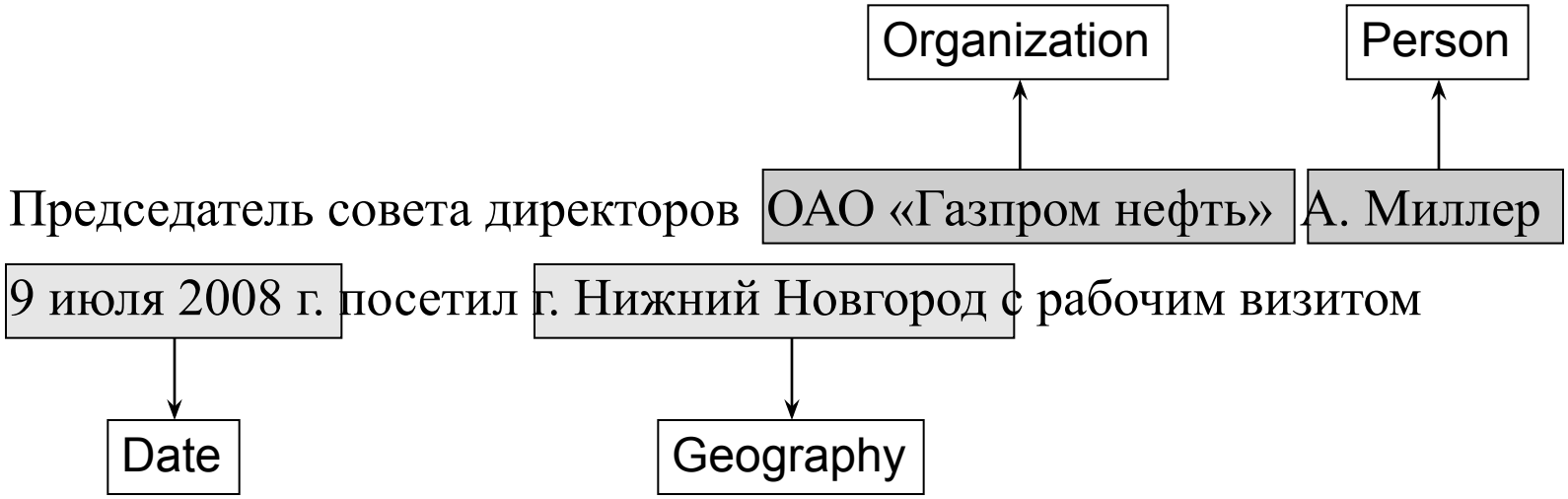


Принцип наследования

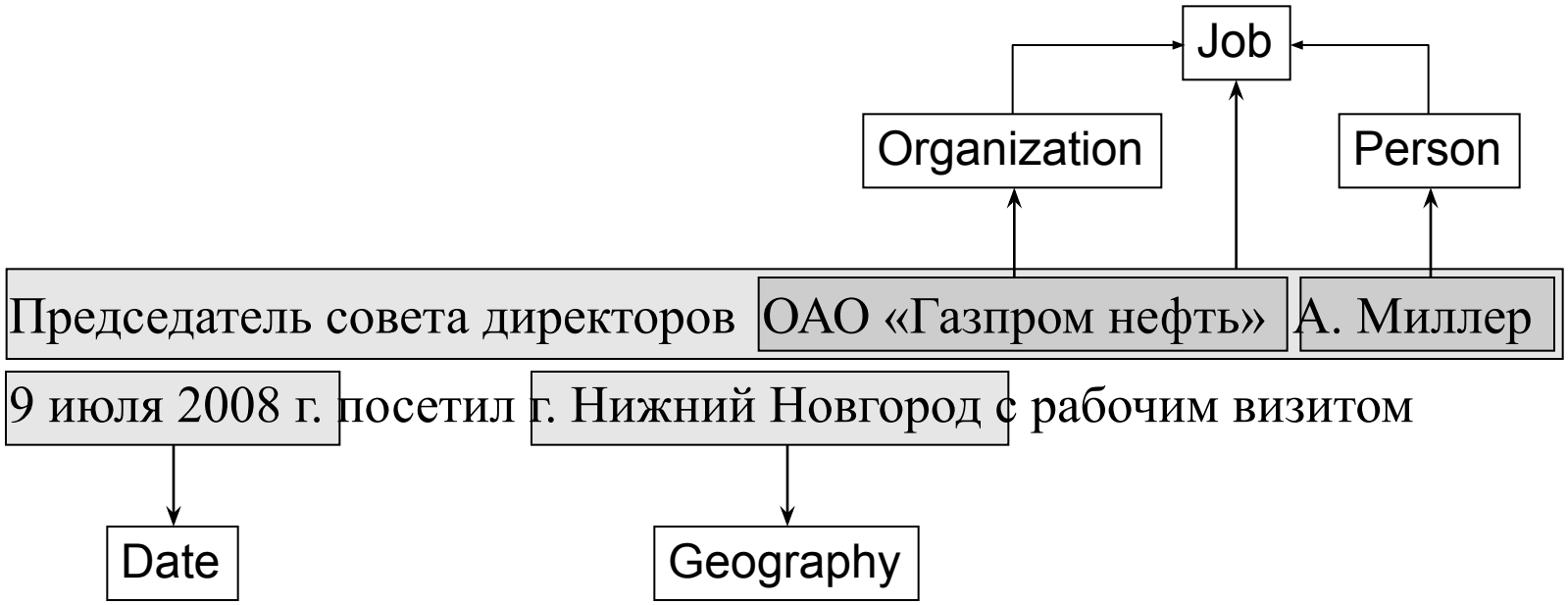
Председатель совета директоров ОАО «Газпром нефть» А. Миллер

