

Ефименко И.В.

Irina.Efimenko@avicomp.ru



**ОБРАБОТКА ЕСТЕСТВЕННОЯЗЫКОВЫХ
ТЕКСТОВ: ОНТОЛОГИЧНОСТЬ В ЛИНГВИСТИКЕ
И ДИСКУРСИВНОСТЬ В ИЗВЛЕЧЕНИИ ЗНАНИЙ**

План презентации

- ❑ **Введение**
- ❑ **Свойства дискурса**
- ❑ **Понятие контактности
и Shallow-подход**
- ❑ **Разграничение релевантных и нерелевантных данных,
разрешение конфликтов**
- ❑ **Онтологии:
интерпретация лингвистических данных**
- ❑ **Заключение**

Введение

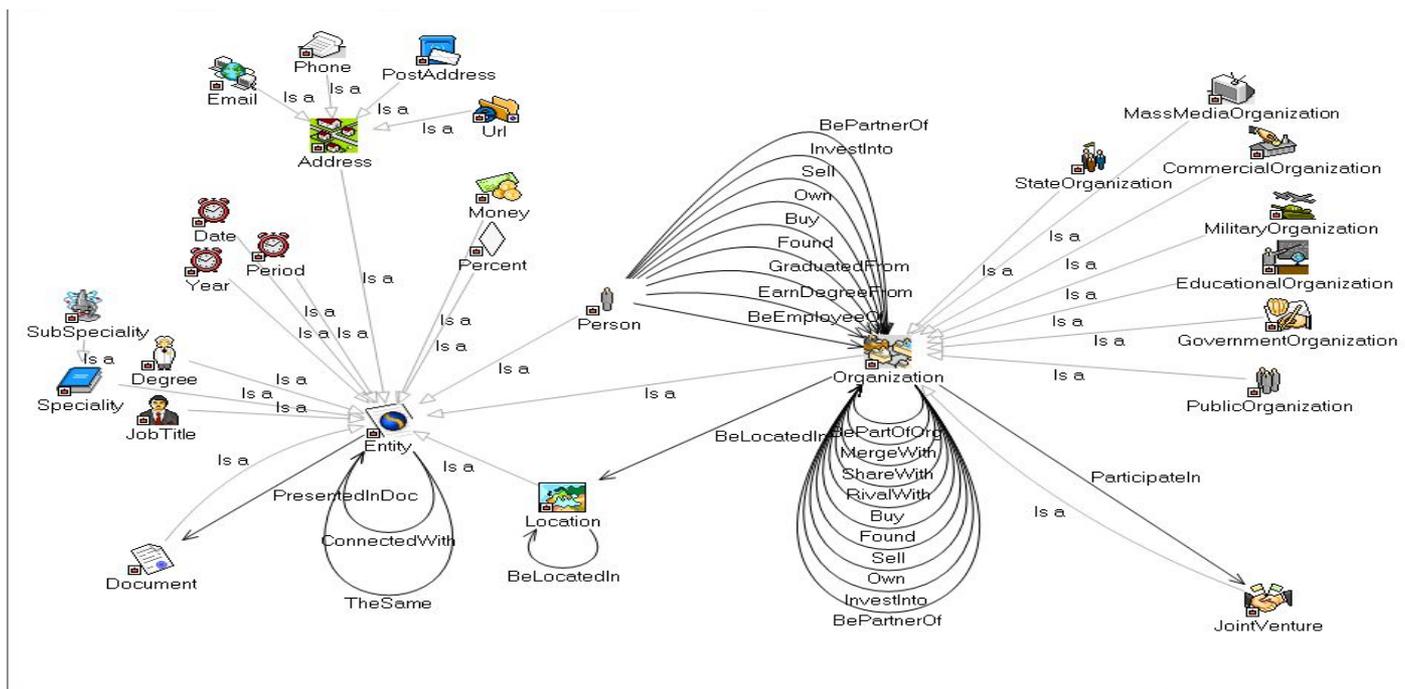
- ❑ Онтология как фильтр
- ❑ Shallow-подход и дискурс, имитация синтаксического анализа
- ❑ Лексические vs. предметные онтологии, обращение к экстралингвистическим данным
- ❑ Дискурсивный подход vs. анализ отдельных фрагментов

Введение

- Принципы работы многоязыковых систем семейства OntosMiner:
 - ❖ Анализ под управлением онтологий
 - ❖ Модифицированный Shallow-подход. Принцип «контактности» (закономерности развертывания дискурса)
 - ❖ Использование онтологических знаний
 - на этапе формирования модели
 - при интерпретации лингвистических явлений (в частности, разрешении неоднозначности, проявляющейся на различных уровнях автоматической обработки).
 - ❖ Дискурс и онтология: разрешение кореференции и анафоры
 - ❖ Понятие «аннотации» как служебного и/или семантического ярлыка. «Технологические приемы».

