Методы и технологии формирования систем формализованных знаний на основе интеллектуального анализа естественно-языковых текстов

Аспирант: Малоземова М.Л.
Научный руководитель: Шишаев М.Г., д.т.н., глав.н.с.

Цель работы

- разработка методов и технологий формирования систем формализованных знаний на основе интеллектуального анализа естественно-языковых текстов

Задачи работы

- Изучение современного состояния проблематики формирования систем формализованных знаний на основе интеллектуального анализа естественно-языковых текстов.
- Обзор современных подходов к формированию систем формализованных знаний.
- Разработка методических основ применения интеллектуального анализа естественно-языковых текстов для формирования систем формализованных знаний.
- Разработка технологии формирования системы формализованных знаний на основе анализа естественно-языковых текстов.
- Экспериментальная проверка эффективности технологии.

Извлечение отношений для обучения онтологий

- Извлечение отношений – обнаружение в текстовых данных отношений между сущностями – <*субъект, отношение, объект*>
- Онтология – концептуальная модель предметной области, разделяемая некоторой группой агентов (люди, организации, ИС)
- Извлечение отношений – один из ключевых этапов процесса обучения онтологий (ontology learning):
  - извлечение таксономических отношений
  - извлечение нетаксономических отношений

Процедура извлечения отношений

1. Формирование дерева синтаксического анализа предложения
2. Обход дерева и формирование n-грамм (комбинаций слова с текущего уровня дерева и связанных с ним слов с дочерних уровней)
3. Выявление среди n-грамм вероятных понятий предметной области (с помощью Word2Vec модели)
4. Извлечение отношений между n-граммами с применением корректирующих процедур:
   - (компания Роснефть, *купила*, context:{}) и (*купила*, акции, context:{}) → (компания Роснефть, акции, context:{купила})
   - (нефтяная *компания*, *компания*, context:{is_kind_of, parent: компания, child: нефтяная компания})

Процедура извлечения отношений

Оценка эффективности

2 тестовых набора – предложения и эталонные отношения:

- 1 набор – для оценки извлечения отношений произвольного типа (500 образцов)
- 2 набор – для оценки извлечения таксономических отношений (75 000 образцов)

Результаты:

1. Оценка извлечения отношений произвольного типа:

   precision = 0.016, fullness = 0.052

2. Оценка извлечения таксономических отношений:

   precision = 0.128, fullness = 0.207

Дальнейшие направления исследования

- Повышение эффективности процедуры извлечения отношений:
  - Анализ влияния метапараметров Word2Vec модели
  - Обучение иного классификатора
- Продолжение исследования в соответствии с поставленными задачами

Спасибо за внимание!