

Технологии извлечения знаний из неструктурированных текстов

Разработка практических заданий и лабораторных работ для
дисциплин блока "Компьютерная лингвистика»

Карпов Николай
Бонч-Осмоловская Анастасия
Сибирцева Вера
Савченко Андрей
Малафеев Алексей

Постановка задачи

Текст

Модель

Структурированные данные

МОСКВА, 15 мая - РИА Новости.
 Руководитель Росатома Сергей Кириенко 19-23 мая в ходе поездки в США проведет ряд рабочих встреч, посвященных двустороннему сотрудничеству в области мирного использования атомной энергии, говорится в сообщении пресс-службы Росатома. Планируется, что Кириенко 22 мая проведет переговоры с министром энергетики США Самуэлем Бозманом и руководителем комиссии по ядерному регулированию США Нильсом Дингом.

ТИПЫ ОБЪЕКТОВ И
ТИПЫ ОТНОШЕНИЙ

РАБОТАТЬ В ОРГАНИЗАЦИИ

ОРГАНИЗАЦИЯ

ПЕРСОНА

Сергей Кириенко

Росатом

Билибинская АЭС

Работает в

Владеть

Руководитель Росатома Сергей Кириенко 19-23 мая в ходе поездки в США проведет ряд рабочих встреч...

Распознавание словосочетаний

The screenshot shows the dixGUI application interface. At the top, there is a menu bar with 'Файл', 'Настройки', 'Утилиты', and 'Помощь'. Below it is a toolbar with navigation icons. The main window is titled 'Анализ словосочетаний' and contains a text input field with the sentence: 'В этом году директор Челябинского металлургического комбината вышел на пенсию.' Below the input is an 'Обновить' button. To the right is a 'Словари словосочетаний' panel with a list of dictionaries: 'Словосочетания: должности', 'Словосочетания: организации', 'Словосочетания: адреса', 'Словосочетания: время', 'Тестирование: Русские словосочетания', 'Словосочетания: отношения', 'Словос...', and 'Словос...'. The 'Словосочетания: должности' dictionary is selected. Below the input and dictionary panels is a grid of green boxes representing recognized words and their grammatical information. The word 'директор' is highlighted in red. Below the grid is a text field showing the original sentence with the recognized words highlighted in green. At the bottom, there is a 'Lookup' section with the following text: 'Lookup: string=директор; length=8; kind=word; lang=cyr; orth=lowercase; base=директор; WORDFORM=директор; ZINDEX=мо 1с 1"; ACCPL=4; DICTID=51; ID=17474; SID=21; OFFSET=12; POS=N; ANIM=anim; GEND=m; NMB=sg; CAS=nom; majorType=jobTitle;'. Three yellow callout boxes with arrows point to the input text, the dictionary list, and the 'jobTitle' label in the lookup results.

