

# Исправление опечаток типа склейка-разрезание

Тигунова Анна

Кафедра Анализа  
данных Яндекс

МФТИ, 2015

# Постановка задачи

# Задача

- Запрос: ни с кем

# Задача

- Запрос: ни с кем
- Эталон: ни с кем

# Задача

- Запрос: ни с кем
- Эталон: ни с кем
- Исправление: ни с кем ?

# Задача

- Запрос: ни с кем
- Эталон: ни с кем
- Исправление: ни с кем ?  
ни с кем ?

# Задача

- Запрос: ни с кем
- Эталон: ни с кем
- Исправление: ни с кем ?

ни с кем ?

ни с кем ?

# Типы опечаток

- Опечатки: 10-12% потока\*
  - Ошибки орфографии 66.7%
  - Сегментация 13.7%
  - Раскладка клавиатуры 9.1%
  - Транслитерация 2.1%
  - Другие и комбинации 8.4%

\* Cucerzan S., Brill E. Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users //EMNLP. – 2004. – Т. 4. – С. 293-300



# Типы опечаток

- Опечатки: 10-12% потока\*
  - Ошибки орфографии 66.7%
  - Сегментация 13.7%
  - Раскладка клавиатуры 9.1%
  - Транслитерация 2.1%
  - Другие и комбинации 8.4%

(Ручная разметка, наборы 2011-2013 гг)

\* Cucerzan S., Brill E. Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users //EMNLP. – 2004. – Т. 4. – С. 293-300

# Типы опечаток

Над ними много работали (ML, много признаков...)

- Опечатки: 10-12% потока\*
  - Ошибки орфографии 66.7%
  - Сегментация 13.7%
  - Раскладка клавиатуры 9.1%
  - Транслитерация 2.1%
  - Другие и комбинации 8.4%

Второй по величине

(Ручная разметка, наборы 2011-2013 гг)

\* Cucerzan S., Brill E. Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users //EMNLP. – 2004. – Т. 4. – С. 293-300

