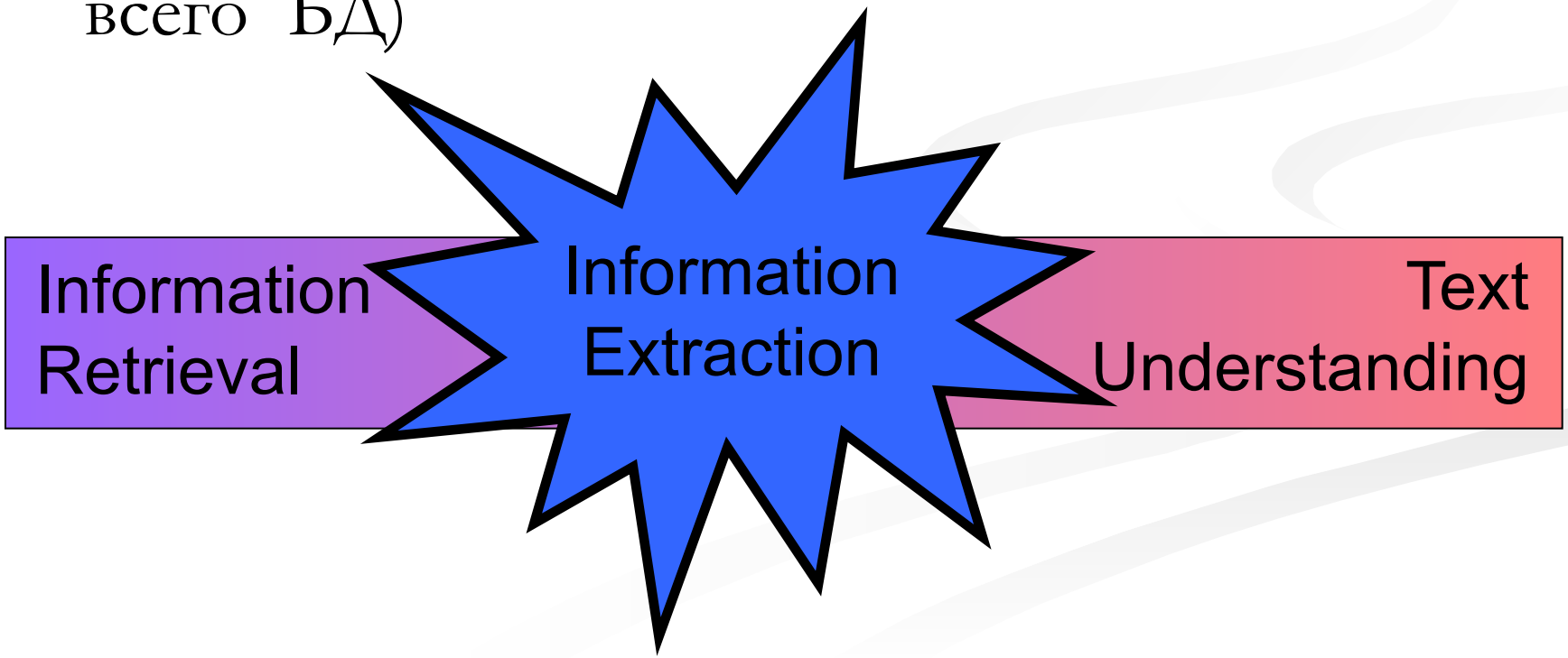
Извлечение информации

Лидия Михайловна Пивоварова

Системы понимания текста

Введение

Information Extraction – извлечение из текста информации определенного типа и представление ее в заданном формате (чаще всего БД)



Мотивация

- Пополнение баз данных (и баз знаний)
- Получение входных данных для работы других систем
- Привлечение внимания эксперта к значимым аспектам информации

Содержание

- Message Understanding Conference
- Извлечение информации: основные подходы
- Named Entity Recognition
- Извлечение отношений
- Наш опыт

MUC (Message Understanding Conference), 1987-1997

– выработка общих подходов к методологии и способам оценки систем извлечения информации из текста.

