Поиск упоминаний лиц в новостных текстах с использованием модели условных случайных полей

Алексей Подобряев
Исследовательский центр искусственного интеллекта
Институт программных систем им. А.К.Айламазяна Российской академии наук,
г. Переславль-Залесский
alex@alex.botik.ru

ИСИДА-Т - система извлечения информации из текстов

- Инженерный подход к извлечению информации;
- Пошаговая обработка текста обрабатывающими ресурсами;
- Частичный синтаксический анализ;
- Правила обработки текста на языке PSL;
- **Ресурс знаний** - словарь и связанная с ним база знаний, представляющая собой семантическую сеть концептов (концепты представляются в виде фреймов) и отношений

Коллекция Persons-1000

http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

1000 документов

10 600 вхождений имен

150 000 слов

CRF

- классификация элементов последовательностей,
- учитывает структуру последовательности,
  а не классификация каждого элемента последовательности в отдельности.
- В нашем случае последовательности − это предложения.
- элементы последовательностей – токены, т.е. отдельные слова, знаки препинания и т.п.

- нельзя считать слова предложения независимыми,
- каждое слово необходимо рассматривать в контексте.
- достаточно далекие слова можно считать независимыми.

Признаки

Графематические

- использованный для написания алфавит,
- является ли первая буква прописной,
- есть ли прописные буквы внутри токена,
- является ли данный токен знаком препинания и каким,
- количество символов в токене.

Глобальный

- Доля вхождений данного токена в текст
  (считается по совпадению хотя бы одной из возможных канонических форм).

Признаки

Морфологический

Часть речи — получается с помощью HMM-алгоритма.

Обучен на коллекции национального корпуса русского языка (примерно миллион словоупотреблений)

Признаки

Признаки, основанные на знаниях

- является ли токен аббревиатурой из словаря аббревиатур.
- является ли данный токен возможным характерным префиксом фамилии (например, «д’», «де», «О’», «Тер-»),
- является ли данный токен характерным словом, обычно находящимся перед фамилией («господин», «месье», «товарищ»).

Реализовано в виде конечного автомата, определяемого словарем (окончаний или характерных слов) с учетом морфологии.

Признаки

- Обозначение должностей, профессий, ролей

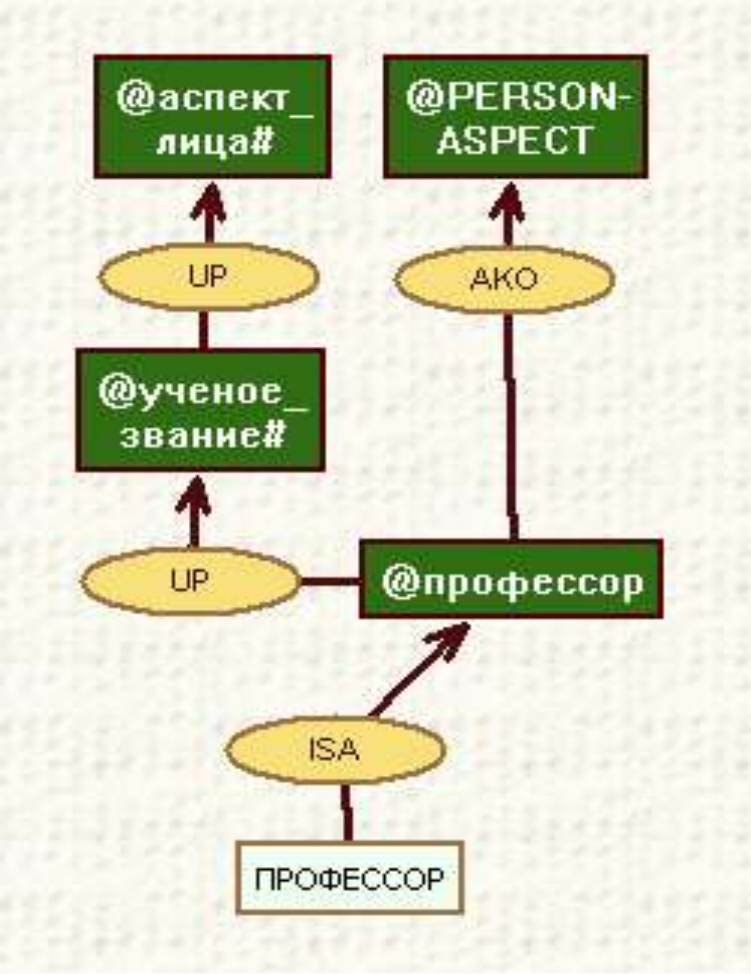
Ресурс знаний − словарь и связанная с ним база знаний, представляющая собой семантическую сеть концептов (концепты представляются в виде фреймов) и отношений

Модель знаний состоит объектов четырех типов:

концепты, экземпляры концептов,

типы отношений, экземпляры отношений.

Признаки



В качестве признака берется наличие направленного пути от данного экземпляра концепта по отношениям типа «ISA», «AKO» и «UP» до концептов «@PERSON_ASPECT» и/или «@аспект_лица#»

Результаты

	С использованием РЗ	Без использования РЗ
Точность	88,1	83,35
Полнота	83,98	77,42
F1-мера	85,99	80,27