# Классический подход

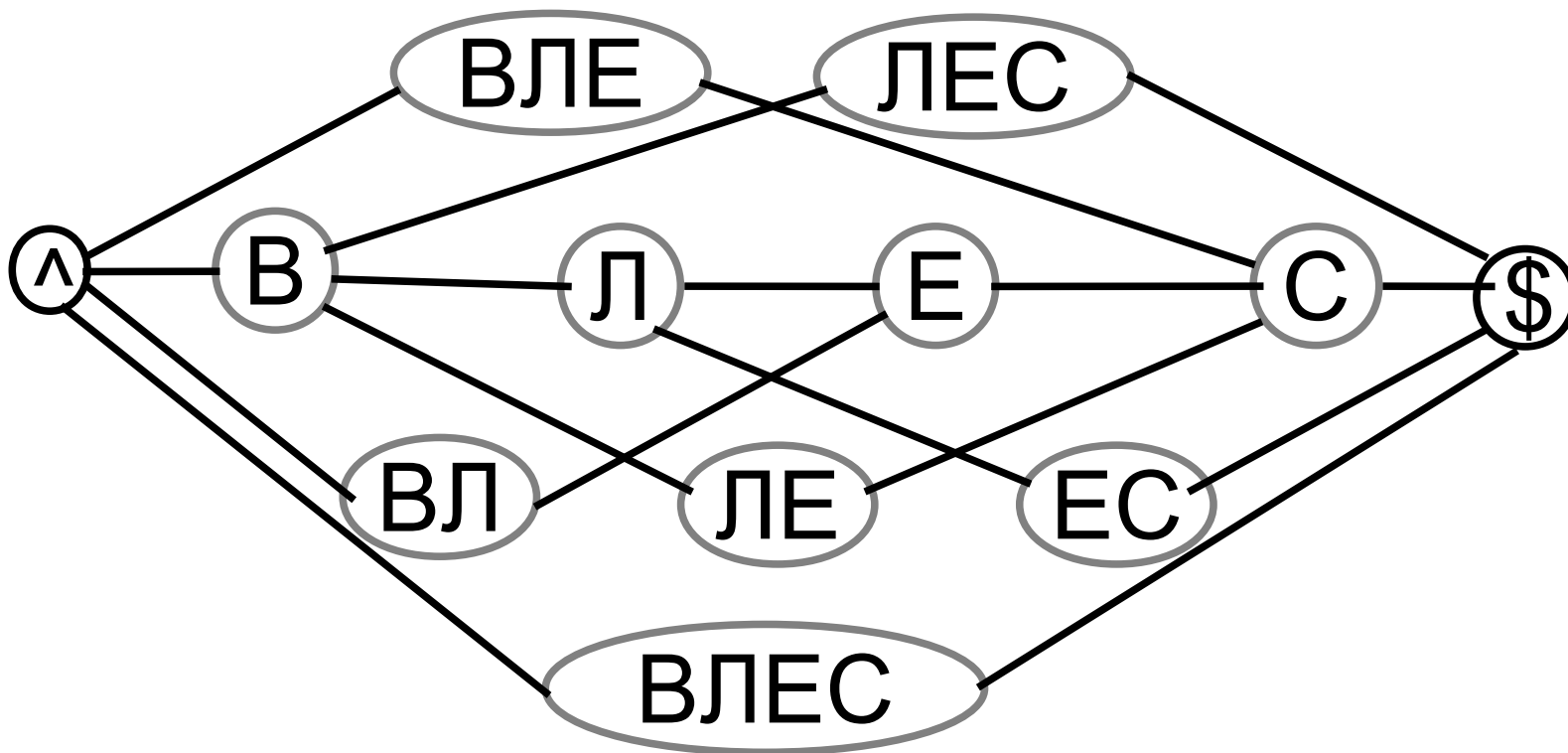
# Классический подход

- Разбиение на слова согласно вероятности по языковой модели
- Вероятностная модель задает вероятность текстовой последовательности в языке
- Ищем  $\operatorname{argmax} P(w_1, \dots, w_n)$
- Описан в \*

\* Russell S., Norvig P., Intelligence A. A modern approach //Artificial Intelligence. Prentice-Hall, Englewood Cliffs. – 1995. – Т. 25.

# Классический подход

- 1) Составляем граф из вариантов разбиения слова



# Классический подход

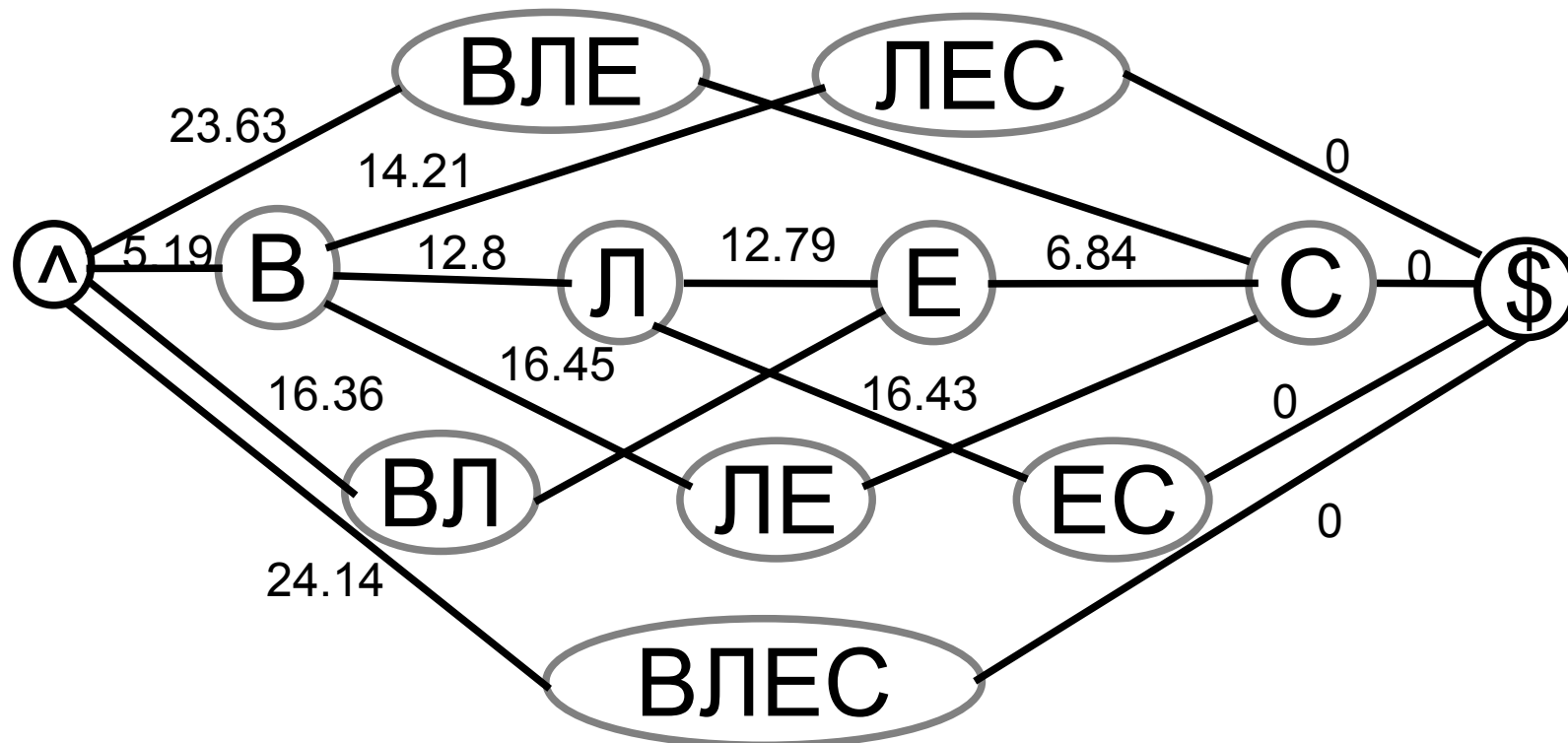
## 2) Биграммный вес по языковой модели:

на ребре  ставим  $-\log P(w_2|w_1)$

- длина пути —  $-\log P(w_1, \dots, w_n)$
- все пути задают возможные сегментации запроса

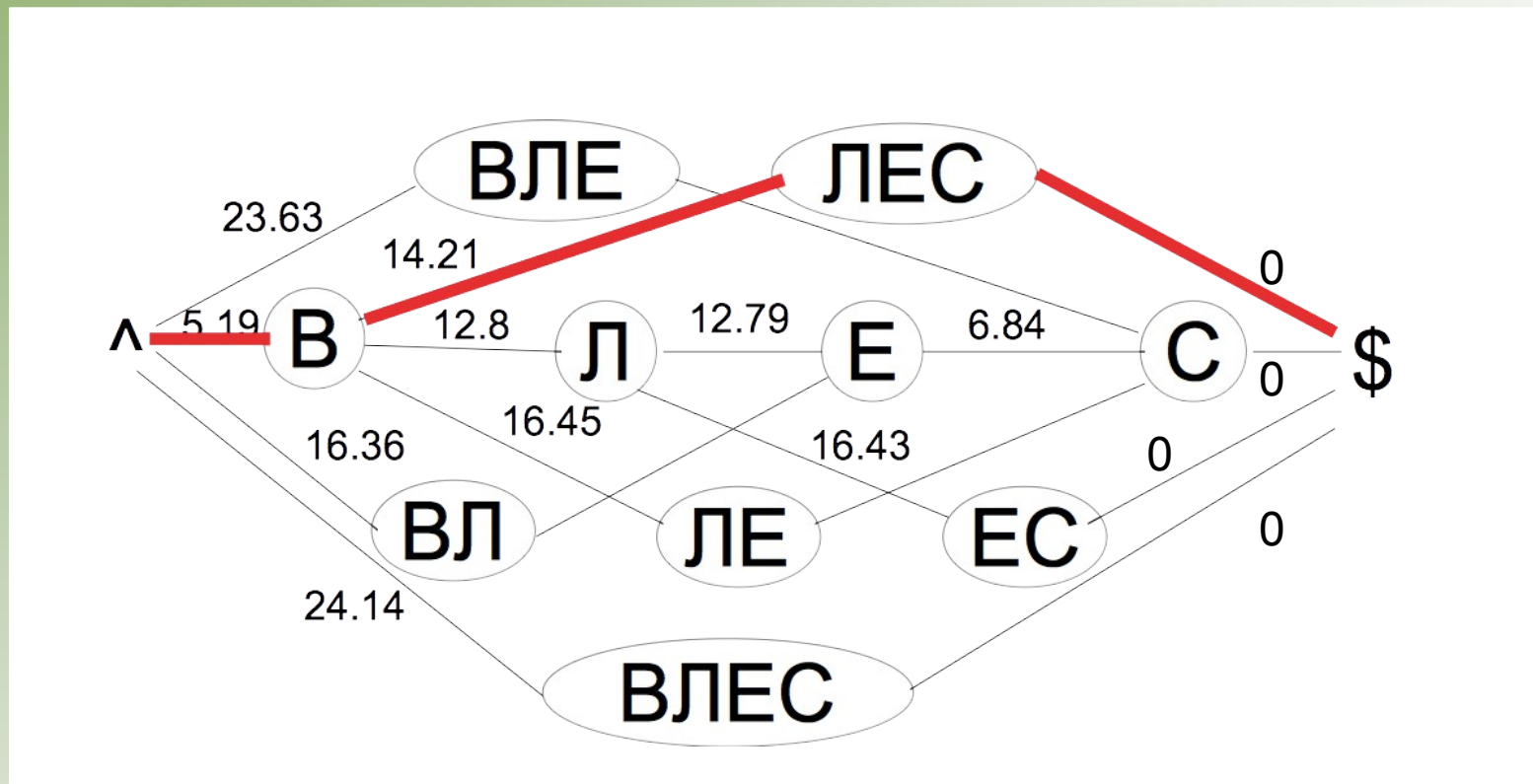
# Классический подход

- 2) Биграммный вес по языковой модели



# Классический подход

- 3) Кратчайший путь в графе (алгоритмы Дейкстры, Витерби...)





# Предлагаемый подход

# Предлагаемый подход: переранжирование

## 1) Генерация гипотез

запрос

точь вточь како ни

# 1) Генерация гипотез

запрос	Гипотезы
точь вточь како ни	<p>точь в точь как они точь в точь какони точь вточь как они точь в точь как он и точьв точь как они точьвточь как они точь в точь как о ни точь в точькак они точь в точь ка кони точь в точь какон и</p> <p>...</p> <p>...</p>

## 2) Вычисление признаков

Гипотезы	Признаки			
точь в точь как они	5	74.1163	56.8264 0	57.3556
точь в точь какони	4	74.1163	56.8264 0	62.9588
точь вточь как они	4	74.1163	56.8264 0	60.8347
точь в точь как он и	6	74.1163	56.8264 0	62.5344
точьв точь как они	4	74.1163	56.8264 0	62.5738
точьвточь как они	3	74.1163	56.8264 0	48.6236
точь в точь как о ни	6	74.1163	56.8264 0	66.7779
точь в точькак они	4	74.1163	56.8264 0	67.0415
точь в точь ка кони	5	74.1163	56.8264 0	68.7043
точь в точь закон и	5	74.1163	56.8264 0	66.9126
точь в точь како ни	5	74.1163	56.8264 0	70.6371
точь в точь ка к они	6	74.1163	56.8264 0	74.3051
т очь в точь как они	6	74.1163	56.8264 0	78.0459
точь в точькак он и	5	74.1163	56.8264 0	72.2203
точь вточь какони	3	74.1163	56.8264 0	66.4379
точь в точь како н и	6	74.1163	56.8264 0	74.8975
точьв точь какони	3	74.1163	56.8264 0	68.1769
точь в точь ка ко ни	7	74.1163	56.8264 0	71.0383

### 3) Машинное обучение

1	"точь в точь как они" , "точь в точь как они"]	5	74.1163	56.8264	0	57.3556