9 июля 2008 г. посетил г. Нижний Новгород с рабочим визитом

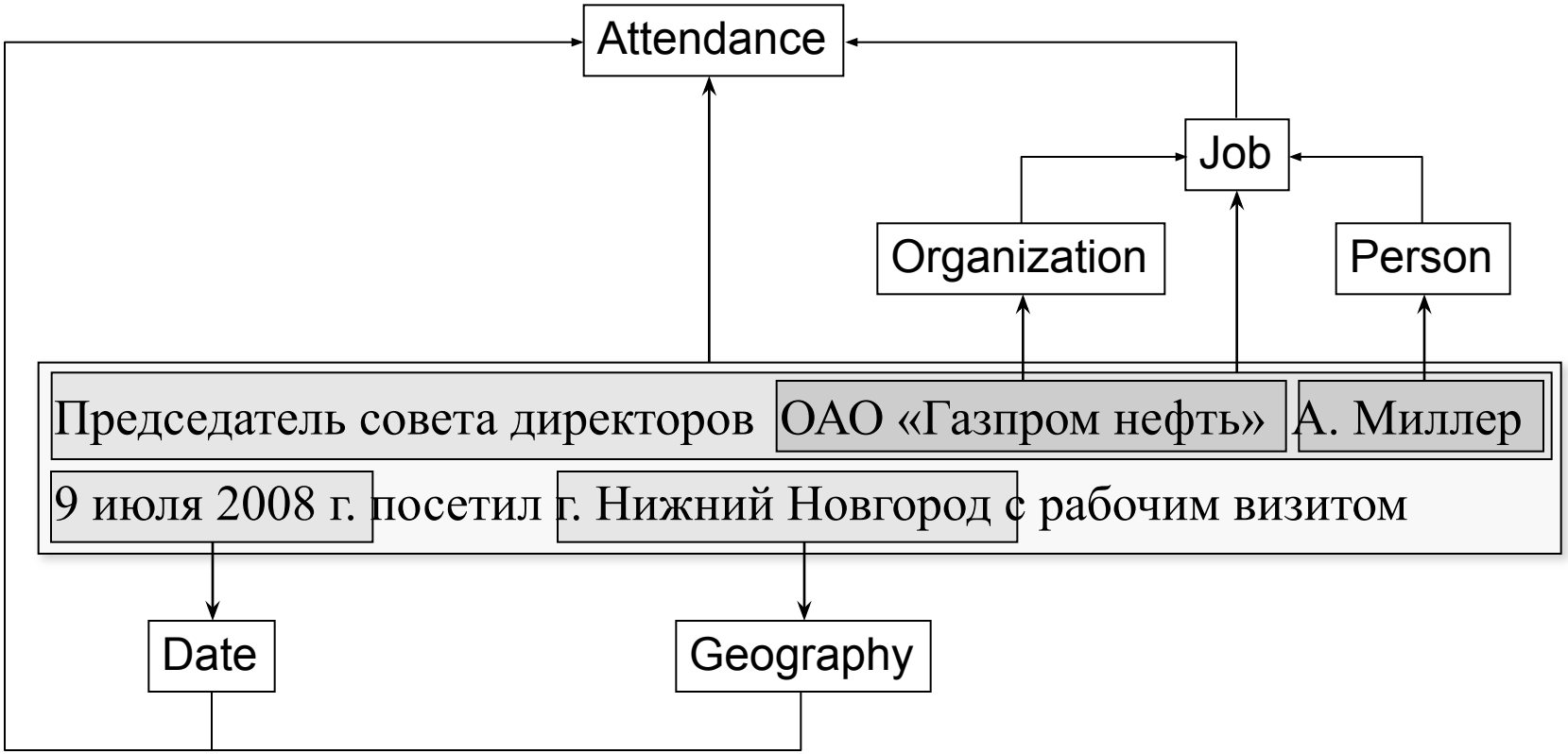
Принцип наследования



Принцип наследования

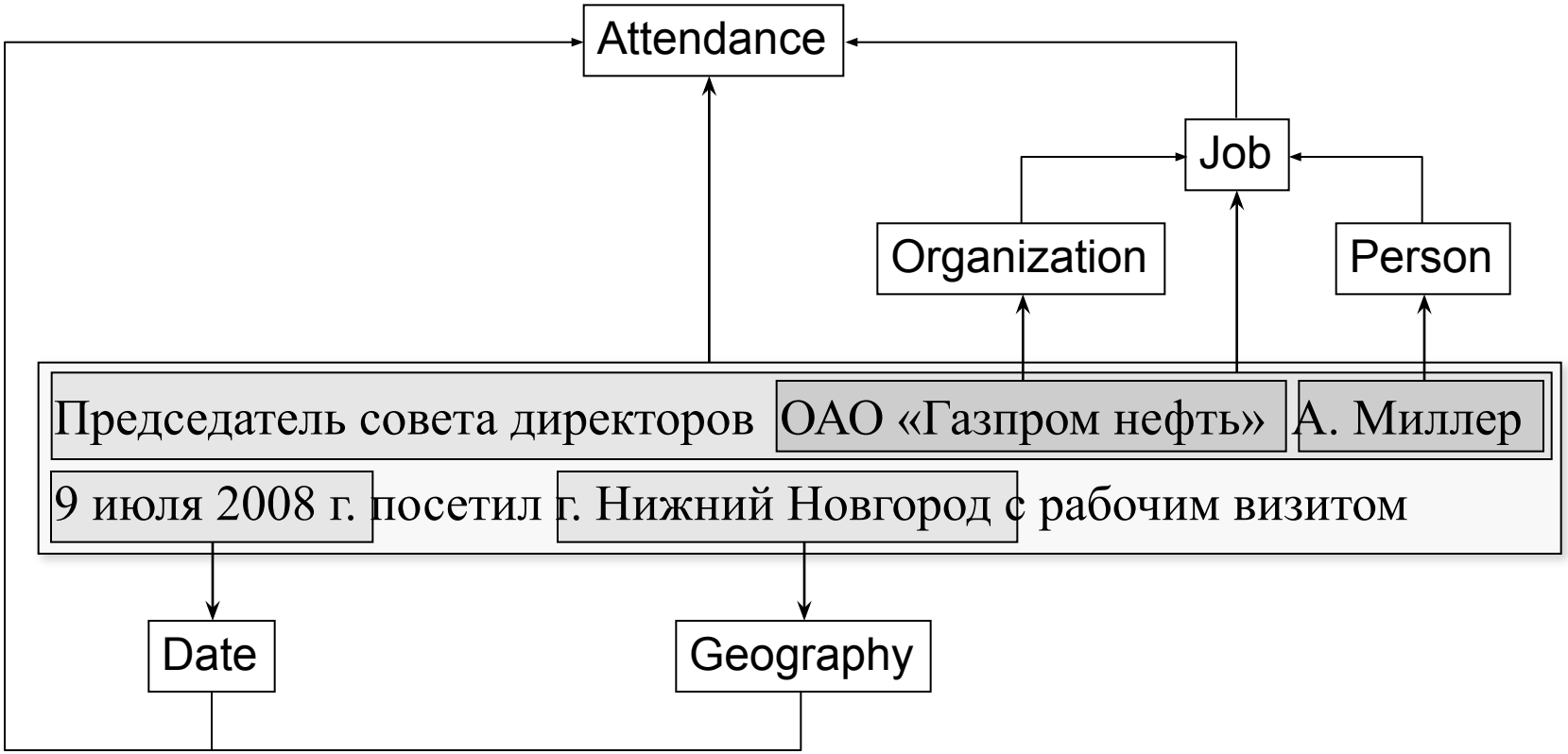


Принцип наследования



Регулярные выражения as is?

