Яндекс

Автоматическое исправление опечаток в поисковых запросах без учета контекста

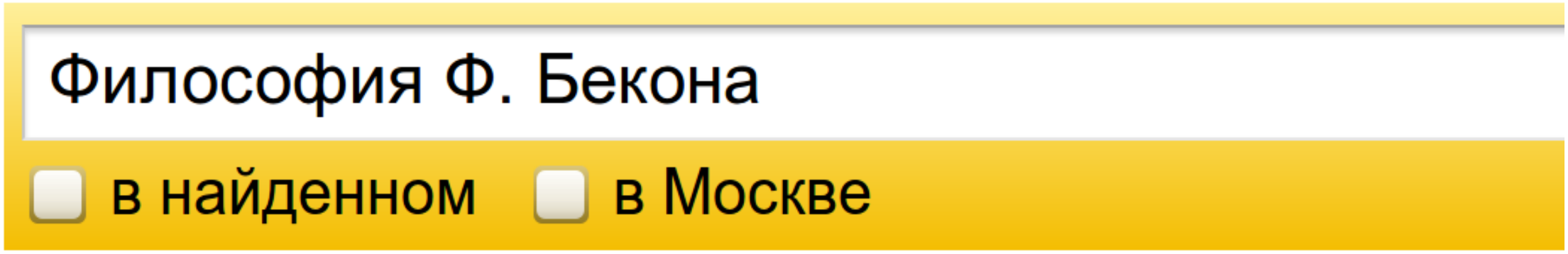
Панина М. Ф., Байтин А. В., Галинская И. Е.

Опечатки в запросах

- Подсказка
- Автозамена
- Типы опечаток

Подсказка

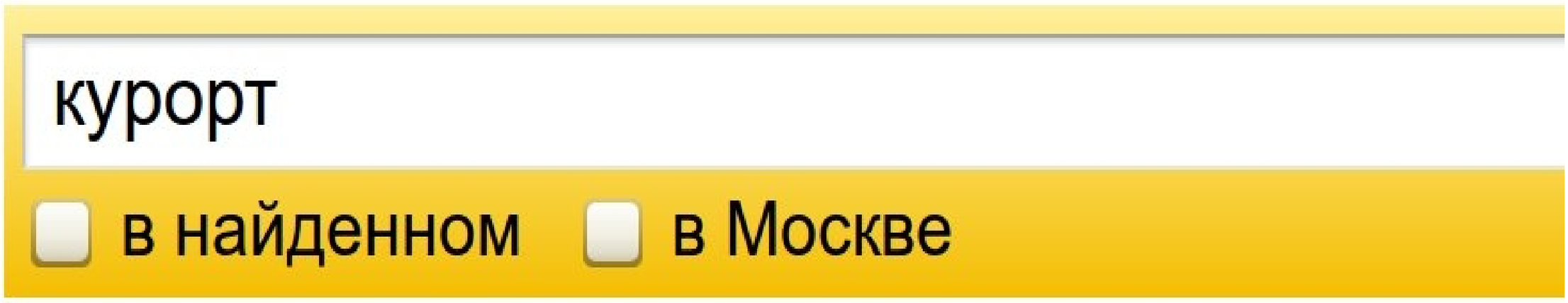
- выдача по исходному запросу
- требуется дополнительный клик



Быть может, вы искали: «**Философия Ф. Бэкона**»

Автозамена

- выдача на исправленный запрос
- высокая точность исправлений



В запросе «куррорт» была исправлена опечатка.