0	"точь в точь как они" , "точь вточь как он и"]	4	74.1163	56.8264	0	62.9588
0	"точь в точь как они" , "точь в точь какони"]	4	74.1163	56.8264	0	60.8347
0	"точь в точь как они" , "точь вточь как они"]	6	74.1163	56.8264	0	62.5344
0	"точь в точь как они" , "точь в точь как он и"]	4	74.1163	56.8264	0	62.5738
0	"точь в точь как они" , "точьвточь как они"]	3	74.1163	56.8264	0	48.6236
0	"точь в точь как они" , "точьвточь как они"]	6	74.1163	56.8264	0	66.7779
0	"точь в точь как они" , "точь в точь как о ни"]	4	74.1163	56.8264	0	67.0415
0	"точь в точь как они" , "точь в точькак они"]	5	74.1163	56.8264	0	68.7043
0	"точь в точь как они" , "точь в точь ка кони"]	5	74.1163	56.8264	0	66.9126
0	"точь в точь как они" , "точь в точь какон и"]	5	74.1163	56.8264	0	70.6371
0	"точь в точь как они" , "точь в точь како ни"]	6	74.1163	56.8264	0	74.3051
0	"точь в точь как они" , "точь в точь ка к они"]	6	74.1163	56.8264	0	78.0459
0	"точь в точь как они" , "т очь в точь как они"]	5	74.1163	56.8264	0	72.2203
0	"точь в точь как они" , "точь в точькак он и"]	3	74.1163	56.8264	0	66.4379
0	"точь в точь как они" , "точь вточь какони"]	6	74.1163	56.8264	0	74.8975
0	"точь в точь как они" , "точь в точь како н и"]	3	74.1163	56.8264	0	68.1769
0	"точь в точь как они" , "точьвточь какони"]	7	74.1163	56.8264	0	71.0383
0	"точь в точь как они" , "точь в точь как о н и"]	5	74.1163	56.8264	0	66.0136
0	"точь в точь как они" , "точь вточь как он и"]	4	74.1163	56.8264	0	72.1835
0	"точь в точь как они" , "точь вточь ка кони"]	6	74.1163	56.8264	0	76.8516
0	"точь в точь как они" , "точь в точь ка кон и"]	2	74.1163	56.8264	0	54.2268
0	"точь в точь как они" , "точь в точь к ак они"]	6	74.1163	56.8264	0	78.0649