Дискурс: Линейность

- Непрерывность, связность, смысловое единство дискурса
- «Связность» онтологии, интерпретация изолированных объектов



Понятие контактности и Shallow-подход:

«Типология ошибок»

□ Нарушение целостности шаблона, «ошибка первого рода»

❖ Пример 1: *Синицына (в девичестве Орлова) Анна-Мария Гузермес, выпускница Одесского сельскохозяйственного техникума и участник конференции «Сделаем «Красную Книгу» белой», является менеджером картеля «Лига Охраны Перелетных Птиц».*

- Аннотации объектов: Синицына (в девичестве Орлова) Анна-Мария Гузермес (тип: [Лицо](#)); Одесского сельскохозяйственного техникума (тип: [Организация](#)); возможно, конференции «Сделаем «Красную Книгу» белой» (тип: [Организация](#)), менеджером (тип: [Должность](#)) и картеля «Лига Охраны Перелетных Птиц» (тип: [Организация](#)).

Понятие контактности и Shallow-подход:

«Типология ошибок»

- Пример 1: *Синицына (в девичестве Орлова) Анна-Мария Гузермес, выпускница Одесского сельскохозяйственного техникума и участник конференции «Сделаем «Красную Книгу» белой», является менеджером картеля «Лига Охраны Перелетных Птиц».*

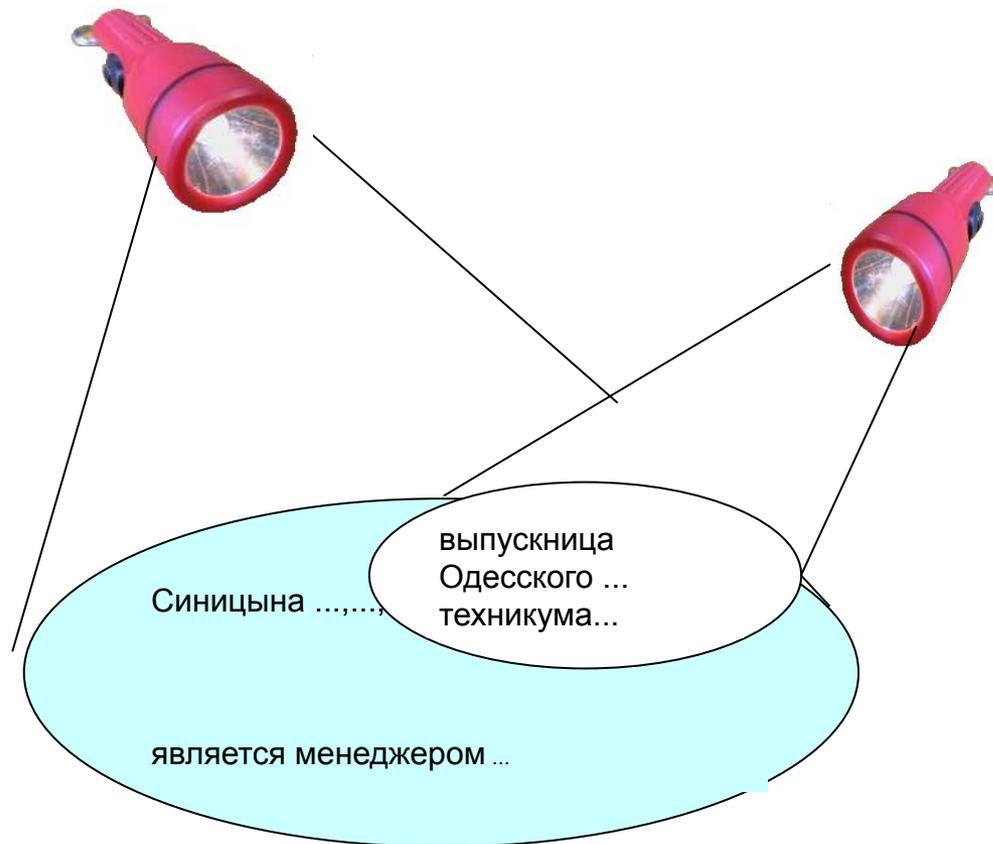
- Входные аннотации на последующих этапах обработки (идентификация связей): Лицо, Организация, Должность и служебная аннотация, маркирующая онтологический предикат (в данном случае, глагол «являться» в определенной форме).
- Схема шаблона: {Лицо (в соответствующей грамматической форме)}, {"являться" в 3 л. ед.ч.}, {Должность (в соответствующей грамматической форме)}, {Организация (в соответствующей грамматической форме)}.

