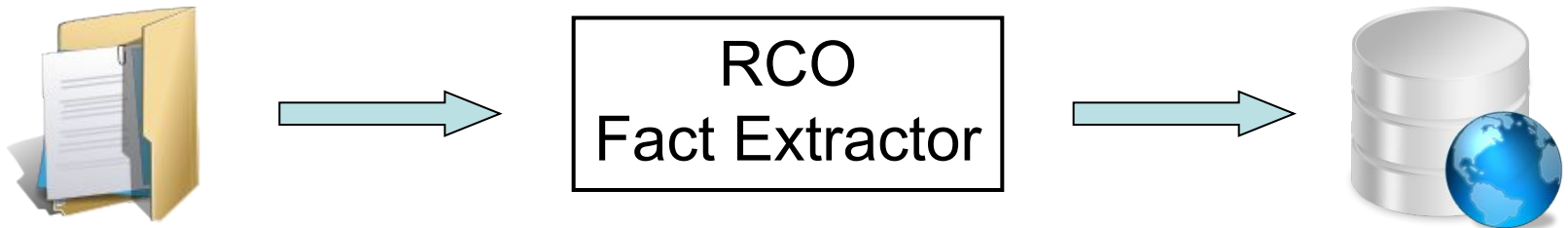


RCO Fact Extractor SDK

Основные этапы обработки текста

Назначение:

выделение из текста структурированной информации на основе правил и шаблонов.



На выходе:

- различные классы сущностей, упомянутых в тексте: персоны, организации, география, предметы, действия, атрибуты и др.;
- сеть отношений, связывающих эти сущности;
- грамматическая информация о составляющих текста;
- семантическая интерпретация результатов разбора – поиск описаний ситуаций, удовлетворяющих заданным семантическим шаблонам.

Этапы обработки текста

1. Токенизация
2. Газеттер
3. Морфологический анализ
4. Семантический словарь
5. Кейп (CAPE – C Annotation Patterns Engine)
6. Модуль выделения именованных объектов, в том числе predetermined пользователем объектов
7. Синтаксический анализ
8. Поиск фактов
9. Разбор таблиц

Токенизация

- Кодировка 1251 и 1252
- Форматы html и текст
- Категории токенизации:
 - текстовый блок
 - абзац
 - предложение
 - слово (токен)
- Типы токенов:
 - знак препинания
 - русское слово
 - латинское слово
 - специальная конструкция

Газеттер

Осуществляет поиск слов и словосочетаний с учётом словоформ. Найденным терминам присваиваются указанные в словаре атрибуты. При обнаружении многословного термина, его слова «склеиваются». При пересечении цепочек слов-кандидатов на склеивание вычисляется «оптимальное» покрытие текста цепочками.

Фрагмент словаря должностей:

Person:Position

главный MAIN врач MSYN главврач

главврач MSYN

генеральный MAIN директор MSYN гендиректор

гендиректор MSYN

Примеры из словарей газеттера

- Можно указывать грамматические значения для неизвестных слов:

**врио SYN {SpeechPartDetailed="NounAnimateM",WordBase="ВРИП",
Case="Any",Number="Singular",Person="Third",Gender="Masculine"}**

- Можно указывать все словоформы для неизвестных нестандартно склоняющихся слов:

Event:Event

**pr-кампания SYN
{SpeechPartDetailed="NounF",Case="Nominative",Number="Singular",
Person="Third",Gender="Feminine"}**

**pr-кампании SYN
{SpeechPartDetailed="NounF",Case="Generative",Number="Singular",
Person="Third",Gender="Feminine"}**

**pr-кампании SYN
{SpeechPartDetailed="NounF",Case="Dative",Number="Singular",
Person="Third",Gender="Feminine"}....**

Морфологический анализ

Определение грамматических характеристик слова (часть речи, падеж, число, род, лицо и т.д.)

В основном словаре:

- 110 тыс. слов (52 тыс. существительных, 24 тыс. глаголов, 33 тыс. прилагательных, остальное – наречия, служебные, наименования, имена, фамилии, география)
- 743 приставки для правил точного анализа неизвестных слов
- 162 окончания для правил точного анализа неизвестных слов

В дополнительном словаре: 27 тыс. фамилий и 23 тыс. имён.