## 4) Переранжирование на основе оценок

1	["точь в точь как они" , "точь в точь как они"]	3.4121476579
0	["точь в точь как они" , "точь в точь как они"]	0.581342481
0	["точь в точь как они" , "точь в точь как они"]	0.7685733726
0	["точь в точь как они" , "точь в точь как они"]	0.5955000104
0	["точь в точь как они" , "точь в точь как они"]	0.8610631743
0	["точь в точь как они" , "точь в точь как они"]	0.7473076925
0	["точь в точь как они" , "точь в точь как они"]	0.4688090921
0	["точь в точь как они" , "точь в точь как они"]	0.4578994831
0	["точь в точь как они" , "точь в точь как они"]	1.374624071
0	["точь в точь как они" , "точь в точь как они"]	-2.272490095
0	["точь в точь как они" , "точь в точь как они"]	-2.414676031
0	["точь в точь как они" , "точь в точь как они"]	-2.276600887
0	["точь в точь как они" , "точь в точь как они"]	1.266194126
0	["точь в точь как они" , "точь в точь как они"]	-0.5097612801
0	["точь в точь как они" , "точь в точь как они"]	-0.7546528759
0	["точь в точь как они" , "точь в точь как они"]	-0.9306838672
0	["точь в точь как они" , "точь в точь как они"]	2.087454543
0	["точь в точь как они" , "точь в точь как они"]	-1.823030004
0	["точь в точь как они" , "точь в точь как они"]	-2.733688294
0	["точь в точь как они" , "точь в точь как они"]	-2.733688294
0	["точь в точь как они" , "точь в точь как они"]	-2.976098019
0	["точь в точь как они" , "точь в точь как они"]	-3.000218687
.....		
.....		

# 1. Генерация гипотез

- Заменяем алгоритм так, чтобы он находил 30 лучших исправлений

# 1. Генерация гипотез

- Заменяем алгоритм поиска в графе так, чтобы он находил 30 лучших исправлений
- Мотивация
  - полнота для 1 варианта исправления - 74%
  - полнота среди 30 вариантов исправления - 91%



## 2. Вычисление признаков

- Фрагменты — естественная структура для опечаток типа сегментации
- Находятся выравниванием

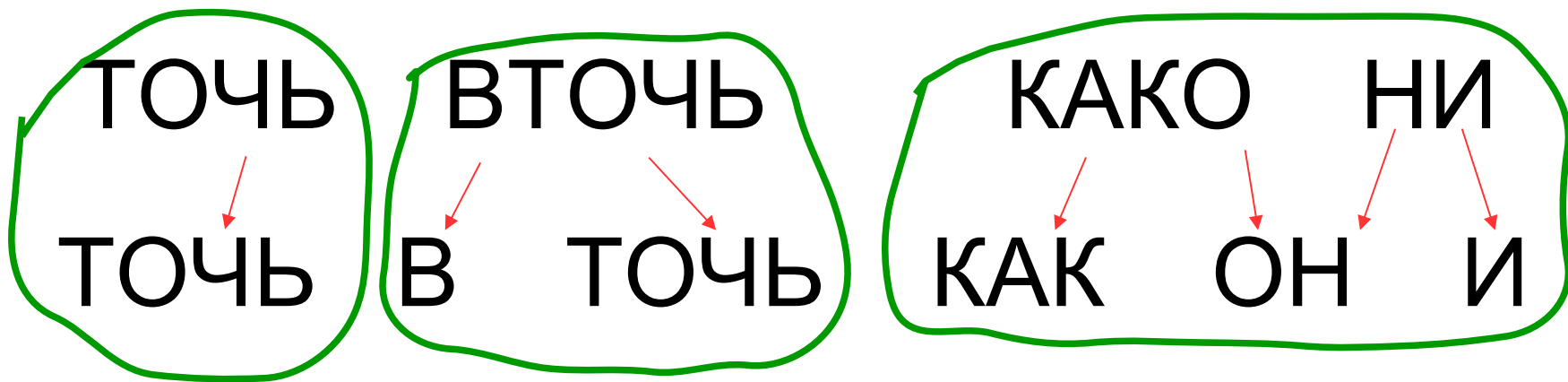
## 2. Вычисление признаков

- Фрагменты — естественная структура для опечаток типа сегментации
- Находятся выравниванием

