Извлечение отношений из текстов на естественном языке

Чернявский Илья

СПбНИУ ИТМО
Руководитель: А.А.Шалыто

Санкт-Петербург
2012г

Отношения

Отношения - тройки вида «слова отношения»(аргумент1, аргумент2).
Примеры:

- Президент сделал заявление о начале проведения реформ
  «сделал заявление о» («президент» , «начале проведения реформ»)
- Анна вышла замуж за Петра
  «вышла замуж за» («Анна» , «Петра»)

Цель

Предложить метод для извлечения отношений из текстов на русском языке.

Известные решения

Для английского языка:

- TextRunner [M.Banko, O.Etzioni, 2008]
- KnowItAll

Подход к решению

1. Обучить компонент «Chunker»

   Пушкин происходил из разветвлённого дворянского рода.

   [NP Пушкин] [VP происходил] [NP из разветвлённого дворянского рода]
2. Обучить компонент «Extractor»

   «происходил из»(Пушкин, разветвлённого дворянского рода)

Подход к решению

1. Обучить компонент «Chunker»

   Пушкин происходил из разветвлённого дворянского рода.

   [NP Пушкин] [VP происходил] [NP из разветвлённого дворянского рода]
2. Обучить компонент «Extractor»

   «происходил из»(Пушкин, разветвлённого дворянского рода)

Подход к решению

1. Обучить компонент «Chunker»

   Пушкин происходил из разветвлённого дворянского рода.

   [NP Пушкин] [VP происходил] [NP из разветвлённого дворянского рода]
2. Обучить компонент «Extractor»

   «происходил из»(Пушкин, разветвлённого дворянского рода)

Подход к решению

1. Обучить компонент «Chunker»

   Пушкин происходил из разветвлённого дворянского рода.

   [NP Пушкин] [VP происходил]
   [NP из разветвлённого дворянского рода]

2. Обучить компонент «Extractor»

   «происходил из»(Пушкин, разветвлённого дворянского рода)

Подход к решению

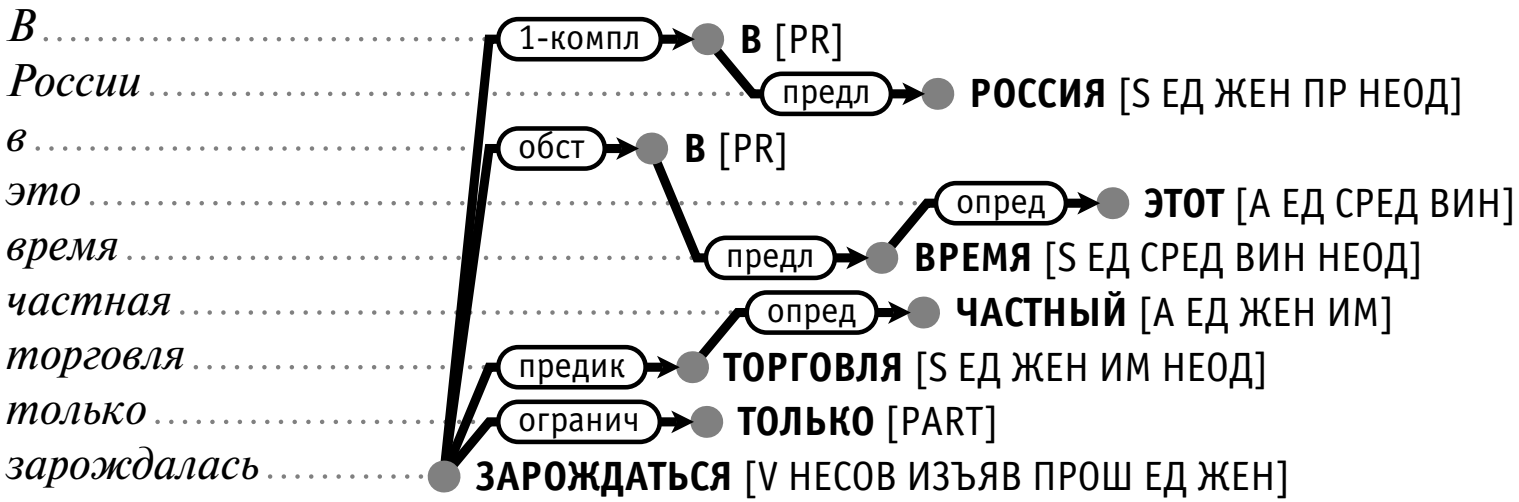
1. Обучить компонент «Chunker»

   Пушкин происходил из разветвлённого дворянского рода.