Неизвестные слова анализируются в приближенной морфологии по правилам на известные приставки/окончания и на основе частоты суффиксов и окончаний известных слов.

Семантический словарь

Навешивает на сущности текста семантические категории и определяет принадлежность к семантическому ряду.

Основные категории:

- контекст места (дома, везде, далеко, здесь)
- контекст времени (весной, зачастую, завтра, когда-нибудь)
- предметные (деревня, надкус, покупатель)
- событийные (использовать, использование, инвестировать, инвестирование, укус)
- признаковые (сила, сильный, бодливость, бодучесть)
- одушевлённые/неодушевлённые (дядя/дуб)
- материальные/нематериальные (жаба/жадность)
- естественные/искусственные (залив/замок)
- имена собственные/нарицательные (Петя/мальчик)
- собирательные (множество, ряд, стог)
- обозначение части (вершина, край, половина)
- единицы измерения (неделя, тонна, март)

Примеры семантических рядов:

- КОРЫСТОЛЮБИЕ, ЗЛАТОЛЮБИЕ
- КОРОЛЬ, КОРОЛЕВНА, КОРОЛЕВИЧ, КОРОЛЕВА
- ЧЕРТ, ЧЕРТЯКА, ЧЕРТЯГА, ЧЕРТУШКА, ЧЕРТИХА, ЧЕРТИК, ЧЕРТЕНОК

Кейп (CAPE)

Выделение в тексте сущностей с помощью специальных правил и регулярных выражений. Правила написаны на специальном языке, который транслируется в конечный автомат.

Примеры сущностей:

- даты: 03.01.1981, с 1-го мая, вчера и сегодня, 22.02.2013г., 2012-2013гг.
- денежные суммы: 1р., 5 руб., 10 рублей 20 коп., 3\$
- номера телефонов: 916-123-45678, 8(495)-987-65, тел. 345-35-45
- адреса: г. Москва, луж Набережная, 6А; ул. Красина 24кв1
- ссылки на нормативно-правовые акты: пп.7 ч.3 КОАП от 03.03.2000г.

Вход – цепочка токенов/сущностей с набором атрибутов.

Правило – ограничения на атрибуты токенов/сущностей в цепочке.

Результат – объединение цепочки в новую сущность, изменение атрибутов сущностей в цепочке.

Rule: EMail_Rule

```
( ({Token.Text =~ "[0-9A-z\.\_\-]+@[0-9A-z\.\_\-]+"}):value ):EAddress
```

```
--> :EAddress.Token = { Type = "Word", SemanticType = "Special:Email", Rule = "EMail_Rule" },  
:EAddress.Cape = { Value = :value.Token.Text }
```


Примеры правил CAPE

Правила могут основываться на предыдущих правилах. В данном примере используется семантический тип, определяющийся правилами для выделения дат:

Rule: DateOfBirth_Rule2

```
( ( {Token.Text =^ "дата"} {Token.Text =^ "год"} ) {Token.Text =^ "рождения"}  
  ( {Token.Text == ":"} {Token.Text == "-"} ) ?  
  ( {Token.SemanticType == "Time:Date"} ) :value  
):DateOfBirth
```

```
--> :DateOfBirth.Token = { Type = "Word", SemanticType = "Special:DateOfBirth", Rule =  
"DateOfBirth_Rule2" },  
  :DateOfBirth.Cape = { Value = :value.Token.Text }
```

Есть возможность использовать макросы и фильтры, наследовать атрибуты:

Rule: MetroStationName_Rule

```
( ( METRO_KEY_FULL (QUOT)? {Token.Filter =< "Metro:Name"} (QUOT)? ) ):metro  
--> :metro.Token = { Type = "Word", SemanticType = "Geoplace:Metro", Text = :metro.Token.Text,  
Rule = "MetroStationName_Rule"},  
  :metro.Morph = { :morph_info.Morph }
```

Модуль выделения именованных объектов

- Выделяет имена персон, названия организаций и географические наименования по общим правилам, опираясь на морфологию и ключевые слова.