Понятие контактности и Shallow-подход: «Типология ошибок»

- **Ошибочная интерпретация шаблона,
«ошибка второго рода»**
 - ❖ Пример 2: Лю Чю Хе Сянь Вань является автором модуля, который много лет успешно работает в системе «Биг Пис» (из предыдущего контекста при этом следует, что «Биг Пис» - название компании).
 - Шаблон: {Лицо (в соответствующей грамматической форме)}, {"работать" в 3 л. ед. ч.}, {Организация (в соответствующей грамматической форм, с предлогом)}
 - Интерпретация: «Лю Чю Хе Сянь Вань много лет успешно работает в системе «Биг Пис»

Наличие ограничений на семантику актантов не является решением

Разграничение релевантных и нерелевантных данных: фокус внимания



Методы разрешения конфликтов:

Пример списка с атрибутами (фрагмент реального текста)



Установлены члены международного синдиката «Золотой мак»:

- ❑ Мгерабишвили Зураб Вахтангович, 1943 г.р., ур. и житель г. Поти, Грузия, лидер синдиката, женат на Мгерабишвили А. К.
- ❑ Могулиев Абдулхайр Магомедович, 17 марта 1984 года рождения, уроженец Согдийской области Таджикистана, житель кишлака Одурван.
- ❑ Чон Ду Хван, гражданин Кореи, 1939 г.р., курьер, брат гражданина Кореи Ли Ю Тинь, верховного жреца «Группы раскаявшихся флибустьеров Капитана Флинта»
- ❑ Братья Кукушкины – Сергей Анатольевич, 1978 г.р., и Петр Анатольевич, 1980 г.р., уроженцы Белгородской области, проживают: Республика Северная Осетия-Алания, г.Ардон, ул. Железнодорожная, д.5 кв. 1. Оба числятся грузчиками в ООО «Ближний свет» (Республика Северная Осетия-Алания, г.Ардон, ул. Железнодорожная, д.5)
- ❑ Ли Си Цин, гр. КНР, постоянно проживает в Ташкенте, Узбекистан, хозяин городского рынка «Бешкеш»
- ❑ Абдуллаев Кодир Исмоилович, 15.10.66 г.р., ур. г. Андижан, Узбекистан, проживает в Узбекистане: г.Корасув, ул.Навруз д. 28, кв. 2, безработный, его женою является известная Ибрагимова Насибахон Шухратовна, 9 марта 1980 г.р., уроженка и жительница г.Корасув, ул.Навруз д. 28, кв. 2, медсестра городской больницы № 4

Методы разрешения конфликтов:

Пример списка с атрибутами

- ❑ *Необходимо установить связь типа «являться сотрудником, работать» между Организацией и каждым из лиц, являющимися вершинами элементов списка.*
- ❑ *Недопустимо появление связи типа «являться сотрудником, работать» между Организацией и другими лицами, фигурирующими в тексте, но при этом не являющимися вершинами элементов списка.*
- ❑ *Дополнительные маркеры вершин списка могут отсутствовать.*

Методы разрешения конфликтов :

Использование имен и атрибутов аннотаций

- *Приписывание атрибутов*
 - ◆ Организация (Лицо.attr == "1", (Лицо)*)+
- *Переименование аннотаций*
 - ◆ Организация (Лицо1)+
- *«Захват» нерелевантных фрагментов*
 - ◆ Организация (Элемент списка)+
- *Включение во входные данные «лишних» аннотаций*
 - ◆ Input: **Comma...**

Методы разрешения конфликтов :

Интерпретация списочных структур

- ❑ **Гвинджи Фануэл Таванда (Gvindgy Fanuel Tavanda);**
- ❑ **Горезваримва Портия (Goredfrimva Portiya);**
- ❑ **Мпоко Луринда; Нтандо Анние Дзиямо Тадуру, 1981 г.р.**
- ❑ **Такавира-Куун Клаудиус;**
- ❑ **Сбанда Тобекиле (Sibanda Tobekili), 22.05.1982 г.р.**

Онтологии:

интерпретация лингвистических данных

□ 1. Восстановление имплицитной информации

◆ А) Восстановление эллипсисов.

- «До IBM, г-н X работал Microsoft» (две связи одного типа - "работать, быть сотрудником" - с одним общим актантом)

◆ Б) Восстановление ситуаций за рамками текста.

- «В этом году г-н X стал главным редактором газеты "Известия"»
- «В этом году г-н X стал программистом Oracle»

Онтологии:

интерпретация лингвистических данных

□ 2. Интерпретация типа временной сущности

- ❖ *Задача взаимного расположения на оси времени извлекаемых событий*
- ❖ *Динамическое изменение интерпретации элементов шкалы в рамках дискурса*
 - «Иванов был уволен из МВД в 1985 году» -> «Иванов работал в МВД до 1985 года»

Заключение

- ❑ **Целесообразность автоматической обработки естественных языковых текстов под управлением предметных онтологий в контексте восприятия входного текста как целостного дискурса**
- ❑ **Необходимость использования экстралингвистической информации при интерпретации лингвистических данных**
- ❑ **Совмещение двух концепций, являющихся в настоящее время наиболее актуальными в смежных, с точки зрения ЕЯ-систем, дисциплинах: онтологически-ориентированных методов в области ИИ и информационных технологий и дискурсивного анализа в лингвистике**
 - ❖ **Новый класс подходов к автоматической обработке естественного языка**

***Спасибо
за внимание!***