   [NP Пушкин] [VP происходил] [NP из разветвлённого дворянского рода]
2. Обучить компонент «Extractor»

   «происходил из»(Пушкин, разветвлённого дворянского рода)

- Обучение модели MaxEnt и модели CRF.
- Обучающее множество – 30000 предложений.



Пример обучения chunker'а

Chunker:

...
Мы S B-NP
придумали V B-VP
постоянно ADV B-NP
действующий V I-NP
источник S I-NP
горючей A I-NP
смеси S I-NP
и CONJ I-NP
зажигалку S I-NP
...

Пример обучения chunker'а

Chunker:

...
Мы S B-NP
придумали V B-VP
постоянно ADV B-NP
действующий V I-NP
источник S I-NP
горючей A I-NP
смеси S I-NP
и CONJ I-NP
зажигалку S I-NP
...

Пример обучения chunker'а

Chunker:

...
Мы S B-NP
придумали V B-VP
постоянно ADV B-NP
действующий V I-NP
источник S I-NP
горючей A I-NP
смеси S I-NP
и CONJ I-NP
зажигалку S I-NP
...

Пример обучения chunker'а

Chunker:

...
Мы S B-NP
придумали V B-VP
постоянно ADV B-NP
действующий V I-NP
источник S I-NP
горючей A I-NP
смеси S I-NP
и CONJ I-NP
зажигалку S I-NP
...

Пример обучения extractor'a

[NP Деньги на программы здравоохранения] [VP изымаются в частности] [NP из местного бюджета]

Extractor:

...
Деньги S O
на PR O
программы S O
здравоохранения S O
изымаются V IN
в частности ADV O
из PR IN
местного A O
бюджета S O
...

Пример обучения extractor'а

[NP Деньги на программы здравоохранения] [VP изымаются в частности] [NP из местного бюджета]

Extractor:

...
Деньги S O
на PR O
программы S O
здравоохранения S O
изымаются V IN
в частности ADV O
из PR IN
местного A O
бюджета S O
...

Точность и полнота

Точность (precision) – доля найденных правильных отношений среди всех найденных.
Полнота (recall) – доля найденных правильных отношений среди всех правильных.
F-Measure:

$$FMeasure = \frac{2 * (precision * recall)}{precision + recall}$$

Результаты обучения

Chunker:

- **MaxEnt**
  Precision: 0.7590
  Recall: 0.7887
  F-Measure: 0.7736
- **CRF**
  Precision: 0.6885
  Recall: 0.7315
  F-Measure: 0.7094

Результаты обучения

Extractor:

- **Эвристический классификатор**
  Precision: 0.7352
  Recall: 0.1424
  F-Measure: 0.2386
- **Байесовский классификатор**
  Precision: 0.6229
  Recall: 0.3247
  F-Measure: 0.4269
- Features:
  - Есть ли существительное в им. падеже;
  - Длина chunk'ов;
  - Есть ли глагол;
  - Есть ли предлог;
  - Согласованность по числу;
  - Расстояние между chunk'ами и др.

Пример

Пушкин [**провёл в**] Царскосельском лицее шесть лет
он [**обращается к**] элегиям
в сентябре он [**прибывает в**] Кишинёв
Новый начальник [**снисходительно относился к**] службе Пушкина
Пушкин [**вступает в**] масонскую ложу в Кишеневе
сам [**пишет в**] своём дневнике
поэт [**пытается обратиться к**] российской древности
поэт [**подаёт**] прошение об отставке
Пушкин [**едет через**] Нижний Новгород
Она [**была начата в**] Болдине
Прошение [**было принято с**] отказом
Пушкин [**был связан**] службой в Петербурге
Пушкин [**отказался от**] неё

Применения

- Представление документов в формате RDF (W3C Semantic Web).
- Поисковые системы (RevMiner).
- Автоматическое построение онтологий.

Результаты

- Проведен анализ эффективности моделей MaxEnt и CRF, эвристического и байесовского классификаторов.
- Построен прототип системы извлечения отношений.

Вопросы

Спасибо за внимание!
Вопросы?