Примеры:

- Иванов А. М., Петра Сергеевича Капицы, г-н Кириенко, И. Крапивин
- АО «МММ», комбинат «Россельмаш», завод металлоконструкций им. Ленина
- г. Москва, Владимирская и Новгородская области

- Производит поиск референтных упоминаний объектов (Путин = президент РФ = глава России)

- Устанавливает кореферентность (Мы пошли к Иванову. Он рассказал всё.)

- «Схлопывает» упоминания одного и того же объекта в разных местах текста.

Примеры:

- Никита Сергеевич Хрущов поднялся на трибуну..... В своей речи Хрущов...
- Банк Уралсиб отчитался за год... Убытки банка составили...

- Идентифицирует объекты, описанные в формате XML.

Примеры XML-описаний объектов

Пример xml-описания для объекта «Путин», тип «персона»:

```
<object id="ПУТИН ВЛАДИМИР ВЛАДИМИРОВИЧ" type="person">
<fields>
<field name="gender">мужской</field>
<field name="last name" modify="yes">Путин</field>
<field name="first name" modify="yes">Владимир</field>
<field name="middle name" modify="yes">Владимирович</field>
</fields>
<desc>
<syn type="normal">преемник Ельцина</syn>
<syn type="context">российский президент</syn>
<syn type="context">наш президент</syn>
<syn type="context">президент Российской Федерации</syn>
<syn type="context">глава России</syn>
<syn type="context">президент РФ</syn>
<syn type="context">глава правительства</syn>
<syn type="context">премьер-министр</syn>
<syn type="context">премьер</syn>
</desc>
</object>
```

Примеры XML-описаний объектов

Пример xml-описания для объекта с типом «организация»:

```
<object id="ПРОМТОРГБАНК" type="organization">
<fields>
<field name="gender">мужской</field>
<field name="full name" modify="yes">Акционерный коммерческий Промышленно-
торговый банк</field>
</fields>
<desc>
<syn type="normal" case="any">АК Промторгбанк (ЗАО)</syn>
<syn type="normal" case="any">ЗАО "Акционерный коммерческий Промышленно-
торговый банк"</syn>
<syn type="normal" case="any">ЗАО АК Промторгбанк</syn>
<syn type="normal" case="any">ЗАО "АК Промторгбанк"</syn>
<syn type="normal">Промышленно-торговый банк</syn>
<syn type="normal">Промторгбанк</syn>
</desc>
</object>
```

Синтаксический анализ

Синтаксический разбор предложения в терминах дерева зависимостей. Установление синтактико-семантических связей между словами и их ролей (субъект, объект, предикат и т.д.).