	Год	**Источники**	**Предметная область**
MUC-1	1987	военные сводки	военно-морские операции
MUC-2	1989	военные сводки	военно-морские операции
MUC-3	1991	новости	Террористическая активность
MUC-4	1992	новости	Террористическая активность
MUC-5	1993	новости	совместные предприятия, производство
MUC-6	1995	новости	смена лидеров на рынке
MUC-7	1997	новости	крушения самолетов, запуски ракет

Дорожки MUC

- **Named Entity recognition** - выделение именованных сущностей
- **Coreference resolution -** разрешение кореференции
- **Template Element construction -** добавление атрибутов к сущностям, найденным на этапе NE, с использованием CR
- **Template Relation construction** – выявление связей между отдельными сущностями
- **Scenario Template production** – построение полного описания события (факта) путем объединения результатов TE и TR

Блестящая красная ракета была запущена во вторник. Это изобретение доктора Биг Хеда. Хед - штатный научный сотрудник Билд Рокет Инкорпорейтед.

Named Entity recognition:

доктор Биг Хед, Хед, Билд Рокет Инкорпорейтед

ракета, вторник…

Coreference resolution:

доктор Биг Хед ≈ Хед

это → ракета

Template Element construction:

Ключ	Объект	Цвет	Светоотражательные свойства
0267	Ракета	Красная	Блестящая

Блестящая красная ракета была запущена во вторник. Это изобретение доктора Биг Хеда. Хед - штатный научный сотрудник Билд Рокет Инкорпорейтед.

Template Relation construction:

Ключ	Объект	Цвет	Светоотражательные свойства	Изобретен
0267	Ракета	Красная	Блестящая	7824

Ключ	ФИО	Степень	Работает	Должность
7824	Биг Хед	Доктор	2345	452

Scenario Template production:

Ключ	Тип события	Объект	Дата
18	Запуск	0267	Вторник

Оценка

$$\text{Recall} = N_{correct} / N_{all\text{-}correct}$$

$$\text{Precision} = N_{correct} / (N_{correct} + N_{incorrect})$$

$$\text{F-mera} = (\beta^2+1)*r*p/(\beta^2 * r+p)$$

- **Named Entity recognition** F<94%
- **Coreference resolution** F<62%
- **Template Element construction** F<87%
- **Template Relation construction** F<76%
- **Scenario Template production** F<51%

Дальнейшее развитие

- ACE (Automatic Content Extraction) 1999 – 2008
- По сравнению с MUC:
  - более детальная таксономия сущностей
  - для всех систем обязательна интерпретация метонимических связей
  - требуется семантический анализ обрабатываемого текста
- Text Analysis Conference (TAC) – настоящее время

Содержание

- Message Understanding Conference
- Извлечение информации: основные подходы
- Named Entity Recognition
- Извлечение отношений
- Наш опыт

Основные подходы

- Задача всегда предельно конкретна:
  - определенный тип текста
  - искомая информация представлена в виде набора полей для заполнения
- Текст, включающий такую информацию, предельно шаблонный
- Поиск осуществляется при помощи набора образцов

Образцы

- Состав образцов:
  - Лексика, семантика
  - Частичный синтаксис
  - Близость, взаимное расположение частей
- Формат:
  - Зависит от формата представления текста в системе
  - Часто используются специальные языки (грамматики)
- Построение образцов:
  - Вручную
  - Машинное обучение (bootstrapping)
  - Обобщение образцов с привлечением словарных и/или онтологических ресурсов

Машинное обучение

- Pro:
  - не требует большого количества ручного труда по написанию правил
  - система более гибкая, ее легко перенастроить
- Contra:
  - требуется большой обучающий корпус, правильно и полностью размеченный
  - сложно отследить в каком именно месте возникла ошибка и исправить ее «точечно»

Правила

- Pro:
  - Может быть предпочтительна в случае сложной предметной области и/или отсутствия лингвистических ресурсов
- Contra
  - Большая ручная работа, требующая специальной квалификации
  - Трудно перенастраивать

- Возможны (и даже предпочтительны) гибридные подходы

Содержание

- Message Understanding Conference
- Извлечение информации: основные подходы
- Named Entity Recognition
- Анафора и кореференция
- Наш опыт