- Отсутствие механизмов повторного использования уже написанных выражений (наследования): можно лишь подставить одно выражение в другое



Регулярные выражения as is?

- Отсутствие механизмов повторного использования уже написанных выражений (наследования): можно лишь подставить одно выражение в другое

26/02/2010; ds@dictum.ru; 85 кг. — *Хорошо, но…*

Вчера **заместителю управляющего делами президента Российской Федерации Павлу Бородину** … — ***?!***

Регулярные выражения as is?

- Отсутствие механизмов повторного использования уже написанных выражений (наследования): можно лишь подставить одно выражение в другое
- Отсутствие специфических возможностей: проверка вхождения слов и их цепочек в заданные множества, работа с грамматическими значениями слова…

26/02/2010; ds@dictum.ru; 85 кг. — *Хорошо, но…*

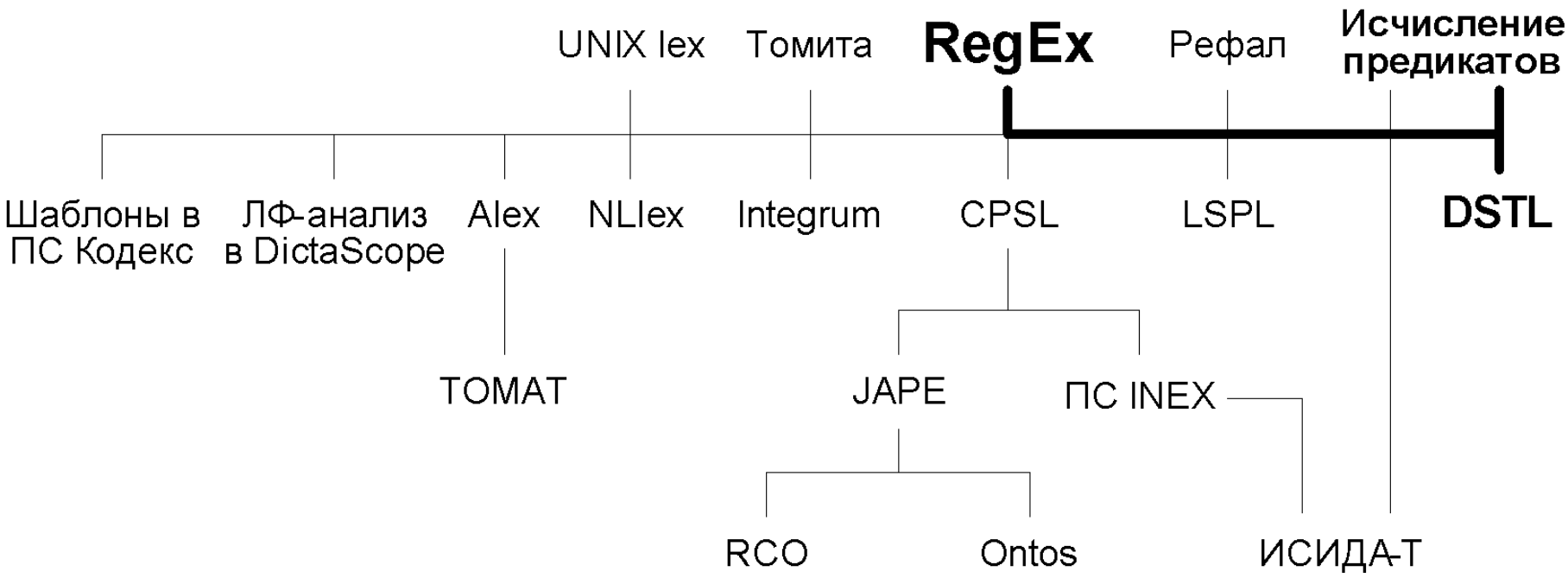
Вчера **заместителю управляющего делами президента Российской Федерации Павлу Бородину** … — ***?!***

Регулярные выражения as is?

- Отсутствие механизмов повторного использования уже написанных выражений (наследования): можно лишь подставить одно выражение в другое
- Отсутствие специфических возможностей: проверка вхождения слов и их цепочек в заданные множества, работа с грамматическими значениями слова…
- Быстрый рост сложности выражений (для их составителя)
- Нетривиальная обработка разделителей (переносы строк, пробелы) и их сочетаний
- Увеличение времени анализа с ростом количества описаний: каждое описание (регулярное выражение) приходится применять к тексту отдельно

Машинное обучение? □ Об этом позже

История



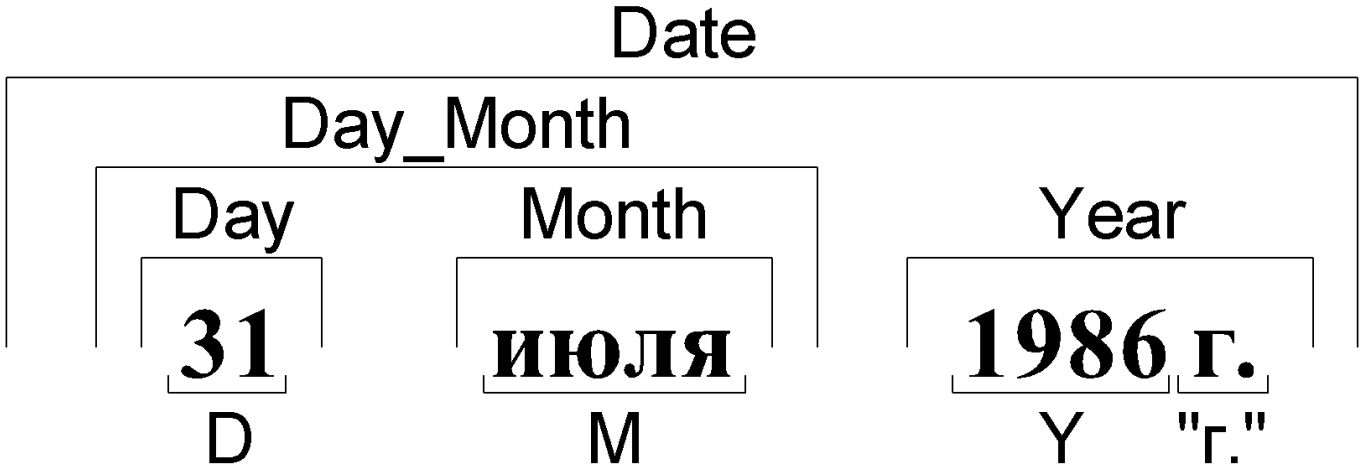
DSTL = Шаблоны + Наследование + Предикаты