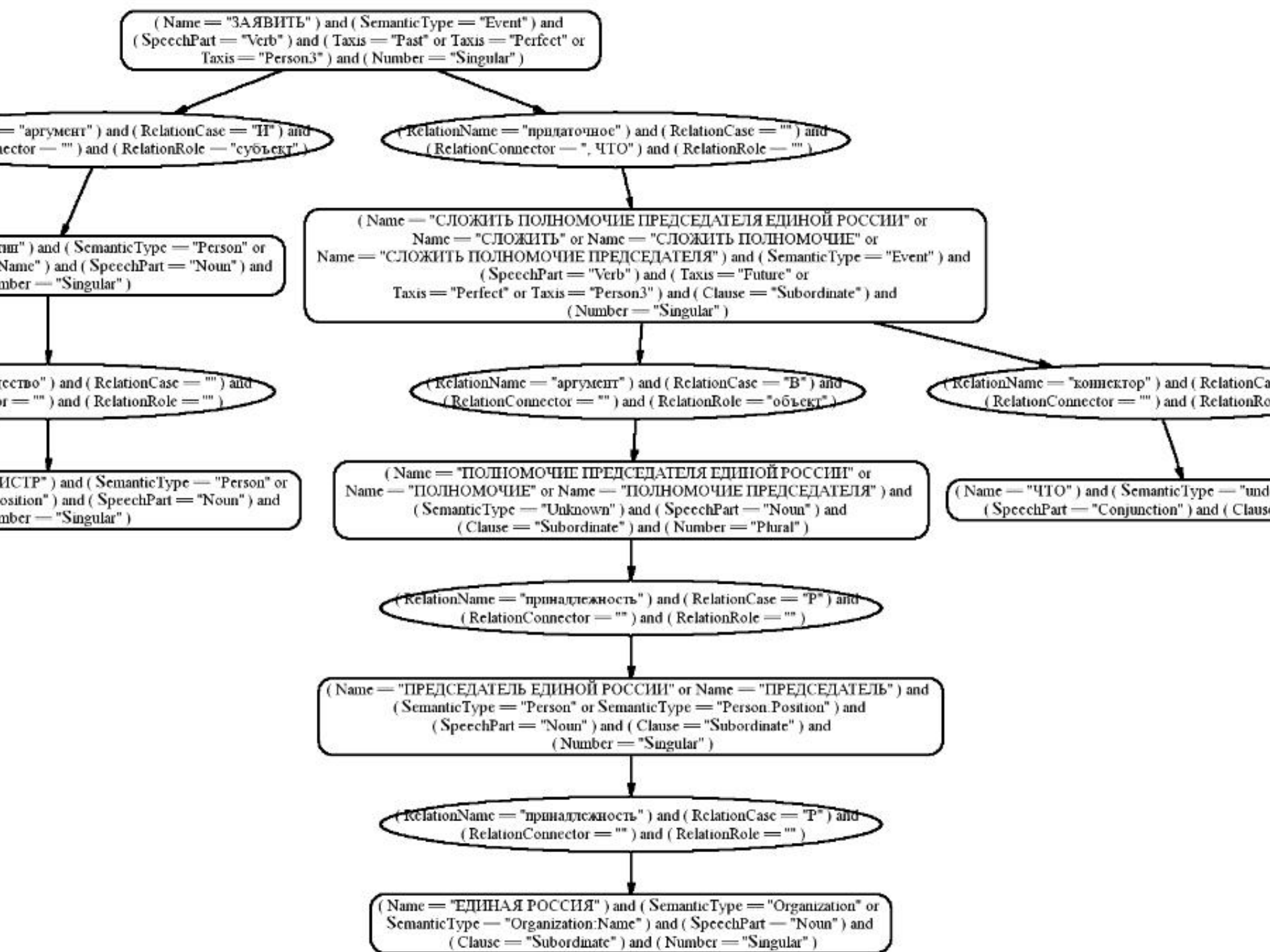
(НАПОМНИМ ,->(ВЧЕРА->_VP(_NP(_Position_:ПРЕМЬЕР-МИНИСТР<-ВЛАДИМИР ПУТИН<-:_Person_)<-_PP(НА ВСТРЕЧЕ<-_PP(С АКТИВОМ))->_VP(ПАРТИИ->ЗАЯВИЛ))))

_S'(, ЧТО<-_PP(ПОСЛЕ ИНАУГУРАЦИИ)->_PP(В КАЧЕСТВЕ ПРЕЗИДЕНТА)->_VP(СЛОЖИТ<-_NP(ПОЛНОМОЧИЯ<-_NP(ПРЕДСЕДАТЕЛЯ<- " ЕДИНОЙ РОССИИ "))))

Пример1: *Напомним, вчера премьер-министр Владимир Путин на встрече с активом партии заявил, что после инаугурации в качестве президента сложит полномочия председателя «Единой России».*

_VP(_NP(ОБЪЕМ<-_NP(ПРОДАЖ<-ХОЛДИНГА OZON))->_hVP(_(ВЫРОС<-НА 78%) И _(СОСТАВИЛ<-8,8 МЛРД РУБ.)))