Свойства запросов к yandex.ru

12% с опечатками

Из них **84%** с одной ошибкой

80% содержат только русские слова

Типы опечаток

Ошибки в словах

- Пропуск буквы *(кро**с**овки → кро**сс**овки)*
- Вставка буквы *(фломасте**к**р → фломасте**р**)*
- Замена буквы (эксперемент → *эксперимент*)
- Перестановка букв *(пространтсво → пространство)*

Слитно-раздельное написание

- Пропуск пробела *(купит**ьд**иван → купит**ь д**иван)*
- Вставка пробела *(по**л г**ода → по**лг**ода)*

Раскладка клавиатуры

(rfr cltkfnm cfqn → как сделать сайт)

Транслитерация

- «Русские» слова латинскими буквами

  (kak pekeklju4it' raskladku → как переключить раскладку)
- «Английские» слова кириллическими буквами

  (май нейм из → my name is)

Распределение типов опечаток

Контекст ошибки

- Контекстная классификация
- Эксперимент

Контекстная классификация ошибок

- Контекстно-независимые (КНЗ) – в любом контексте исправляются одинаково (*бук**е**нист → бук**и**нист*)

- Контекстно-зависимые (КЗ) – в разных контекстах исправляются по-разному

- КЗ исправления

*с**кч**ать фильм → ск**а**чать фильм*
*как не с**кч**ать в отпуске → как не ск**у**чать в отпуске*

- КЗ опечатки

*клод м**о**не → клод м**о**не*
*эдуард м**о**не → эдуард м**а**не*

Не являются КЗ:

- **Ошибки согласования (согл)** - слово с опечаткой и его исправление являются разными формами одного и того же слова (*лет**о*** → *лето**м***)

- **Орфоварианты (орфвар)** – оба написания слова употребимы (*к**э**йтлин* → *к**е**йтлин*)

Доля контекстно-зависимых ошибок

Данные:
- из 10 000 запросов выбрали уникальные опечатки типа “ошибки в словах” и их исправления *(q = куппить, c = купить)*

- вручную удалили слова с ошибками СОГЛ и ОРФВАР

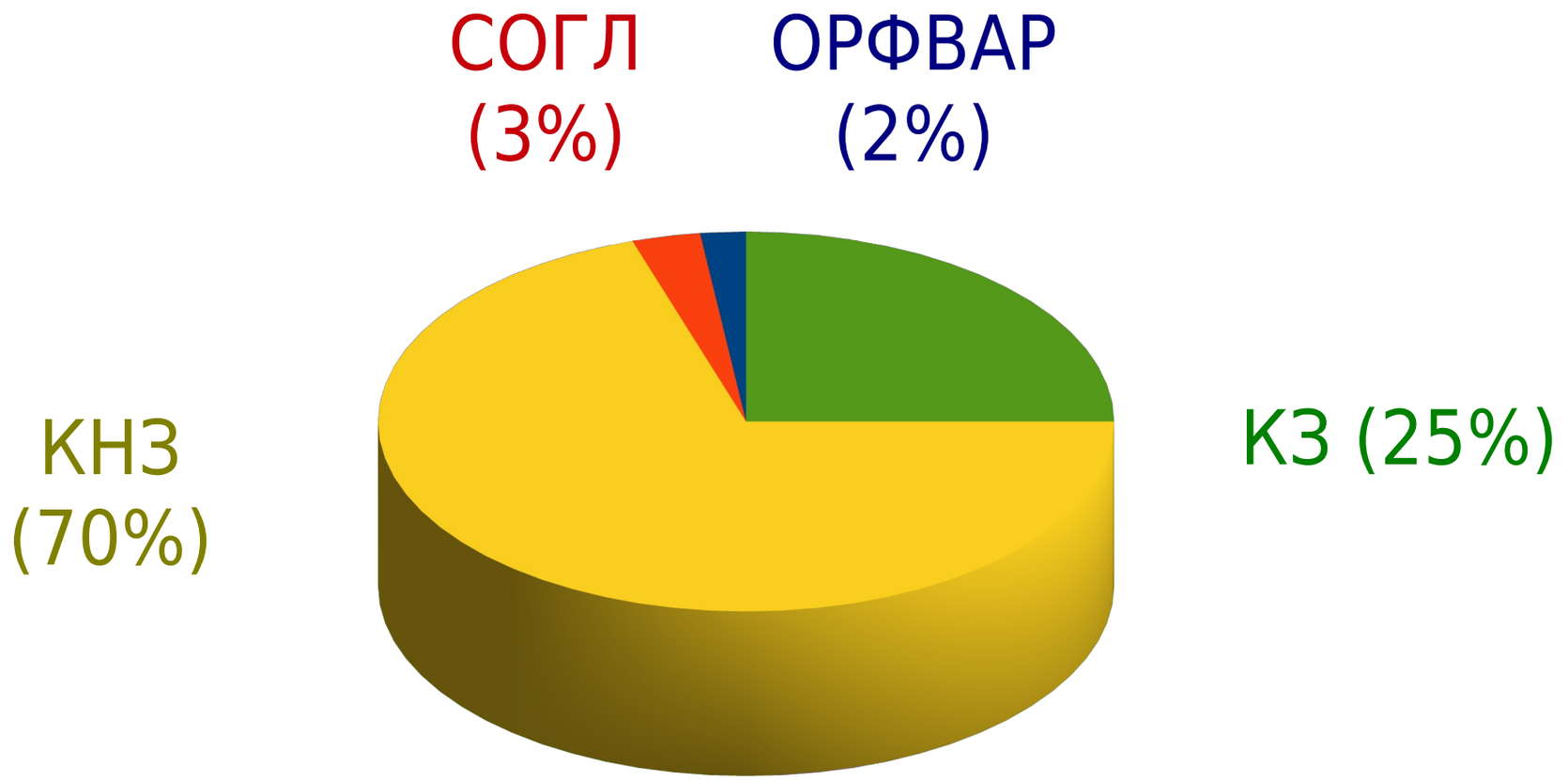
Эксперимент:
- аналитик исправил ошибки в словах, не зная их исправлений в запросах