Извлечение именованных сущностей

- Named Entity:
  - Стандартные примеры: персоналии, географические названия, организации…
  - Для биологических текстов: названия генов, белков, ферментов…
- Не только для Information Extraction: ответы на вопросы, извлечение мнений, реферирование…
- Named Entity Recognition: Information Extraction в миниатюре; проще, потому что не нужно извлекать связи между понятиями

Основные подходы

- Основанный на знаниях:
  - список имен собственных
  - регулярные выражения, описывающие именованные сущности
  - образцы, описывающие контекст
- Машинное обучение
  - обучающий корпус
  - определение характерных свойств
  - поиск по этим свойствам

Содержание

- Message Understanding Conference
- Извлечение информации: основные подходы
- Named Entity Recognition
- Извлечение отношений
- Наш опыт

Извлечение отношений между понятиями

- Отношения:
  - Таксономические – РОД-ВИД, ЧАСТЬ-ЦЕЛОЕ…
  - Специфические для предметной области – СТРАНА-СТОЛИЦА, БЕЛОК-ФЕРМЕНТ…
- В тексте определяются:
  - Свойствами именованных сущностей
  - Лексическими свойствами контекста
  - Синтаксическими свойствами контекста
- Извлечение:
  - Правила (образцы) vs. машинное обучение
  - Поиск: начиная с именованных сущностей vs. Начиная с отношений

Анафора и кореференция

- Извлечение информации в масштабах текста
- Кореференция: возможно использование экстралингвистической информации
- Анафора: невозможно использование экстралингвистической информации
  - Вокруг местоимения отыскиваются существительные-кандидаты
  - Проверяется согласование
  - Статистики и эвристики
- Во многих систем не разрешается ни анафора, ни даже кореференция – трудоемкие алгоритмы, низкое качество

Содержание

- Message Understanding Conference
- Извлечение информации: основные подходы
- Named Entity Recognition
- Извлечение отношений
- Наш опыт:
  - Система фактографического поиска в газетных текстах
  - Система автоматического пополнения онтологии на основе энциклопедических и толковых словарей

Система фактографического поиска в газетных текстах

- Рубашкин В. Ш., Капустин В. А., Пивоварова Л. М., Чуприн Б. Ю. **Методы извлечения фактографической информации из текстов. Опыт разработки**. // Megaling'2007 Горизонты прикладной лингвистики и лингвистических технологий– Симферополь: Изд-во ДиАйПи, 2007.
- Пивоварова Л.М. **Фактографический анализ текста в системе поддержки принятия решений** // Вестник Санкт-Петербургского университета Сер. Филология, востоковедение, журналистика. 2010. Вып. 4 - 190-197

Система Factors

Поиск факторов

Добавить образец | Достоверно

Документ 101 | Документ 102 | Документ 103 | Документ 10

На предстоящем саммите Россия - ЕС в Гааге МИД РФ намерен затронуть ситуацию с правами русских в Латвии.

Состояние ситуации | Факторы

116

ID	Текст
1	Бюджет Латвии
2	ВВП Латвии
3	Уровень налогов в Латвии
4	Внешний долг Латвии
5	Внутренний долг Латвии
6	Объем финансирования обороны в Латвии
7	Объем финансирования науки и образования в Латвии
8	Затраты на содержание госаппарата в Латвии

Синтаксическая структура документа | Синтаксическая структура образцов | Образцы

ID фактора	ID образца	Текст образца	Степень достоверности
1	6	Бюджет Латвии	Достоверно
1	110	бюджет ... государство	Достоверно
2	111	внутренний валовой продукт ... государство	Достоверно
3	7	Уровень налогов в Латвии	Достоверно
3	112	налоговые поступления ... государство	Достоверно
4	113	внешний долг ... государство	Достоверно
5	114	внутренний долг ... государство	Достоверно
6	115	вооруженные силы ... расходы	Достоверно
6	116	оборона ... расходы	Достоверно
7	117	наука (как отрасль)* ... расходы	Достоверно
7	118	образование ... расходы	Достоверно

Образец | по документу | по образцу | по окрестности

Система, основанная на знаниях – используется онтология IntTez - http://inttez.ru/

Постановка задачи

Задача: извлечение из текстов СМИ информации общественно-политической тематики.

