

# Построение словаря сокращений

Амосов Федор, СПбГУ

Руководитель: Серебряков Сергей, 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  $\chi^2$
  - ▶ Алгоритмы, учитывающие специфику сокращений
    - ▶ scaled likelihood ratios (LR)
    - ▶ scaled mutual information (MI)

## Тестирование на данных проекта «Открытый корпус» (~ 73000 предложений)

%	точность	полнота	F1–мера
по частоте	34	8	13
по длине	55	38	45
t–test	28	21	24
test $\chi^2$	47	31	37
MI	70	42	52
LR	<b>71</b>	<b>46</b>	<b>56</b>

# Токенизация английских текстов (1)

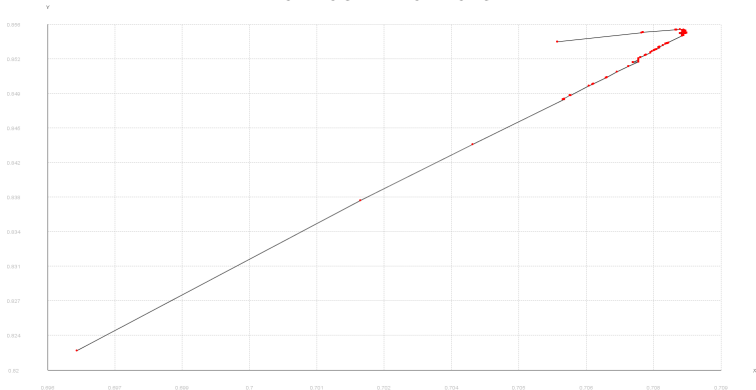
- ▶ Брауновский корпус (~ 60000 предложений)
- ▶ Сокращения, извлеченные методом LR (282 кандидата)

Mr.	314
Mrs.	232
Dr.	84
St.	72
U.S.	70
Jr.	34
etc.	26
Fig.	25
p.m.	24
...	...

- ▶ На сколько улучшится работа токенайзера FreeLing с различными наборами наилучших сокращений из этого списка?

## Токенизация английских текстов (2)

ТОЧНОСТЬ—ПОЛНОТА



%	точность	полнота	F1-мера
без сокращений	79.9	68.8	73.9
по умолчанию	85.1	70.5	77.1
<b>LR</b>	<b>85.5</b>	<b>70.8</b>	<b>77.5</b>

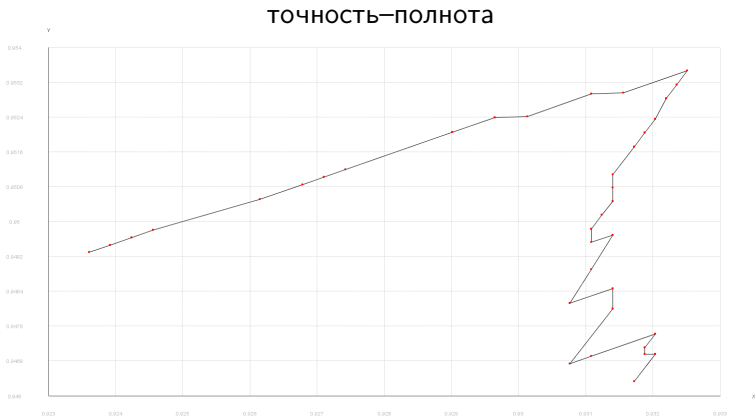
# Токенизация биомедицинских текстов (1)

- ▶ Корпус MedTag (~ 6300 предложений)
- ▶ Сокращения, извлеченные методом LR (111 кандидатов)

al.	8
e.g.	7
i.e.	6
J.	5
S.E.	5
r.h.	3
Sph.	3
Chem.	2
Ass.	2
i.v.	2
Fig.	2
...	...

- ▶ FreeLing?

## Токенизация биомедицинских текстов (2)



%	точность	полнота	F1-мера
без сокращений	94.6	93.2	93.9
по умолчанию	94.0	91.7	92.9
<b>LR</b>	<b>95.3</b>	<b>93.3</b>	<b>94.3</b>



Спасибо за внимание!