ТОЧЬ	ВТОЧЬ	КАКО	НИ		
↓	↓	↓	↓		
ТОЧЬ	В	ТОЧЬ	КАК	ОН	И

## 2. Вычисление признаков

- Фрагменты — естественная структура для опечаток типа сегментации
- Находятся выравниванием



# Проблемы

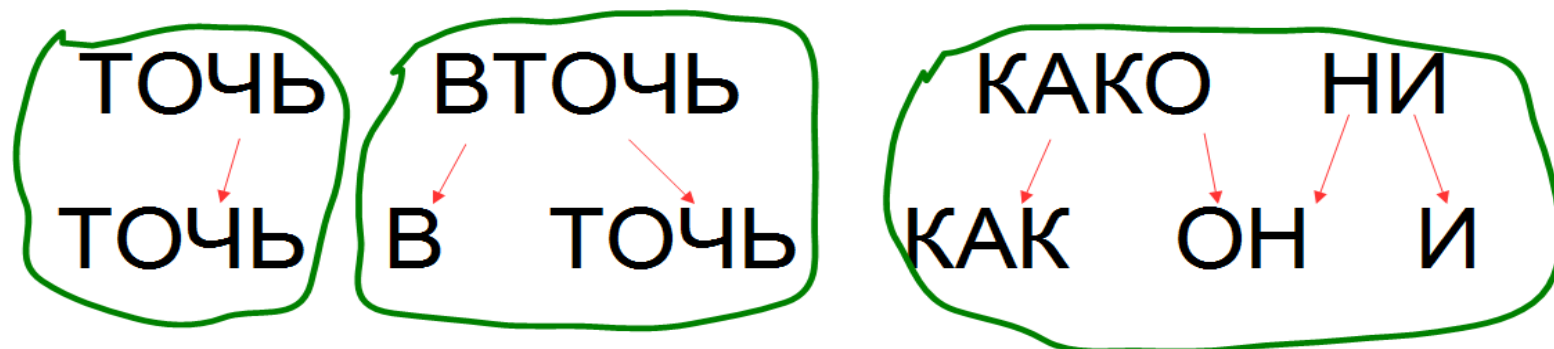
- В пределах одного запроса разное число фрагментов
- Разное число слов в фрагменте
- Сложно придумать признаки на уровне фрагмента

# Решение

- Вычисление признаков по словам
  - Вес по языковой модели
  - Длина слова
  - ...
- Комбинируем внутри фрагмента
- Комбинируем по фрагментам

# Преобразование признаков

Признаки на уровне слов	Признаки на уровне предложения
Вес по языковой модели (LM)	$\sum_s \sum_f LM$
	$\min_s \text{avg}_f LM$
	$\max_s \sum_f LM$
	$\text{avg}_s \sum_f LM$
Длина слова (len)	$\sum_s \sum_f \text{len}$
	$\min_s \max_f \text{len}$
	$\max_s \sum_f \text{len}$



5.72

1.1

5.72

3.3

2.1

0.63

5.72

5.72

3.3

4.9

Вес по языковой модели

Мах внутри фрагмента

Средний по исправлению

- Полный контекст в запросе и исправлении (Благодаря переранжированию)

## 3. Машинное обучение

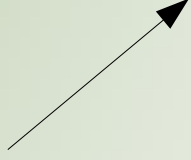
- Задача ранжирования (learning to rank)
- Gradient boosted oblivious decision trees (matrixnet)




### 3. Машинное обучение

- 1.6% опечаток типа сегментации ( $10\% * 16\%$ )

Число  
опечаток  
среди  
запросов



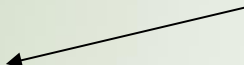
Число опечаток  
типа сегментации  
среди всех  
опечаток