Пример2: *Объем продаж холдинга Ozon вырос на 78% и составил 8,8 млрд руб.*



Поиск фактов

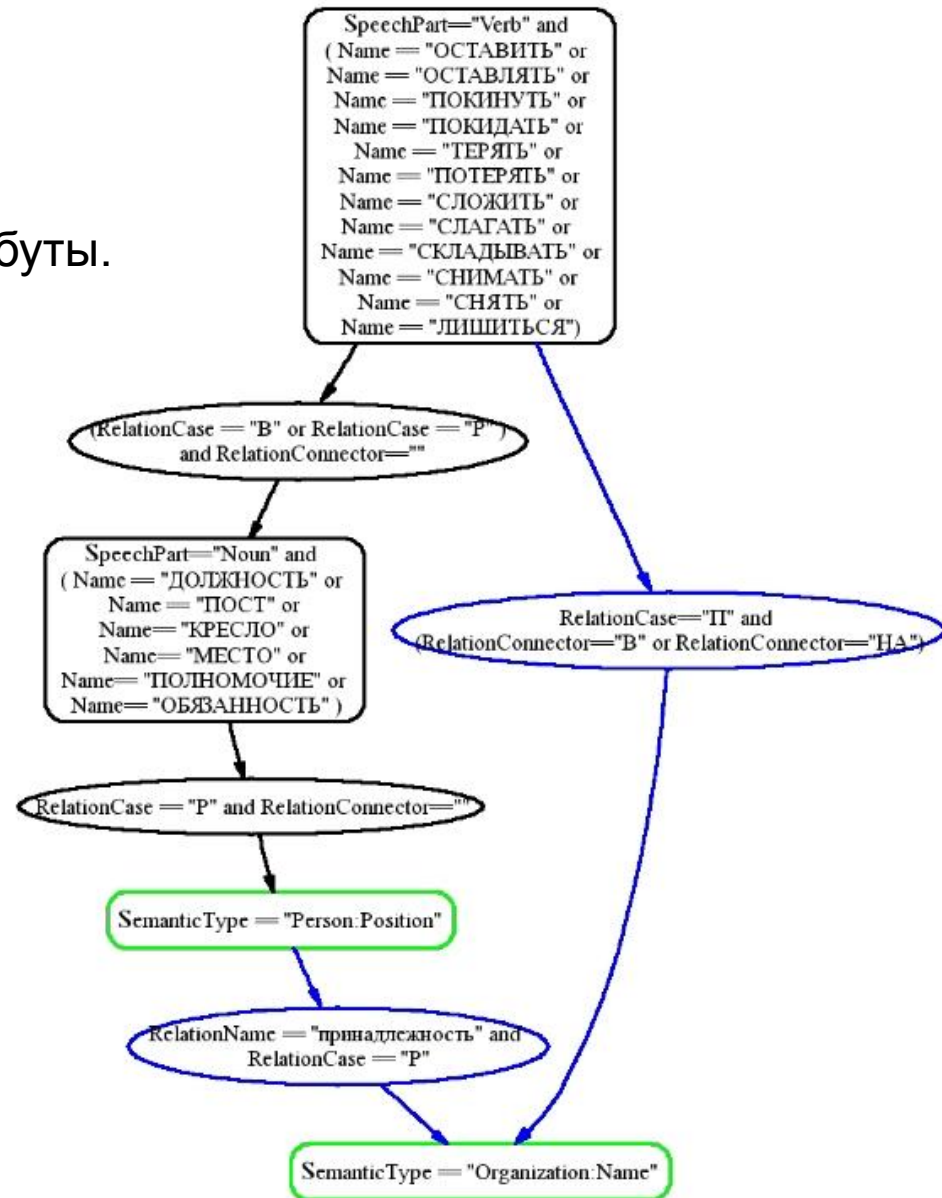
Производится с помощью шаблонов на основе синтаксического разбора предложения.

В графе синтаксического разбора атрибуты.
В графе шаблона у узлов ограничения.

Ищется подграф в графе синтаксического разбора, у которого атрибуты соответствуют ограничениям шаблона

Типы вершин в шаблоне:

- обязательные
- необязательные
- запрещающие



(Name = "ВЫРОС ОБЪЕМ ПРОДАЖИ" or Name = "ВЫРАСТИ" or Name = "ВЫРОС ОБЪЕМ") and (SemanticType = "Event") and (SpeechPart = "Verb") and (Taxis = "Past" or Taxis = "Perfect" or Taxis = "Person3") and (Number = "Singular")

(RelationName = "атрибут-наречие") and (RelationCase = "") and (RelationConnector = "") and (RelationRole = "")

(Name = "НА 78%") and (SemanticType = "Numeric.Percent") and (SpeechPart = "Adverb")