Входной текст

Словари, по которым производится проверка

Результат распознавания: найдены три словосочетания из указанных словарей

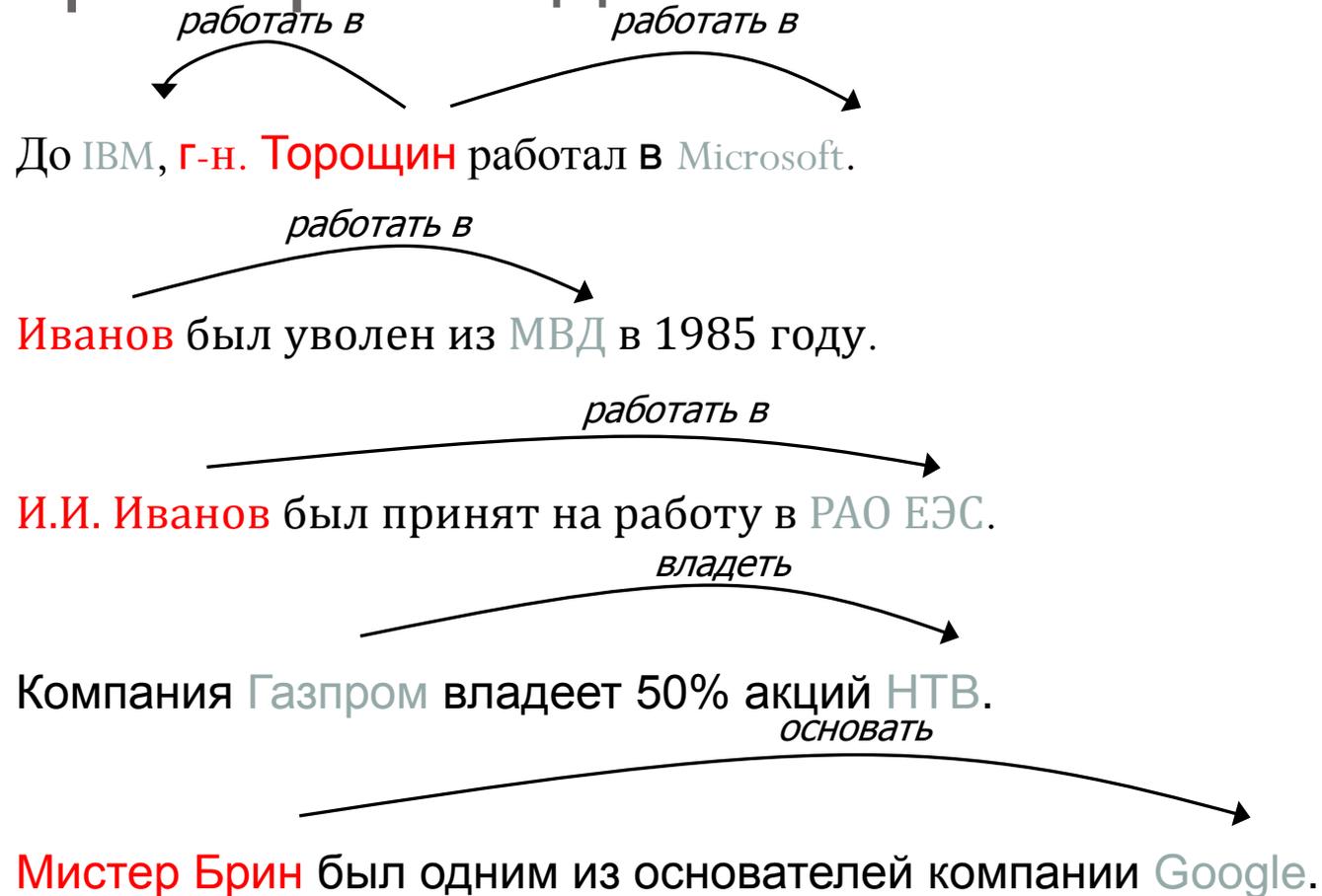
Выделение отношений

- «Осыпание» нерелевантных фрагментов текста
- Пример с отношением «работать»:

```
{Date} | {StartPoint}? {Person}  
{becomeVG.VOICE="act", becomeVG.MOOD="ind"}  
{JobTitle} {Organization }
```

В середине 2003-2004 года бывший немецкий гражданин Хайнц Шиммельбуш неожиданно становится директором малоизвестной компании «Васюки».