# 3. Машинное обучение

- 1.6% опечаток типа сегментации

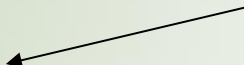
Сильный  
дисбаланс



# 3. Машинное обучение

- 1.6% опечаток типа сегментации

Сильный  
дисбаланс



# 3. Машинное обучение

- 1.6% опечаток типа сегментации

Сильный дисбаланс

У правильный гипотезы 1 если опечатка **была**

1	"точь вточь как они"	"точь в точь как они"	5	74.1163	56.8264	0
0	"точь вточь как они"	"точь вточь как он и"	4	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь какони"	4	74.1163	56.8264	0
0	"точь вточь как они"	"точь вточь как они"	6	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь как он и"	4	74.1163	56.8264	0
0	"точь вточь как они"	"точьв точь как они"	3	74.1163	56.8264	0
0	"точь вточь как они"	"точьвточь как они"	6	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь как о ни"	4	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точькак они"	5	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь ка кони"	5	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь какон и"	5	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь како ни"	6	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь ка к они"	6	74.1163	56.8264	0
0	"точь вточь как они"	"т очь в точь как они"	5	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точькак он и"	3	74.1163	56.8264	0
0	"точь вточь как они"	"точь вточь какони"	6	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь како н и"	3	74.1163	56.8264	0
0	"точь вточь как они"	"точьв точь какони"	7	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь как о н и"	5	74.1163	56.8264	0
0	"точь вточь как они"	"точь вточь как он и"	4	74.1163	56.8264	0
0	"точь вточь как они"	"точь вточь ка кони"	6	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь ка кон и"	2	74.1163	56.8264	0
0	"точь вточь как они"	"точь в точь к ак они"	6	74.1163	56.8264	0

У правильной  
гипотезы  $w < 1$   
если опечатки  
**не было**

0.5	["точь в точь как они" , "точь в точь как они"]	5	74.1163	56.8264	0	57.39
0	["точь в точь как они" , "точь в точь как он и"]	4	74.1163	56.8264	0	62.95
0	["точь в точь как они" , "точь в точь какони"]	4	74.1163	56.8264	0	60.83
0	["точь в точь как они" , "точь вточь как они"]	6	74.1163	56.8264	0	62.53
0	["точь в точь как они" , "точь в точь как он и"]	4	74.1163	56.8264	0	62.57
0	["точь в точь как они" , "точь в точь как они"]	3	74.1163	56.8264	0	48.62
0	["точь в точь как они" , "точь вточь как они"]	6	74.1163	56.8264	0	66.71
0	["точь в точь как они" , "точь в точь как о ни"]	4	74.1163	56.8264	0	67.04
0	["точь в точь как они" , "точь в точь как они"]	5	74.1163	56.8264	0	68.70
0	["точь в точь как они" , "точь в точь ка кони"]	5	74.1163	56.8264	0	66.91
0	["точь в точь как они" , "точь в точь какон и"]	5	74.1163	56.8264	0	70.63
0	["точь в точь как они" , "точь в точь како ни"]	6	74.1163	56.8264	0	74.30
0	["точь в точь как они" , "точь в точь ка к они"]	6	74.1163	56.8264	0	78.04
0	["точь в точь как они" , "т очь в точь как они"]	5	74.1163	56.8264	0	72.22
0	["точь в точь как они" , "точь в точь как он и"]	3	74.1163	56.8264	0	66.43
0	["точь в точь как они" , "точь вточь какони"]	6	74.1163	56.8264	0	74.89
0	["точь в точь как они" , "точь в точь како н и"]	3	74.1163	56.8264	0	68.17
0	["точь в точь как они" , "точь в точь какони"]	7	74.1163	56.8264	0	71.03
0	["точь в точь как они" , "точь в точь как о н и"]	5	74.1163	56.8264	0	66.01
0	["точь в точь как они" , "точь вточь как он и"]	4	74.1163	56.8264	0	72.18
0	["точь в точь как они" , "точь вточь ка кони"]	6	74.1163	56.8264	0	76.85
0	["точь в точь как они" , "точь в точь ка кон и"]	2	74.1163	56.8264	0	54.22
0	["точь в точь как они" , "точь в точь к ак они"]	6	74.1163	56.8264	0	78.06

Таким образом мы варьируем баланс: полнота-точность

# Эксперименты

# Оценка качества

- Точность

$$\text{точность} = \frac{\text{число верно исправленных запросов}}{\text{число исправленных запросов}}$$

- Полнота

$$\text{полнота} = \frac{\text{число верно исправленных запросов}}{\text{число запросов с опечаткой}}$$

# Оценка качества

	Полнота	Точность
Языковая модель	74%	34%
Переранжирование*	77%	34%
Переранжирование**	74%	39%