DSTL: простой пример

```
Year { /* 1986 г. { year = 1986; } */
   T := Y "г." ?;
   C := Length (Y) = 4 & IsNumeric (Y);
   A := { year := Y; };
};
```

Наследование

```
Year { /* 1986 г. { year = 1986; } */
   T := Y "г." ?;
   C := Length (Y) = 4 & IsNumeric (Y);
   A := { year := Y; };
};
```



Наследование

```
Months := { "января": 1, … ,"декабря": 12 };

Day {
   T := D; /* 31 {day: 31} */
   C := IsNumeric (D) & DiapStr (D, 1, 31);
   A := { day := StrToInt (D); };
};

Month {
   T := M; /* июль, июля {month: 7} */
   C := M in Months; /* Months["июля"] = 7 */
   A := { month := Months[M]; };
};

/* 31 июля {day: 31, month: 7} */
Day_Month { T := [Day] [Month]; };
/* 31 июля 1986 г. {day: 31, month: 7, year: 1986} */
Date { T := [Day_Month] [Year]; };
```

Работа с морфологией

механика

{«механик», Сущ, Од, Муж, Род, Ед}
{«механик», Сущ, Од, Муж, Вин, Ед}
{«механика», Сущ, Неодуш, Жен, Им, Ед}

Работа с морфологией

v механика

{«механик», Сущ, Од, Муж, Род, Ед}
{«механик», Сущ, Од, Муж, Вин, Ед}
{«механика», Сущ, Неодуш, Жен, Им, Ед}

1. **Одноместные функции**: проверка существования грамматической формы с заданными характеристиками

```
HasGrammarForm (V, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

Работа с морфологией

v механика

{«механик», Сущ, Од, Муж, Род, Ед}
{«механик», Сущ, Од, Муж, Вин, Ед}
{«механика», Сущ, Неодуш, Жен, Им, Ед}

1. **Одноместные функции**: проверка существования грамматической формы с заданными характеристиками

```
HasGrammarForm (V, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

Работа с морфологией

механика	Александра
{«механик», Сущ, Од, Муж, Род, Ед}	*{«Александр», Сущ, Имя, Муж, Род, Ед}*
{«механик», Сущ, Од, Муж, Вин, Ед}	*{«Александр», Сущ, Имя, Муж, Вин, Ед}*
{«механика», Сущ, Неодуш, Жен, Им, Ед}	*{«Александра», Сущ, Имя, Жен, Им, Ед}*

1. **Одноместные функции**: проверка существования грамматической формы с заданными характеристиками

```
HasGrammarForm (V, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

Работа с морфологией

V механика

{«механик», Сущ, Од, Муж, Род, Ед}
{«механик», Сущ, Од, Муж, Вин, Ед}
{«механика», Сущ, Неодуш, Жен, Им, Ед}

Александра **W**

{«Александр», Сущ, Имя, Муж, Род, Ед}
{«Александр», Сущ, Имя, Муж, Вин, Ед}
{«Александра», Сущ, Имя, Жен, Им, Ед}

1. **Одноместные функции**: проверка существования грамматической формы с заданными характеристиками

```
HasGrammarForm (V, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

2. **Двуместные функции**: (1) из первого и второго слова выбираются подмножества S_1 и S_2 грамматических форм с заданными характеристиками, (2) проверяется, существует ли пара (v_1,v_2) такая, что $v_1 \in S_1$, $v_2 \in S_2$, и обе формы имеют требуемый набор характеристик с попарно совпадающими значениями

```
AreConcordant (Case_, Number_,
   V, PartOfSpeech_, Noun_, Gender_,
Masc_,
   W, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

Работа с морфологией

V механика

{«механик», Сущ, Од, Муж, Род, Ед}
{«механик», Сущ, Од, Муж, Вин, Ед}
{«механика», Сущ, Неодуш, Жен, Им, Ед}

Александра **W**

{«Александр», Сущ, Имя, Муж, Род, Ед}
{«Александр», Сущ, Имя, Муж, Вин, Ед}
{«Александра», Сущ, Имя, Жен, Им, Ед}

1. **Одноместные функции**: проверка существования грамматической формы с заданными характеристиками

```
HasGrammarForm (V, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

2. **Двуместные функции**: (1) из первого и второго слова выбираются подмножества S_1 и S_2 грамматических форм с заданными характеристиками, (2) проверяется, существует ли пара (v_1,v_2) такая, что $v_1 \in S_1$, $v_2 \in S_2$, и обе формы имеют требуемый набор характеристик с попарно совпадающими значениями

```
AreConcordant (Case_, Number_,
    V, PartOfSpeech_, Noun_, Gender_,
Masc_,
    W, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

Работа с морфологией

V механика Александра **W**

{«механик», Сущ, Од, Муж, Род, Ед} ⟷ *{«Александр», Сущ, Имя, Муж, Род, Ед}*

{«механик», Сущ, Од, Муж, Вин, Ед} ⟷ *{«Александр», Сущ, Имя, Муж, Вин, Ед}*

{«механика», Сущ, Неодуш, Жен, Им, Ед} *{«Александра», Сущ, Имя, Жен, Им, Ед}*

1. **Одноместные функции**: проверка существования грамматической формы с заданными характеристиками

```
HasGrammarForm (V, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

2. **Двуместные функции**: (1) из первого и второго слова выбираются подмножества S_1 и S_2 грамматических форм с заданными характеристиками, (2) проверяется, существует ли пара (v_1,v_2) такая, что $v_1 \in S_1$, $v_2 \in S_2$, и обе формы имеют требуемый набор характеристик с попарно совпадающими значениями

```
AreConcordant (Case_, Number_,
    V, PartOfSpeech_, Noun_, Gender_,
Masc_,
    W, PartOfSpeech_, Noun_, Gender_,
Masc_)
```