(RelationName = "аргумент") and (RelationCase = "И") and (RelationConnector = "") and (RelationRole = "субъект")

(Name = "ОБЪЕМ ПРОДАЖИ" or Name = "ОБЪЕМ") and (SemanticType = "Feature") and (SpeechPart = "Noun") and (Number = "Singular")

(RelationName = "принадлежность") and (RelationCase = "P") and (RelationConnector = "") and (RelationRole = "")

(Name = "ПРОДАЖА") and (SemanticType = "Event") and (SpeechPart = "Noun") and (Number = "Plural")

(RelationName = "аргумент") and (RelationCase = "P") and (RelationConnector = "") and (RelationRole = "объект")

(Name = "OZON") and (SemanticType = "Organization" or SemanticType = "Organization.Name") and (SpeechPart = "Noun") and (Number = "Singular")

(SpeechPart = "Verb" and (Name = "ВОЗРАСТАТЬ" or Name = "ВОЗРАСТИ" or Name = "ВЫРАСТИ" or Name = "ПОВЫСИТЬ" or Name = "ПОВЫСИТЬСЯ" or Name = "ПОВЫШАТЬ" or Name = "ПОВЫШАТЬСЯ" or Name = "ПОДНИМАТЬ" or Name = "ПОДНИМАТЬСЯ" or Name = "ПОДНЯТЬ" or Name = "ПОДНЯТЬСЯ" or Name = "ПРЕВЫСИТЬ" or Name = "ПРЕВЫШАТЬ" or Name = "РАСТИ" or Name = "УВЕЛИЧИВАТЬ" or Name = "УВЕЛИЧИВАТЬСЯ" or Name = "УВЕЛИЧИТЬ" or Name = "УВЕЛИЧИТЬСЯ" or Name = "УДВАИВАТЬСЯ" or Name = "УДВОИТЬСЯ" or Name = "УСИЛИВАТЬ" or Name = "УСИЛЯТЬ" or Name = "УСИЛЯТЬСЯ" or Name = "УСИЛИВАТЬСЯ" or Name = "УТРАИВАТЬСЯ" or Name = "УТРОИТЬСЯ")) or (SpeechPart = "Noun" and (Name = "ВОЗРАСТАНИЕ" or Name = "ПОВЫШЕНИЕ" or Name = "ПРЕВЫШЕНИЕ" or Name = "РОСТ" or Name = "УВЕЛИЧЕНИЕ" or Name = "УСИЛЕНИЕ" or Name = "УДВОЕНИЕ" or Name = "УТРОЕНИЕ"))

RelationRole = "субъект" or RelationRole = "объект"

Name = "ОБЪЕМ"

RelationCase = "P" and RelationConnector = ""

Name = "ПРОДАЖА" or Name = "ПОСТАВКА"

RelationCase = "P" and RelationConnector = ""

SemanticType = "Organization"

Разбор таблиц

С помощью правил, написанных на языке САРЕ, связывает сущности из разных полей таблицы и оформляет эти связи в виде фактов.

Пример таблицы:

Наименование организации	ИНН	ОГРН	Адрес регистрации (для иностранных компаний)	Общая стоимость контрактов (тыс. руб.)	Виды контрактов
АвангардСтрой ООО	5609061753	1075658001726		729	Купля-продажа
Автолюкс ООО	6314021059	1026300898645		29	Подряд
Аксиома-Сервис ООО	7701801805	5087746163060		633	Оказание услуг
АЙ ЭС ДЖИ ЗАО	7707329787	1055003031214		314	Оказание услуг
АйСи ИМПЭКС ООО	7701751600	1077761839902		125	Оказание услуг
Ачимгаз ЗАО	8904047896	1068904007578		398	Поставка
Аксоль ОАО ПКФ	3016033920	1023000819676		608	Поставка
Алюминевая продукция ЗАО	6660085435	1026604955882		770	Подряд
Астраханьэнергосбыт ОАО	3017041554	1053000000041		165	Поставка
Бейкер Хьюз Б.В. Компания	9909183206	-	: 125167, РФ, Москва, Ленинградский проспект, д. 37, корп. 9.	163	Оказание сервисных услуг при строительстве скважин