- сравнили исправления слов без учета запроса с исправлениями с учетом запроса

Примеры

Опечатка	Исправление вне запроса	Исправления в запросе	Тип ошибки
клон	*клон*	*африканский слон*	КЗ
нгород	*город*	*сад и огород*	КЗ
фмгур	*фигур*	*правильная фигура*	КН
прогулкв	*прогулка*	*прогулка по реке*	КН
вкусные	(*отфильтровано*)	*вкусный суп*	СОГЛ
кэйтлин	*(отфильтровано)*	*кейтлин*	ОРФВАР

Доля контекстно-зависимых ошибок



- контекст не влияет на исправление **75%** опечаток

Алгоритм оценки надежности исправления запроса

- Описание алгоритма
- Описание признаков

Алгоритм

Вход: Q → C

куппить слона → *купить слона*

недороог → *недорого*

Выход: Оценка надежности: “надежное” или “ненадежное” исправление

Ограничения:

- тип ошибки “ошибки в словах”
- порог по точности
- язык ru/en

Модуль исправления опечаток

$$C' = argmax\, P(Q|C) \cdot P(C)$$

Q — исходный запрос

C — исправление

$P(Q|C)$ — вероятность трансформации Q в C (модель ошибки)

$P(C)$ — вероятность запроса (модель языка)

Основные шаги алгоритма

Выравнивание

Выделяем пары слов опечатка – исправление (q→c)
(куппить → купить, недорооог → недорого)

Фильтрация

- совпадающая лемма
- короткое слово *(длина < 4)*
- язык запроса *(ru/en)*

Классификация

Для каждой пары q → c:
- вычисляем признаки
- определяем принадлежность к классу “надежные” исправления (логистическая регрессия)

Признаки

Вес по языковой модели:

- Словарная 3-грамная модель
- Буквенная 3-грамная модель

Данные:

Запросы к Яндексу за полгода

Словарность:

Бинарный признак

Данные:

Ручные морфологические словари

Признаки

Вероятность написания слова с заглавной буквы:

Отношение частоты написания слова с заглавной буквы к общей частоте слова

<u>Данные:</u>

веб корпус
(100M документов)

Язык слова:

бинарный признак — en/ru

Длина слова

Признаки

Дистанция редактирования:

Взвешенное расстояние Левенштейна

Данные:
Запросы к Яндексу за
полгода

Пример

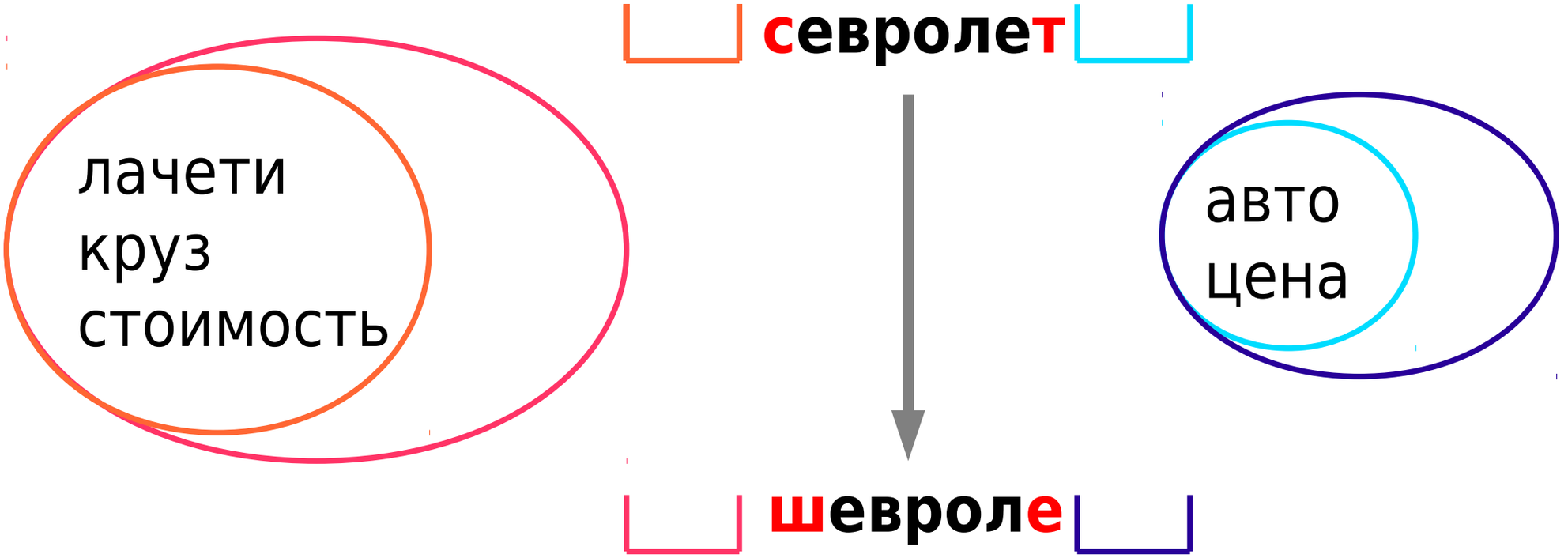
опечатка	исправление	Левенштейн	Взвешенный Левенштейн
высотский	высоцкий	2	3
утка	будка	2	23