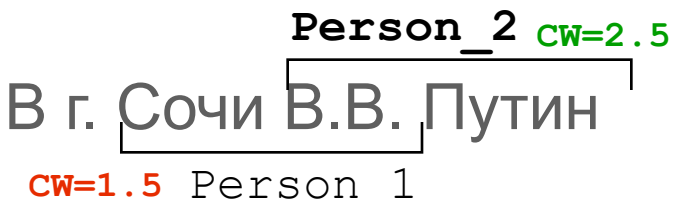
Согласование и нормальная форма

```
N {
   T := W; /* Иван, Петру, Сергеем */
   C := HasGrammarForm (W, {Subtype: Name, Gender: Masc});
   A := { GrV := W.GrV; W := GetInitialForm (W); };
};


...


N_Sn { T := [N] [Sn]; /* Иванам Петровым */
   C := AreConcordant (N, Sn, {Gender, Number, Case}); };
Sn_N { T := [Sn] [N]; /* Петрову Ивану */
   C := AreConcordant (N, Sn, {Gender, Number, Case}); };
N_Pt_Sn { T := [N] [Pt] [Sn]; /* Ивана Михайловича Петрова */
   C := AreConcordant (N, Pt, Sn, {Gender, Number, Case}); };
Sn_N_Pt { T := [Sn] [N] [Pt]; /* Петровым Иваном Михайловичем */
   C := AreConcordant (Sn, N, Pt, {Gender, Number, Case}); };
```

Неоднозначность и конфликты

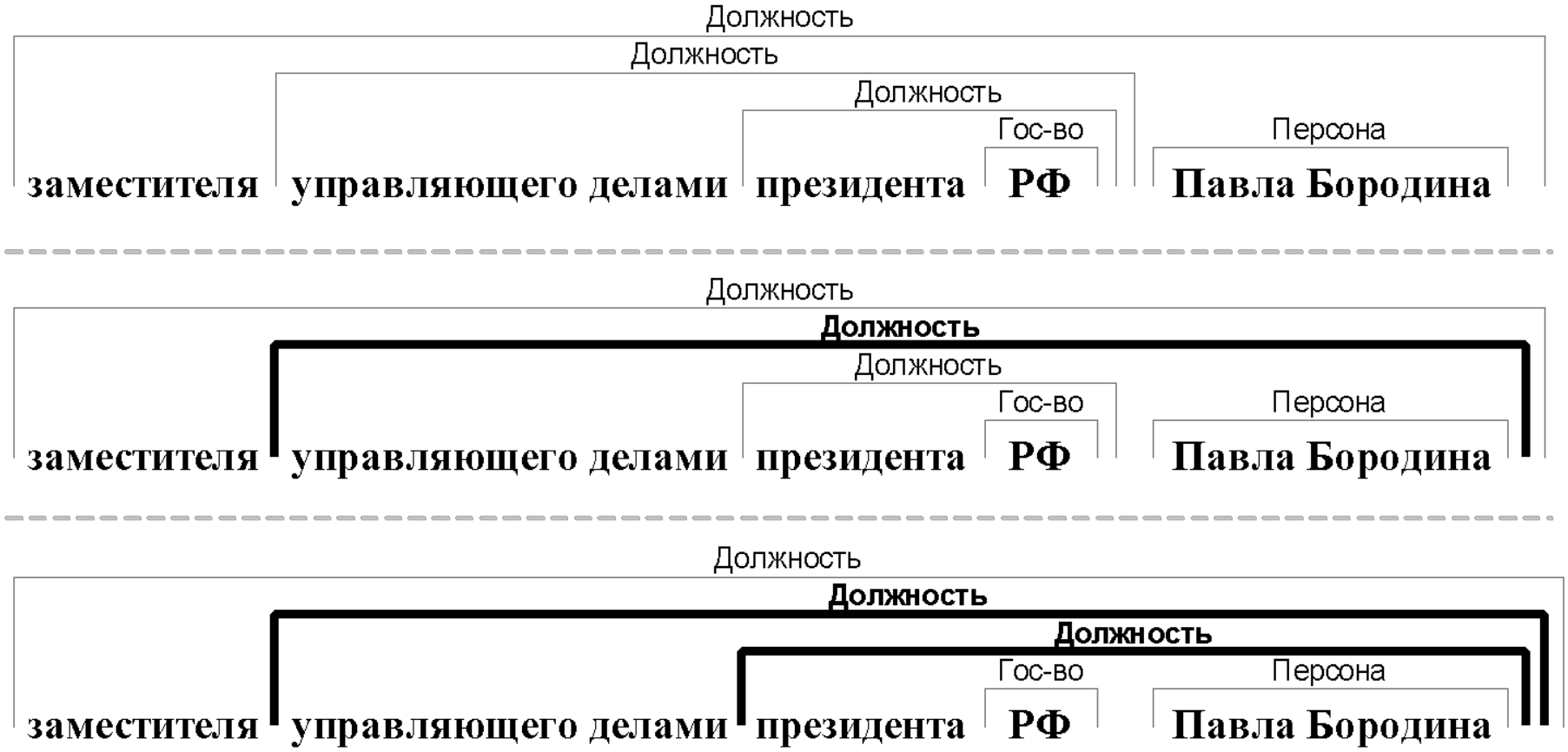
```
SN {
    T := SName;
    C := IsCapitalized (SName)
       & Length (SName) >= 2;
    A := { CW := 1 - (IsVoc (SName)
       & !IsPOS (SName, Surname_));};
};
```


```
NP {
    T := N \. P \.;
    C := Length (N) = 1
       & Length (P) = 1;
    A := { CW := 1; };
};
```




```
Person_1 { T := [SN][NP]; A := {CW := NP.CW + SN.CW;
};};};
Person_2 { T := [NP][SN]; A := {CW := NP.CW + SN.CW +
0.5;};};};
```

Неоднозначность и конфликты



Сравнение языков

CPSL

```
Macro: NOT_QUOTE ({!Token.string == "\""})
Rule: NewspaperName
   ({Token.string =| "газета"} | {Token.string =| "журнал"})
   {Token.string == "\""}
   (({!Token.string == "\"", Morpho.Capitalized == True})
    NOT_QUOTE? NOT_QUOTE? NOT_QUOTE?)
    : newspaperName {Token.string == "\""}
--> :newspaperName.ProperName = {
   kind = "Newspaper", rule = "NewspaperName"}
```

DSTL

```
QUOTE := "\"";

Name : hidden {
   T := First (Other) {0,3};
   C := IsCapitalized (First) &
        First != QUOTE & Other != QUOTE;
};

Newspaper {
   T := Pr QUOTE [Name] QUOTE;
   C := Pr %in {"газета", "журнал"};
   A := { kind := "Newspaper"; newspaperName := Name; };
};
```

