Построение словаря сокращений

Амосов Федор

HP Labs

29 мая 2013 г.

Постановка задачи

- ... чем их ранние и средние вещи. Уже первые научно-фантастич. рассказы и повести С. отмечены вниманием к внутр. миру героев, «реалистичностью» деталей, юмором. Разрабатывая преим. жанр социально-филос. фантастики, к-рый в творчестве С. нередко приобретает черты сатирич. гротеска, авторы отстаивают гуманистич. идеал прогресса во имя человека, предостерегают против бездуховного «благоденствия», выступают против любых форм порабощения, размышляют о роли личности в обществе, об ответственности перед будущим. За последние ...
- ... что тормозить тут нельзя. Никакие "анализы через пару дней"и т.п... Праздник грядёт... В этом году день варенья приходится на понедельник. Что вызывает вопрос ...
- Ну чо, Show must go on, ёпт.

Подходы

- Отсечения,
  - по частоте
  - по длине
- Случайно ли появление слова w и точки в виде $w.$?
  - Стандартные алгоритмы проверки гипотез.
    - t–test
    - Pirson test χ^2
  - Алгоритмы, учитывающие специфику сокращений.
    - scaled likelihood ratios (LR)
    - scaled mutual information (MI)

Результаты

Тестирование на данных проекта «Открытый корпус»
($\sim$ 73000 предложений)

%	точность	полнота	F1–мера
по частоте	34	8	13
по длине	55	38	45
t–test	28	21	24
test χ^2	47	31	37
MI	70	42	52
LR	**71**	**46**	**56**

Практическое применение: токенизация (1)

- Брауновский корпус ($\sim$ 60000 предложений)
- Извлечение сокращений (таблица) методом LR.
- На сколько улучшится работа токенайзера FreeLing с разными наборами наилучших сокращений из этого списка?

Практическое применение: токенизация (2)

График (precision от recall)
Таблица улучшений

Q&A

Спасибо за внимание!