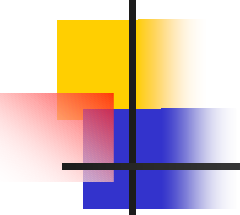


Интегрум на РОМИП-2005. Система извлечения структурированной информации.



Integrum Techno

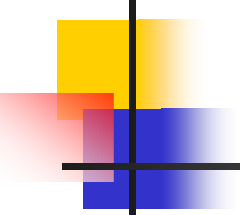
Информационное агентство «Интегрум-Техно»



Структурированная информация

Множество типизированных объектов,
связанных содержательными,
экстралингвистическими отношениями.
(Например, объекты типа ФИО, связанные
отношениями «родственник», «коллега».)

Integrum Techno



Постановка задачи

1. Выделение именованных сущностей

- а. Выделение ФИО
- б. Выделение компаний

2. Выделение фактов

- а. Где работал\работает данный человек?
- б. Какими предприятиями владеет или владела данная организация/персона?

Integrum Techno

- пометы в морфологическом словаре для обозначения имен, отчеств и фамилий
- наличие словарной фамилии необязательно
- отождествление по тексту формально различных фио
- разрешения родовой омонимии для имен Александра, Валерия

Выделение фактов

- **Выделение цепочек**

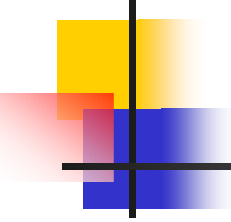
*По словам замдиректора компании «ТрансСофт»
Антон Петрова,*

- **Интерпретация участников по ролям**

ФИО	Должность	Организация
<i>Антон Петрова</i>	<i>замдиректора</i>	<i>«ТрансСофт»</i>

- **Нормализация**

ФИО	Должность	Организация
<i>Антон Петров</i>	<i>заместитель директора</i>	<i>ТрансСофт</i>



Основные способы выражения в тексте

- **Именная группа:**

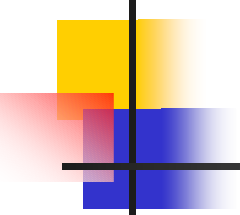
группы ФИО + Должность + Организация
(*А.Петров, директор «Газпрома»*)

- **Предикативная структура:**

вершина клаузы + актанты типа ФИО,
Должность, Организация
(*А.Петров раньше работал директором
«Газпрома»*)

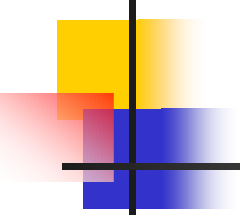
Принцип работы

- **Выделение ключевых слов**
(«управляющий директор», «контора», «месторождение»)
- **Описание синтаксических конструкций, вершинами которых являются ключевые слова**
(группа должности: группа *прил_сущ*, где главное слово «директор»)
- **Описание взаимного расположения и согласования таких синтаксических групп**
(ФДО: (группа ФИО) + «,» +
(группа должности) +
(группа компании) +
«,»)



Инструменты анализа

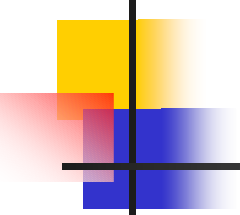
- **Словарь ключевых слов**
(слова и словосочетания, характерные для выделяемых фактов)
- **Язык Тома**
(язык описания синтаксической структуры выделяемых фактов)
- **Шаблоны предикативных структур**
(язык описания грамматических признаков, лексического состава и взаимного расположения актантов и вершины клаузы)



Язык Тома

- В основу парсера положен алгоритм Томиты (GLR-grammar, бесконтекстная грамматика)
- Терминалы:
 - одиночные слова предложения
 - объекты типа фии, дата, число
 - словосочетания из словаря ключевых слов
 - цепочки, распознанные другими грамматиками
- Проверка согласования и приписывание вершины в правых частях правил
- Порождение всех возможных вариантов синтаксического разбора распознаваемой цепочки и выбор лучшего

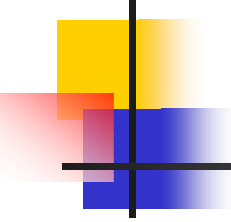
Integrum Techno



Шаблоны для ситуаций

- Используют результаты фрагментационного анализ:
 - дерево построенных фрагментов
 - синтаксические связи между подлежащим и сказуемым, причастием и определяемым существительным
- Работают внутри одной клаузы
- Приписываются вершинам клауз
- Актантами являются цепочки, распознанные специальными грамматиками

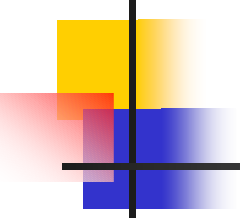
Integrum Techno



Проблемы оценки выделения ФИО

- Нормализация
(«*Николая Непейводы*» -> «*Николай Непейвода*»)
- Восстановление полных имен и отчеств.
(«*Н. Н. Непейвода*» -> «*Николай Николаевич Непейвода*»)

Integrum Techno



Проблемы постановки задачи по выделению фактов (компаний)

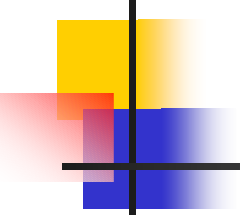
- Что считать за компанию?
(«Комитет Госдумы по конституционному законодательству и государственному строительству», «региональное духовное управление Ямало-Ненецкого автономного округа »)
- Что включать в имя компании?
*(«универмаг Москва» или «Москва»
«РАО ЕЭС» или «ЕЭС»)*
- Нормализация
- Референтные обозначения

Integrum Techno

Проблемы постановки задачи по выделению фактов (должности)

- Должности с определениями типа «бывший»
(«бывшего начальника областного управления ФСБ»)
- «Будущие» должности
(«Анатолий Быков снова станет депутатом Законодательного собрания Красноярского края.»)
(«Кандидат на пост мэра Екатеринбурга банкир Олег Гусев»)
(«кандидат в депутаты»)

Integrum Techno

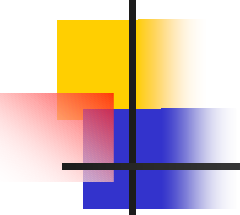


Проблемы оценки фактов

Два способа оценки полноты:

- Традиционная оценка – все вхождения фактов одинаково ценны
- Информационная полнота – одинаковые факты склеиваются, оцениваются разные факты, независимо от частоты встречаемости

Integrum Techno

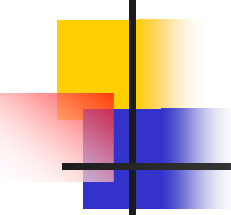


Наши результаты

- Точность по всем фактам: 0.93
- Точность по фактам типа «owns»: 0.67
- Точность по фактам типа «employs»: 0.94

(Другие оценки пока недосчитаны)

Integrum Techno



Основные типы ошибок

- Неправильная должность («автор», «кандидат на...», «бывший ...»)
- Неправильное выделение ФИО
(«руководитель внешнеполитического отдела аппарата президента **Новруз Мамедов**»)
(«член правления предприятия "Viljandi Metall" **Яак Сульг**»)
- Неправильное выделение компании
(«член Научного совета Московского центра Карнеги **Алексей Малащенко** » - компания Карнеги)
- Сложные сочиненные конструкции
- Отрицательные вхождения
(«Осинцев, подчеркнул Крицкий, не является членом "Единой России" »)