Механизм анализа

$A = \{a_1, \ldots, a_n\} = \{1, 2, 3\};\ X \subseteq A$

$T = X_1\ X_2 \ldots\ X_L = \{1, 2\}\ \{1, 2, 3\}\ \{2, 3\}\ \{3\}$

$P = p_1\ p_2 \ldots\ p_m = \{1\}\ \{2, 3\} = p_1\ p_2;\ \overline{P} = \{P_s\}_{s=1}^{M}$

Задача: *найти все вхождения образцов из* $\overline{P}$ *в* T

$$T = \{1, 2\}\ \{1, 2, 3\}\ \{2, 3\}\ \{3\}$$

Механизм анализа

$A = \{a_1, \ldots, a_n\} = \{1,2,3\};\ X \subseteq A$

$T = X_1\ X_2 \ldots\ X_L = \{1,2\}\ \{1,2,3\}\ \{2,3\}\ \{3\}$

$P = p_1\ p_2 \ldots\ p_m = \{1\}\ \{2,3\} = p_1\ p_2;\ \overline{P} = \{P_s\}_{s=1}^{M}$

Задача: *найти все вхождения образцов из* $\overline{P}$ *в* T

p_1

$$T = \{1,2\}\ \{1,2,3\}\ \{2,3\}\ \{3\}$$

Механизм анализа

$A = \{a_1, \ldots, a_n\} = \{1,2,3\};\ X \subseteq A$

$T = X_1\ X_2 \ldots\ X_L = \{1,2\}\ \{1,2,3\}\ \{2,3\}\ \{3\}$

$P = p_1\ p_2 \ldots\ p_m = \{1\}\ \{2,3\} = p_1\ p_2;\ \overline{P} = \{P_s\}_{s=1}^{M}$

Задача: *найти все вхождения образцов из* $\overline{P}$ *в* T

p_1 p_2

$$T = \{1,2\}\ \{1,2,3\}\ \{2,3\}\ \{3\}$$

Механизм анализа

$$A = \{a_1, \ldots, a_n\} = \{1,2,3\};\ X \subseteq A$$

$$T = X_1\ X_2 \ldots\ X_L = \{1,2\}\ \{1,2,3\}\ \{2,3\}\ \{3\}$$

$$P = p_1\ p_2 \ldots\ p_m = \{1\}\ \{2,3\} = p_1\ p_2;\ \overline{P} = \{P_s\}_{s=1}^{M}$$

Задача: *найти все вхождения образцов из* $\overline{P}$ *в* T

p_1 p_2

$$T = \{1,2\}\ \{1,2,3\}\ \{2,3\}\ \{3\}$$

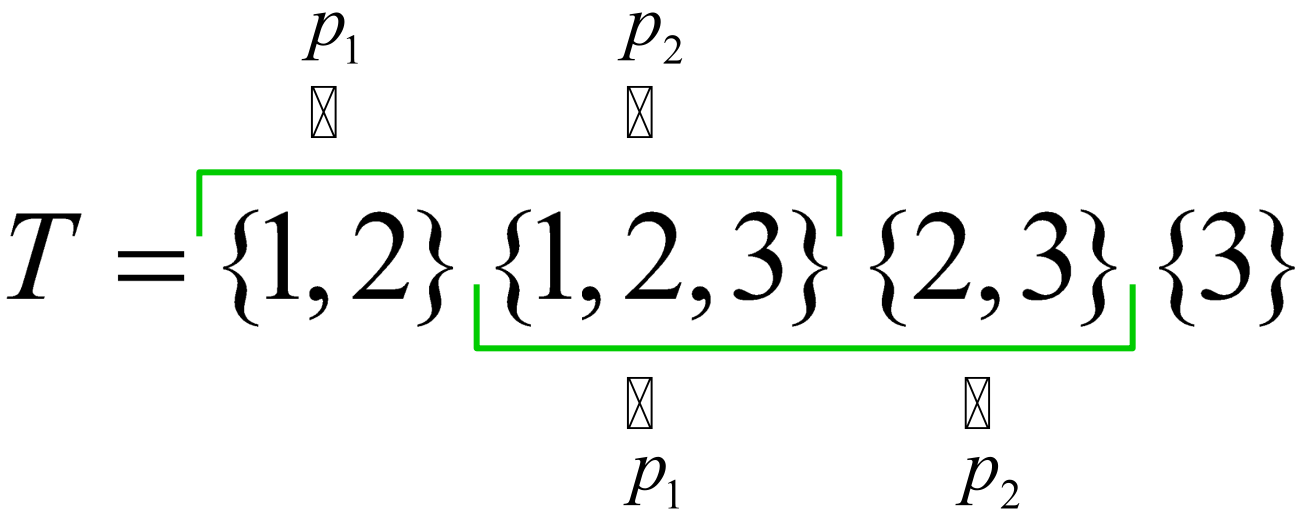
Механизм анализа

$A = \{a_1, \ldots, a_n\} = \{1, 2, 3\};\ X \subseteq A$

$T = X_1\ X_2 \ldots\ X_L = \{1, 2\}\ \{1, 2, 3\}\ \{2, 3\}\ \{3\}$

$P = p_1\ p_2 \ldots\ p_m = \{1\}\ \{2, 3\} = p_1\ p_2;\ \overline{P} = \{P_s\}_{s=1}^{M}$