Примеры выделения отношений

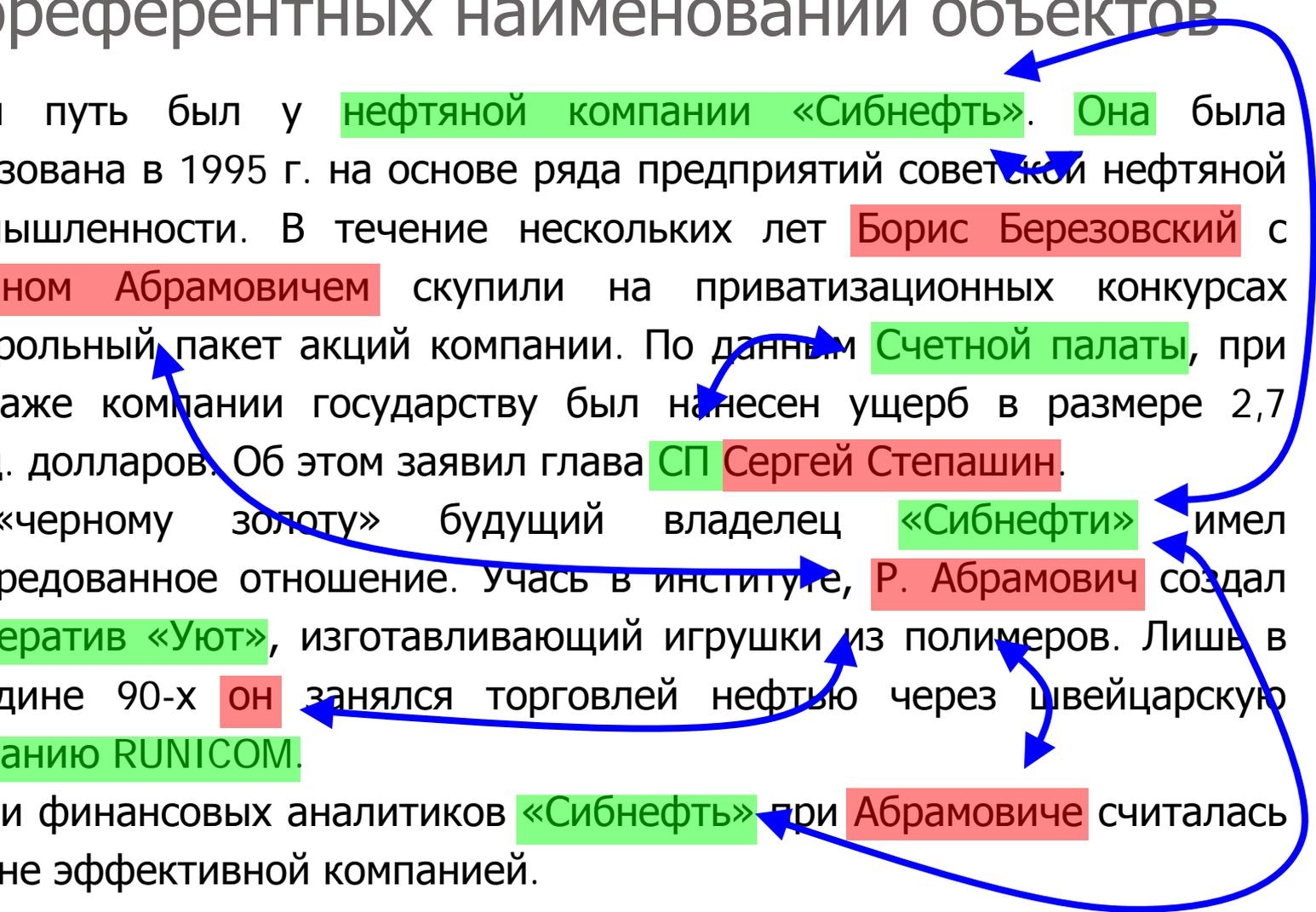


Вторичное распознавание и связывание кореферентных наименований объектов

Иной путь был у **нефтяной компании «Сибнефть»**. Она была образована в 1995 г. на основе ряда предприятий советской нефтяной промышленности. В течение нескольких лет **Борис Березовский** с **Романом Абрамовичем** скупили на приватизационных конкурсах контрольный пакет акций компании. По данным **Счетной палаты**, при продаже компании государству был нанесен ущерб в размере 2,7 млрд. долларов. Об этом заявил глава **СП Сергей Степашин**.

К «черному золоту» будущий владелец **«Сибнефти»** имел опосредованное отношение. Участь в институте, **Р. Абрамович** создал **кооператив «Уют»**, изготавливающий игрушки из полимеров. Лишь в середине 90-х **он** занялся торговлей нефтью через швейцарскую **компанию RUNICOM**.

Среди финансовых аналитиков **«Сибнефть»** при **Абрамовиче** считалась крайне эффективной компанией.

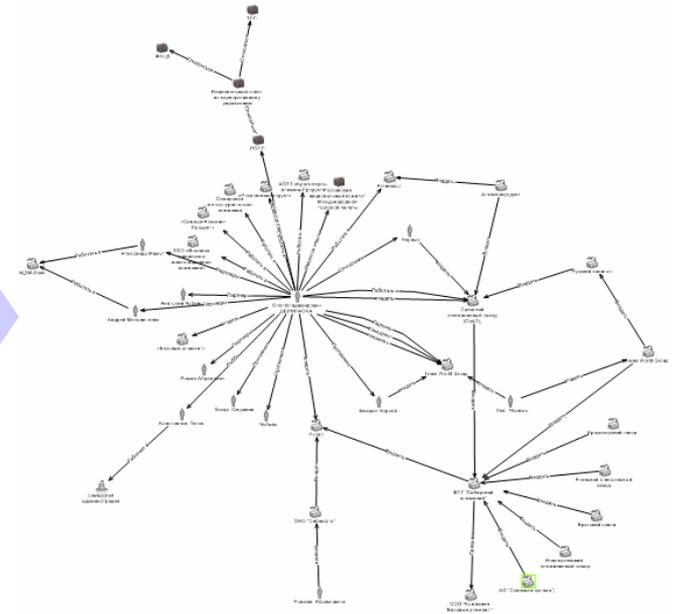
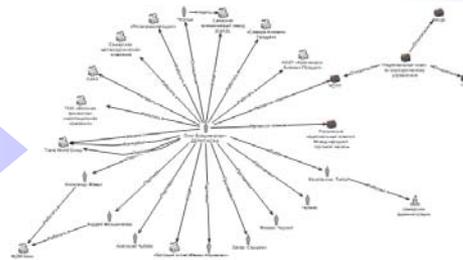
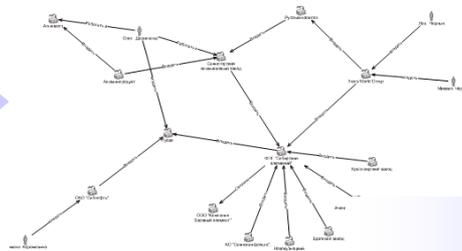


Интеграция информации

После сохранения необходимо выявить и объединить идентичные объекты из разных документов.

МОСКВА, 15 мая - РИА Новости. Руководитель Росатома Сергей Кириенко 19-23 мая в ходе поездки в США проведет ряд рабочих встреч, посвященных двустороннему сотрудничеству в области мирного использования атомной энергии, говорится в сообщении пресс-службы Росатома. Планируется, что Кириенко 22 мая проведет переговоры с министром энергетики США Самуэлом Бодманом и руководителем комиссии по ядерному регулированию США Нильсом Дивозом.

МОСКВА, 15 мая - РИА Новости. Руководитель Росатома Сергей Кириенко 19-23 мая в ходе поездки в США проведет ряд рабочих встреч, посвященных двустороннему сотрудничеству в области мирного использования атомной энергии, говорится в сообщении пресс-службы Росатома. Планируется, что Кириенко 22 мая проведет переговоры с министром энергетики США Самуэлом Бодманом и руководителем комиссии по ядерному регулированию США Нильсом Дивозом.



тексты

отдельные графы

база знаний