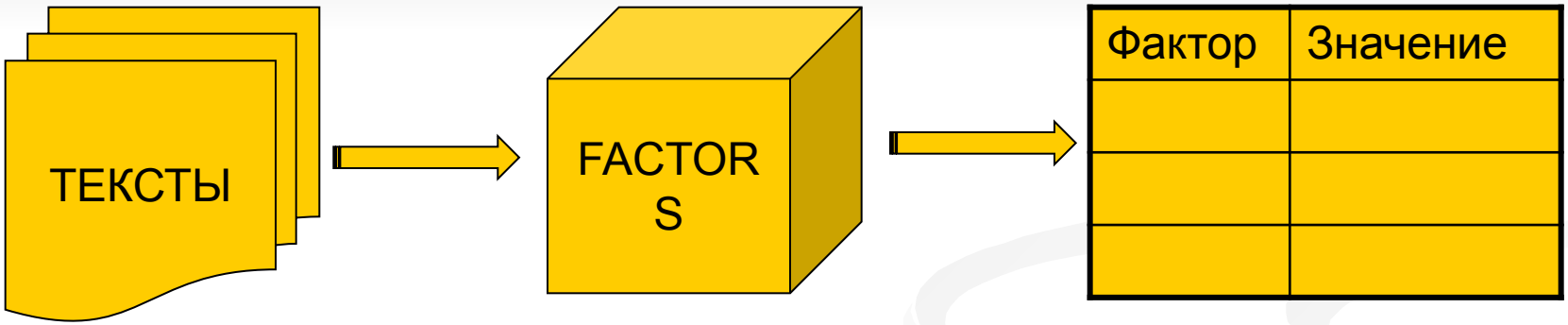
Факторы - различные характеристики общественно-политической ситуации(около 100).

Значения факторов:

- Количественные - *число пенсионеров; средний уровень заработной платы*
- Оценочные - *социальная напряженность; военные угрозы*

Система Factors:

- интеллектуальная среда для поддержки работы эксперта-аналитика с текстами.



Режимы работы:

- Автоматический
- Диалоговый

Функциональность:

1. Последовательное наращивание распознаваемых аспектов содержания в процессе работы эксперта-аналитика с системой.
2. Легкость и простота редактирования и пополнения; визуальное представление информации.
3. Функциональная расширяемость и переносимость на другие проблемные и предметные области.

Образцы

1. Текстовые – выделение в тексте релевантных фрагментов (при анализе может проверяться совпадение синтаксических связей)
2. Концептуальные – сборка образца из концептов **онтологии** (при анализе осуществляется поиск с учетом отношения «общее-частное»)
3. Смешанные

Образцы

- Фактор + значение

  В основном для оценочных факторов

  социальная напряженность → стихийный митинг

- Только фактор

  Для количественных факторов:

  уровень инфляции →

  инфляция составила 4%

Поиск образцов в тексте

население ... право на труд ... ограничение

1) Поиск опорного элемента

население ... право на труд ... ограничение

2) Поиск в окрестности других элементов

население ... право на труд ... ограничение

Для концептов образца – учет синонимов

ограничение = ограниченный, ограничить, ущемление

Параметры поиска предполагают отладку и настройку

Только фактор: поиск значения

Собственный признак фактора – концепт, отвечающий на вопрос «количество (величина) чего?»

Уровень зарплаты → заработная плата

Транспортные издержки → траты

Число пенсионеров → пенсионеры

Онтология:

собственный признак ↔ единица измерения

заработная плата ↔ денежная единица

пенсионеры ↔ без единиц

Общий алгоритм поиска

1) Поиск образца
2) Определение собственного признака и единиц измерения
3) Поиск числа с единицей измерения
4) Проверка соответствия единиц измерения
5) Если число не найдено – поиск слов *большой, маленький, растет, падает* и их синонимов
6) Определение достоверности

Содержание

- Message Understanding Conference
- Извлечение информации: основные подходы
- Named Entity Recognition
- Извлечение отношений
- Наш опыт:
  - Система фактографического поиска в газетных текстах
  - Система автоматического пополнения онтологии на основе энциклопедических и толковых словарей