Задача: *найти все вхождения образцов из* $\overline{P}$ *в* T



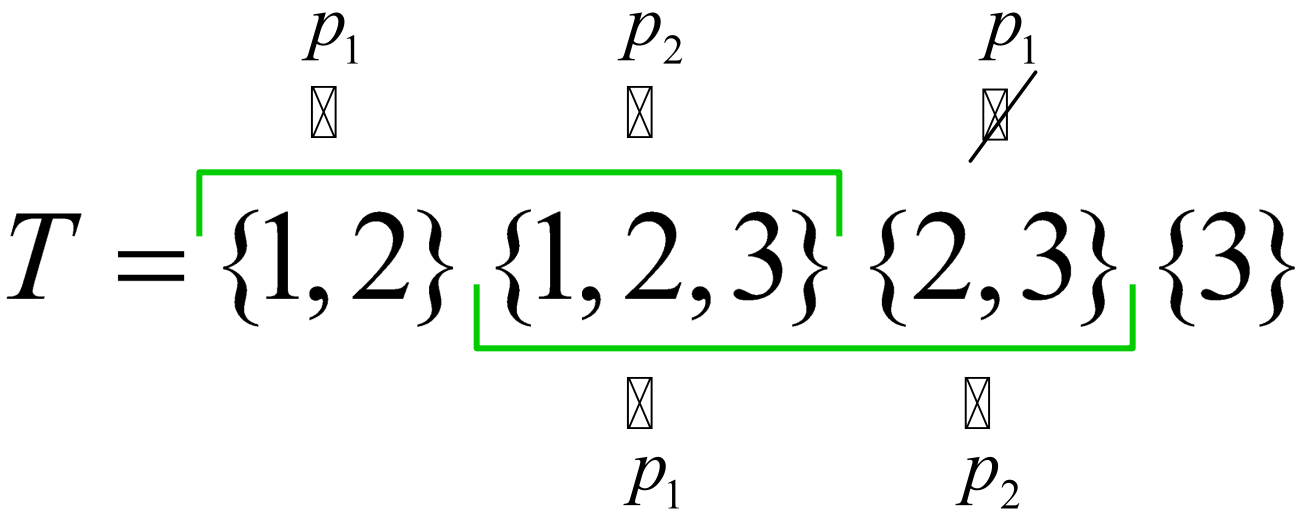
Механизм анализа

$A = \{a_1, \ldots, a_n\} = \{1, 2, 3\};\ X \subseteq A$

$T = X_1\ X_2 \ldots\ X_L = \{1, 2\}\ \{1, 2, 3\}\ \{2, 3\}\ \{3\}$

$P = p_1\ p_2 \ldots\ p_m = \{1\}\ \{2, 3\} = p_1\ p_2;\ \overline{P} = \{P_s\}_{s=1}^{M}$

Задача: *найти все вхождения образцов из* $\overline{P}$ *в* T



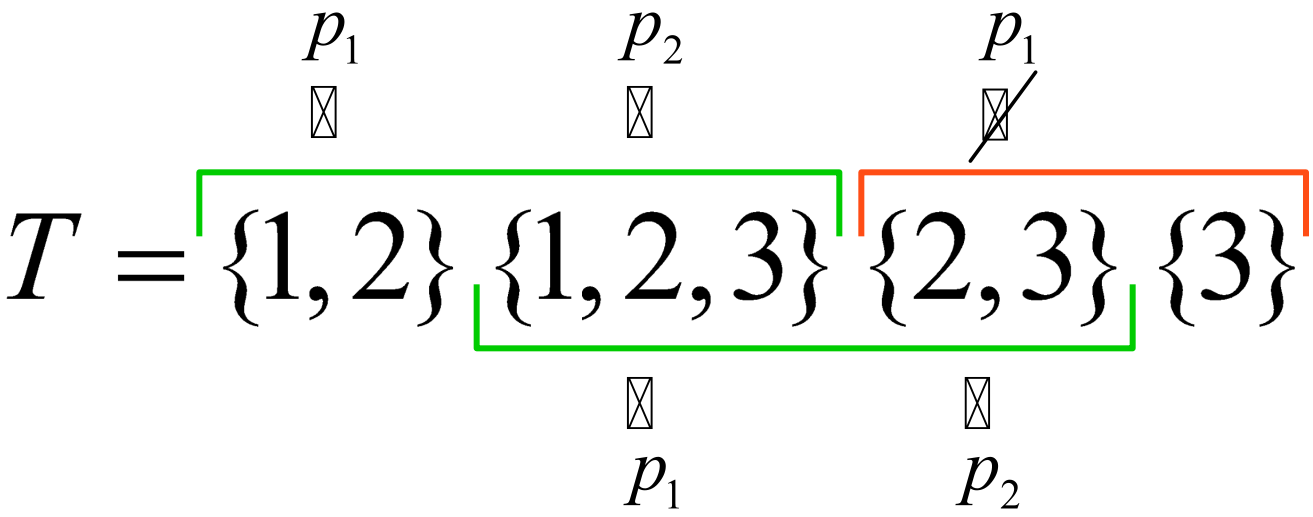
Механизм анализа

$A = \{a_1, \ldots, a_n\} = \{1, 2, 3\};\ X \subseteq A$

$T = X_1\ X_2 \ldots\ X_L = \{1, 2\}\ \{1, 2, 3\}\ \{2, 3\}\ \{3\}$

$P = p_1\ p_2 \ldots\ p_m = \{1\}\ \{2, 3\} = p_1\ p_2;\ \overline{P} = \{P_s\}_{s=1}^{M}$

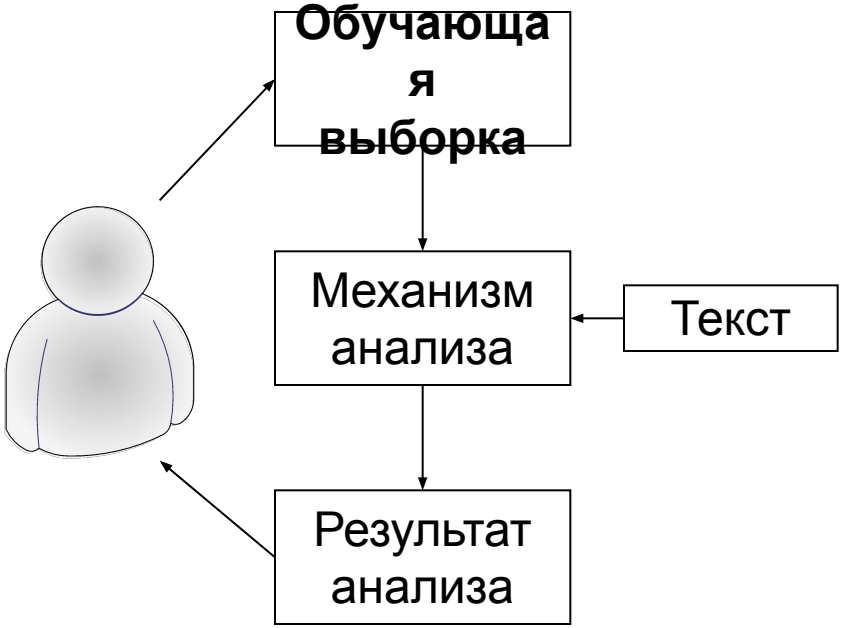
Задача: *найти все вхождения образцов из* $\overline{P}$ *в* T