Признаки. Взаимный контекст

Опечатка

Левый контекст

Правый контекст

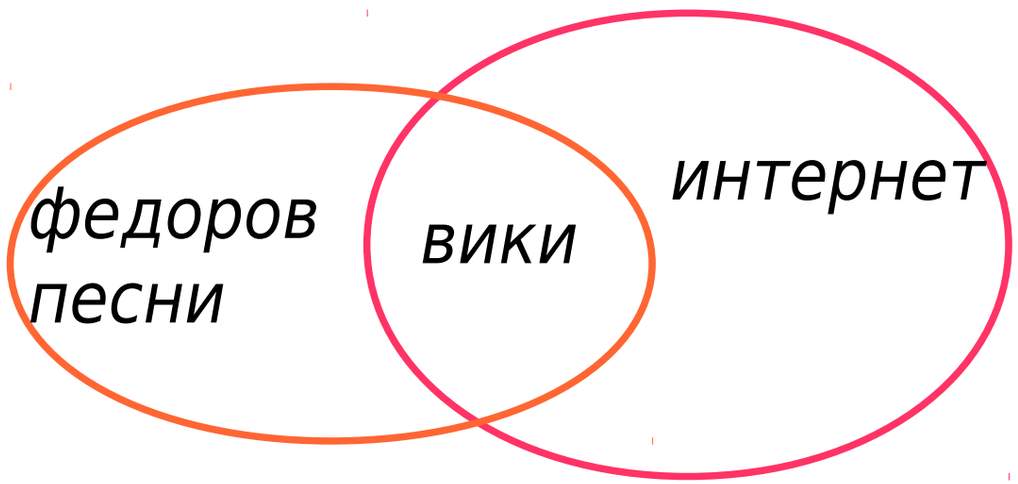


Признаки. Взаимный контекст

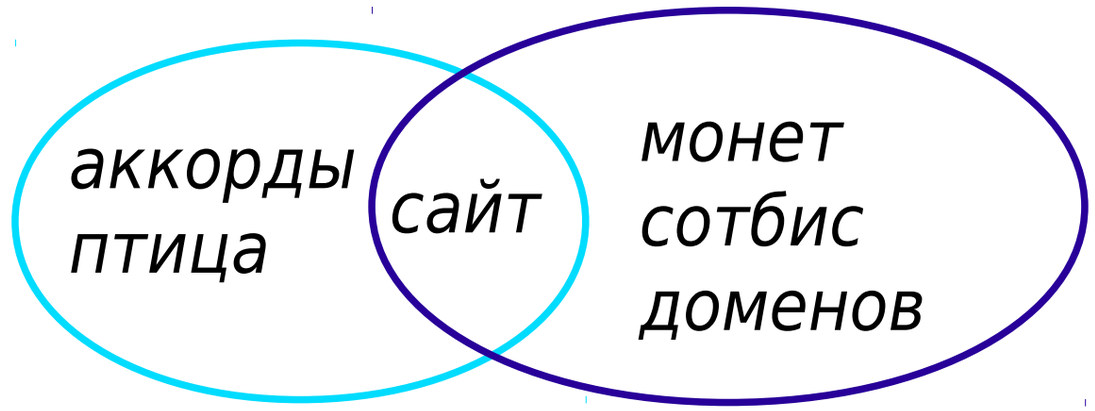
Не опечатка

аукцыон

Левый контекст



Правый контекст



аукцион

Признаки. Взаимный контекст

$$P_c(w_2|w_1) = \sum \frac{P(w|w_1)}{P(w)} \cdot P(w|w_2) P(w_2)$$

w_1 опечатка

w_2 исправление

w общий контекст

$P_c(w_1|w_2)$ вероятность того, что w_1 опечатка w_2

M. Li and Y. Zhang. *Exploring distributional similarity based models for query spelling correction*. ACL'06

Эксперименты

- Данные
- Результаты
- Разбор ошибок

Данные для эксперимента

- 2150 запросов с опечатками типа “ошибки в словах” размечены вручную

- соотношение классов “надежные”:”ненадежные”~1:4

- 16% коротких слов и 2% с одинаковыми леммами

- метрика качества Полнота (Точность фиксирована)

Эксперименты

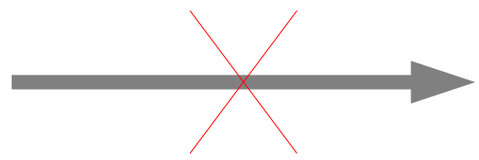
	Порог по точности 90%	Порог по точности 95%
Базовый набор признаков*	36,4%	21,4%
Полный набор признаков	77,3%	54,9%

* базовый набор признаков:
- модель языка
- модель ошибок

Примеры ошибок

Ошибка
false-positive

савенок
фамилия



совенок

<u>Основные проблемы:</u>

- небольшая дистанция редактирования

- частотная опечатка vs редкое слово (ложное срабатывание “взаимного контекста”, редкого слова нет в словаре)

Ошибка
false-negatives

Гипнос бог → Гипноз бог

- опечатка и исправление - словарные слова
- контекстно-зависимая ошибка

Выводы

- большая часть ошибок не зависит от контекста и может быть исправлены автоматически

- классический набор статистических и лексических признаков обеспечивает достаточно высокую полноту при заданной точности

Яндекс

Спасибо