Система автоматического пополнения онтологии на основе

- V. Bocharov, L. Pivovarova, V. Rubashkin, B. Chuprin **Ontological Parsing of Encyclopedia Information**. In Computational Linguistics and Intelligent Text Processing 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010. Proceedings. Lecture Notes in Computer Science. - Springer Berlin / Heidelberg – 2010 – pp. 564 – 579
- Бочаров В.В., Пивоварова Л.М., Рубашкин В.Ш. **Логико-лингвистический анализ текстов определений в энциклопедических и толковых словарях** // Сучасні технології комп'ютерної лексикографії (на матеріалах міжнародної конференції «MegaLing'2009») : Зб. наук. пр. / НАН України, Укр. мовно-інформ. фонд [та ін.]; редкол.: Ю. Д. Апресян [та ін.].— К. : Довіра, 2009
- Рубашкин В.Ш., Бочаров В.В., Пивоварова Л.М., Чуприн Б.Ю. **Опыт автоматизированного пополнения онтологий с использованием машиночитаемых словарей** // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). - М.: Изд-во РГГУ, 2010.

Пополнение онтологий

- Пополнение онтологий – бутылочное горлышко инженерии знаний
- Ontology Learning – автоматическое пополнение онтологии на основе естественно-языковых текстов

Источник

- Российский энциклопедический словарь - Гл. ред.: А. М. Прохоров — М.: Большая Российская энциклопедия, 2001

- исключены персоналии, географические названия и другие имена собственные
- 26375 словарных статей, 21782 различных терминов

Гипотеза

В большинстве случаев родовой по отношению к определяемому термин представлен первым по порядку существительным (именной группой) в именительном падеже.

АГРАФ - нарядная заколка для волос, с помощью которой крепили в прическах перья, цветы, искусственные локоны и т. д.

Примеры

ПЕРИСТИЛЬ - прямоугольный двор, сад, площадь, окруженные с 4 сторон крытой колоннадой.

ЯТАГАН - рубяще-колющее оружие (среднее между саблей и кинжалом) у народов Ближнего и Среднего Востока (известно с 16 в.).

Общий алгоритм анализа

Словарная статья (текст + пометы + сокр.)

↓

Лексикографическая обработка

↓

Словарная статья (текст)

↓

Синтаксический анализ

↓

Дерево зависимостей

↓

Извлечение отношений

↓

Отношения (термин – ключевое слово)

Лексикографическая обработка

- сокращения (разворачиваются в полные слова, если это возможно)
- пометы (удаляются)
- текст в скобках (удаляется)

Лексикографическая обработка

АБРЕКИ - В прошлом у народов Сев. Кавказа изгнанники из рода, ведшие скитальческую или разбойничью жизнь

АБРЕКИ - В прошлом у народов Северного Кавказа изгнанники из рода, ведшие скитальческую или разбойничью жизнь

АКСЕЛЕРАЦИЯ - (В антропологии) ускорение роста и полового созревания детей и подростков

АКСЕЛЕРАЦИЯ - ускорение роста и полового созревания детей и подростков

Синтаксический анализ

- Используются компоненты АОТ
- Упрощённые правила (Tomita-формализм)
- Строится дерево зависимостей

Упрощенные правила

ПРИЛАГАТЕЛЬНОЕ + ИМЕННАЯ ГРУППА

[ANP] -> [ADJ] [NP root]

: $0.grm := case_number_gender($1.grm, $2.type_grm, $2.grm);

ГЕНИТИВНАЯ ГРУППА

[GP] -> [NP root] [NP grm="рд"];

ПРЕДЛОЖНАЯ ГРУППА

[PP] -> [PREP root] [NP];

ИМЕННАЯ ГРУППА

[NP] -> [NOUN];

[NP] -> [NP root] [PP] ;

[NP] -> [PP] | [GP] | [ANP];

ПРИЛАГАТЕЛЬНОЕ + ИМЕННАЯ ГРУППА

ВОДОРОД - ХИМИЧЕСКИЙ ЭЛЕМЕНТ

ХАЛАТ – ВЕРХНЯЯ ОДЕЖДА

ГЕНИТИВНАЯ ГРУППА

АМПЕР - ЕДИНИЦА ИЗМЕРЕНИЯ

АБЗАЦ – ЧАСТЬ ТЕКСТА

ПРЕДЛОЖНАЯ ГРУППА

АВАЛЬ - ПОРУЧИТЕЛЬСТВО ПО ВЕКСЕЛЮ

АКСЕЛЕРОМЕТР – ПРИБОР ДЛЯ ИЗМЕРЕНИЯ УСКОРЕНИЯ

Синтаксический анализ: снятие неоднозначности

	До	После
Лемм / слово	1,27	1,06
Морфологических вариантов / слово	2,26	1,64

Неоднозначность: пример

- ***о чукотском море***

- ***море***
  - МОРЕ (ср.р.)
  - МОР (мр.р.)
  - МОРА (жр.р.)