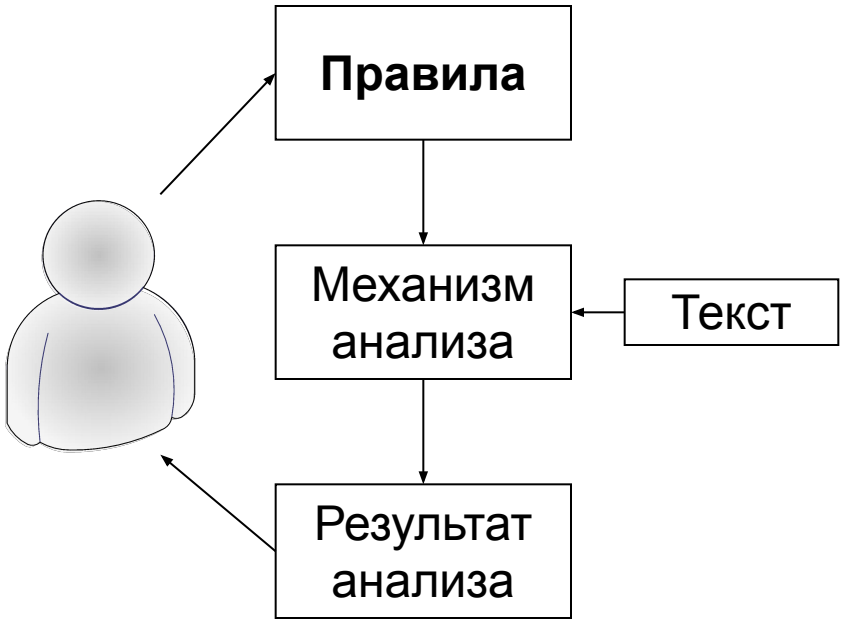
Проблемы и решения

- Правила, составляемые экспертом, дают лучший результат в сравнении с результатом применения машинного обучения (обучение с учителем, распознавание образов …)
- **Проблема:** высокая трудоемкость работы эксперта
- Машинное обучение:
  - Хорошо применимо для распознавания узких классов (напр., в Named Entities Recognition — имена людей — популярно у зарубежных исследователей)
  - Позволяет *распознать* текстовый фрагмент и приписать класс, но не заполнить поля или отразить структуру наследования (следствие — трудность разрешения конфликтов)
  - Обучение — возможно, не менее трудоемко, чем составление правил, и результат иногда недетерминирован для учителя
  - Неполнота обучающей выборки
- **Возможное решение:** возьмем лучшее из обоих подходов

Проблемы и решения

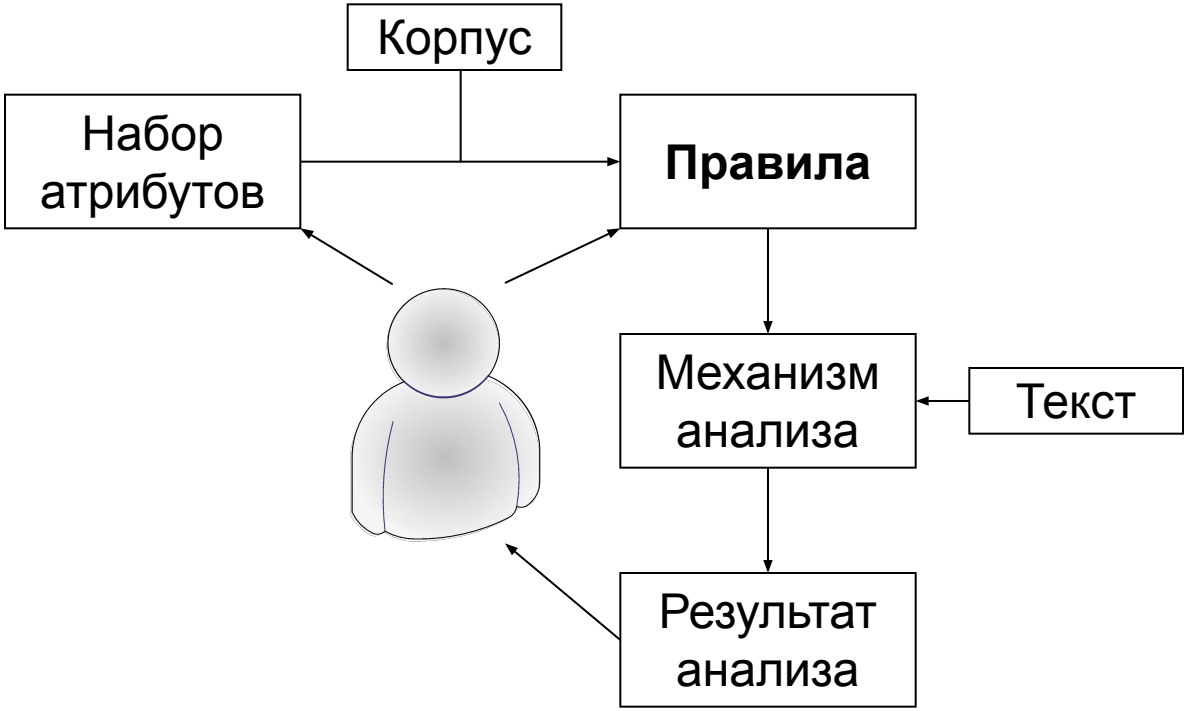


Машинное обучение



Система правил

Проблемы и решения



- Механизм анализа не меняется
- Эксперт формирует набор атрибутов, система выявляет в текстах корпуса устойчивые сочетания

Использование для поиска

Запрос: «**февраль 2010**»

Вхождения образца из запроса в текст

Образцы	День	Месяц	Год
февраль 2010	∅	2	2010
4 февраля 2010	4	2	2010
25-02-2010	25	2	2010
25.02.10	25	2	10
2010 год	∅	∅	2010
Feb 6, 2010	6	2	2010
06-03-2009	6	3	2009

Использование для поиска

- **Проблема** — сравнение объектов сложнее сравнения слов:
  - **Частичное совпадение** (*«2 февраля 2010»* и *«февраль 2010»*)
  - **Частичное несовпадение** (*«февраль 2010»* и *«февраль 2009»*)
  - **«Семантическая» близость** (*«3 февраля 2010»* и *«4 февраля 2010»* ближе, чем *«3 февраля 2010»* и *«3 февраля 2009»*)
- Пусть вместе с базой правил определена функция ***d*(*x*,*y*)**:
  - $d(x,y) = 0$ для одинаковых объектов
  - $d(x,y) = \infty$ для объектов разных классов
  - Частичное совпадение «лучше» частичного несовпадения
- **Решение** — степень схожести вместо булевского равенства:

$$W(x,y) = \frac{1}{1 + d(x,y)}$$

Контакты

Адрес:

603950 Россия, Нижний Новгород,
Проспект Гагарина 23, корпус 7

Тел (факс): +7 (831) 278-67-57

e-mail: ds@dictum.ru

web: www.dictum.ru