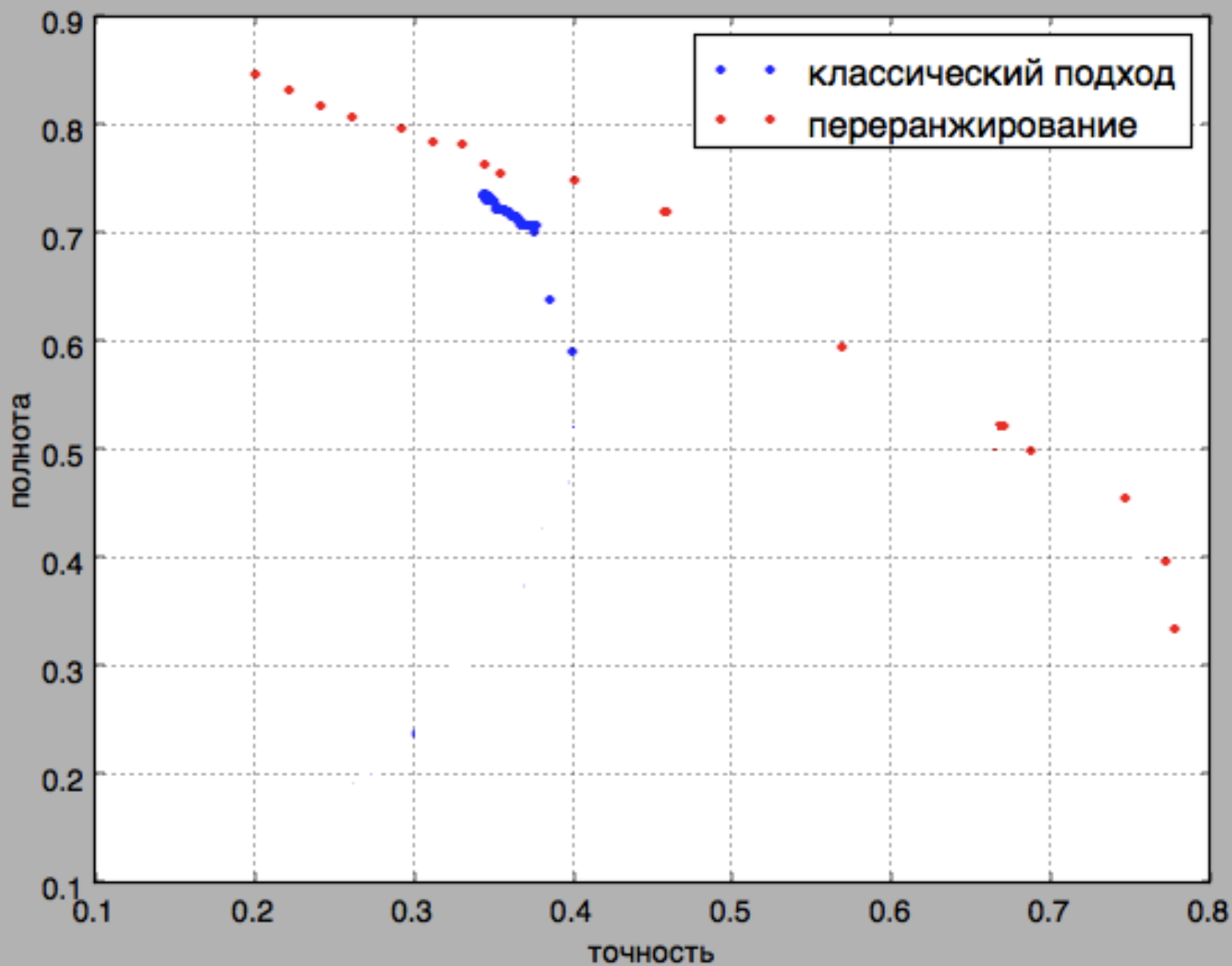
Языковая модель (30 лучших) - 91%

\* Если настраиваем точность как у языковой модели

\*\* Если настраиваем полноту как у языковой модели



# Парето-фронт полнота-точность



# Примеры

	Классический подход	Переранжирование
райодиннадвоих скачать	райодиннадвоих скачать	рай один на двоих скачать
подключение powifi	подключение powifi	подключение по wifi
контрагент	контр агент	контрагент
интернет радиоудлинитель	интернет радио удлинитель	интернет радиоудлинитель
рельеф спортзал	рельеф спорт зал	рельеф спортзал

# Дальнейшие задачи

- встроить в текущую программу проверки орфографии
- придумать больше признаков общего характера
- Придумать больше признаков характерных для поисковых запросов (цифры, URL...)