- МОРА отбрасывается после синтаксического анализа

Отношение между термином и определением

- АВАНПОРТ - внешняя часть порта, предназначенная для стоянки судов, ожидающих подхода к причалам, погрузки и разгрузки.
- ШНЕК - название винтового конвейера.
- ПАРАБОЛОГРАФ - прибор для вычерчивания плоских кривых 2-го порядка (парабол).

Типы выделяемых отношений

Тождество	Same
Обобщение (значение по умолчанию)	Gen
Частный случай (обратное к GEN)	Spec
Часть	Part
Целое	Whole
Назначение	Func
другое	Other

Правила

- приписывается конкретному опорному слову
- описывает на какой тип отношений указывает данное слово
- следует ли сохранять данное слово в качестве опорного или необходимо отбросить его и перейти к следующему, указанному правилом.

Примеры правил: тождество

Обозначение

1. Тип отношения меняется на Same
2. Записывается следующее (по дереву) существительное

СОЦИОСФЕРА - обозначение человечества, общества, а также освоенной человеком природной среды, в совокупности составляющих часть географической оболочки.

Примеры правил: тождество

Явление

1. Записывается «явление»
2. Тип отношения меняется на Same
3. Записывается следующее (по дереву) существительное

СИНЕСТЕЗИЯ - явление восприятия, когда при раздражении данного органа чувств наряду со специфическими для него ощущениями возникают и ощущения , соответствующие другому органу чувств.

Зачем нужен первый пункт

Явление

1. Записывается «явление»…

- *атмосферное явление, физическое явление*
- ИЗОМЕРИЯ - явление, заключающееся в существовании изомеров - соединений, одинаковых по составу и молекулярной массе, но различающихся по строению или расположению атомов в пространстве.

Общий вид правил

1. Записать – <имя отношения> - следующее существительное

2. <имя отношения> - следующее существительное

3. Сложные правила

Примеры правил: обобщение

Род, вид, сорт…

- <Gen> - следующее существительное.

ФИЛЬДЕПЕРС - высший сорт фильдекоса.

ПИДЖИНЫ - тип языков, используемых как средство межэтнического общения в среде разноязычного населения.

Примеры правил: обобщение

Жанр

- Записать - <Gen> - следующее существительное.

МИСТЕРИЯ - жанр средневекового западноевропейского религиозного театра.

Примеры правил: часть

Совокупность

- <Part> - следующее существительное.

АРХИВ - совокупность документов, образовавшихся в результате деятельности учреждений, предприятий и отдельных лиц.

Примеры правил: часть

Скопление

- Записать - <Part> - следующее существительное.

ГАНГЛИЙ - анатомически обособленное скопление нервных клеток , волокон и сопровождающей их ткани .

НО:

ПНЕВМОТОРАКС - скопление воздуха или газов в полости плевры.

Примеры правил: целое

Часть

- <Whole> - следующее существительное.

АЛГЕБРА - часть математики , развивающаяся в связи с задачей о решении алгебраических уравнений.

Примеры правил: целое

Участок

- Записать - <Whole> - следующее существительное.

АНТИКОДОН - участок транспортной РНК, состоящий из трех нуклеотидов.

НО:

ИМЕНИЕ - земельный участок с усадьбой.

Примеры правил: инструмент/назначение

Метод, способ

- <Func> - следующее существительное.

ЗАИЛЕНИЕ - метод мелиорации песчаных земель.

СГРАФФИТО - способ декоративной отделки стен, при котором рисунок процарапывается в верхнем слое штукатурки и обнажается нижний слой, отличающийся по цвету.

Примеры правил: инструмент/назначение

Орудие

- Записать - <Func> - следующее существительное.

ПЕРЕМЕТ - орудие лова рыбы (главным образом хищной).

НО:

артиллерийское орудие

орудие труда

«Сложные» правила

Инструмент, прибор, аппарат…

1. Записать
2. Перейти к следующему предлогу
3. Если это ***для***: Func – следующее существительное.

ФЕН - электрический аппарат для сушки волос.

Другие типы отношений

- Записать - <Other> - следующее существительное.

АБОРТ - прерывание беременности в сроки до 28 недель (то есть до момента, когда возможно рождение жизнеспособного плода).

ХОМИНГ - способность животного возвращаться со значительного расстояния на свой участок обитания, к гнезду, логову и т. д.

Другие типы отношений

характеристика	распространение
признак	переход
свойство	извлечение
число	превращение
показатель	введение
степень	выделение
количество	возникновение
характер	нарушение
масса	прерывание
состояние	развитие
способность	образование
место	увеличение
источник	уменьшение

Правила: резюме

- 18 правил
- 91 опорное слово, для которого существуют правило
- 8484 статей, для которых используются
- 4679 различных опорных слов
- 1978 опорных терминов

Оценка

- Экспертная оценка, 200 словарных статей

- 90% случаев (179 статей) решения совпали с результатами, полученными автоматически
- 21 случай ошибок:
  - 16 случаев – неточности алгоритма
  - 5 случаев – опорное слово отсутствует в тексте определений

- АБРАЗИВНЫЙ ИНСТРУМЕНТ - служит для механической обработки (шлифование, притирка и другие).
- АВОГАДРО ЗАКОН - в равных объемах идеальных газов при одинаковых давлении и температуре содержится одинаковое число молекул.
- АБИТУРИЕНТ - в большинстве стран - оканчивающий среднее учебное заведение.

Пополнение онтологии

Результаты логико-лингвистического анализа представляются в виде таблицы

ПИДЖИН	язык	GEN	*Текст определения*
ЗАИЛЕНИЕ	мелиорация	FUNC	*Текст определения*
ФЕН	аппарат	GEN	*Текст определения*
ФЕН	сушка	FUNC	*Текст определения*
ПАРСЕК	единица	GEN	*Текст определения*

единица измерения

Процедура пополнения

- Указание базового концепта онтологической таксономии
- Формирование энциклопедической выборки
- Добавление терминов выборки
- Постредактирование

Пополнение онтологии: пример

Базовый концепт: СУДНО

Энциклопедическая выборка:

балкер баржа барк барка баркас баркентина брандвахта брандер бриг бригантина бригантина газовоз газотурбоход галера галион глиссер джонка дизель-электроход землесосный снаряд землечерпальный снаряд иол катамаран катамаран кеч килектор клинкер клипер ковчег корабль военный коч кунгас ледокол лихтер лодка нис пароход парусное судно плашкоут понтон приз прорезь рыбоконсервная плавучая база рыбопромысловая база сейнер скампавея струг суда обеспечения судно на воздушной подушке судно на подводных крыльях судно научно-исследовательское тендер теплоход траулер тримаран турбоход шлюп шхуна электроход яхта

С учетом отношения *НИЖЕ*:

ШЛЮПКА: *баркас вельбот гичка туз*

БАРЖА: *шаланда*

Источники

- Gaizauskas, R., Wilks, Y., 1998. Information Extraction: Beyond Document Retrieval - http://www.aclclp.org.tw/clclp/v3n2/v3n2a2.pdf
- Cunningham, H. Information Extraction, Automatic - http://gate.ac.uk/sale/ell2/ie/main.pdf
- Appelt D. Introduction to information extraction - AI Communications 12 (1999) 161–172
- Feldman R., Sanger J. The Text MiningHandbook – Cambridge University Press, 2007
- Dan Jurafsky From Languages to Information. Lecture 15: Relation Extraction - http://www.stanford.edu/class/cs124/
- Dan Jurafsky From Languages to Information. Lecture 7: Named Entity Tagging - http://www.stanford.edu/class/cs124/
- Татьяна Ландо Автоматическое извлечение фактов из текста на примере сервиса Яндекс.Пресс-портреты – http://mathlingvo.ru/nlpseminar/archive/s_32