Исправление ошибок написания

Spelling correction

Использованы картинки А. Байтина (Яндекс. Группа опечаток и машинного перевода)

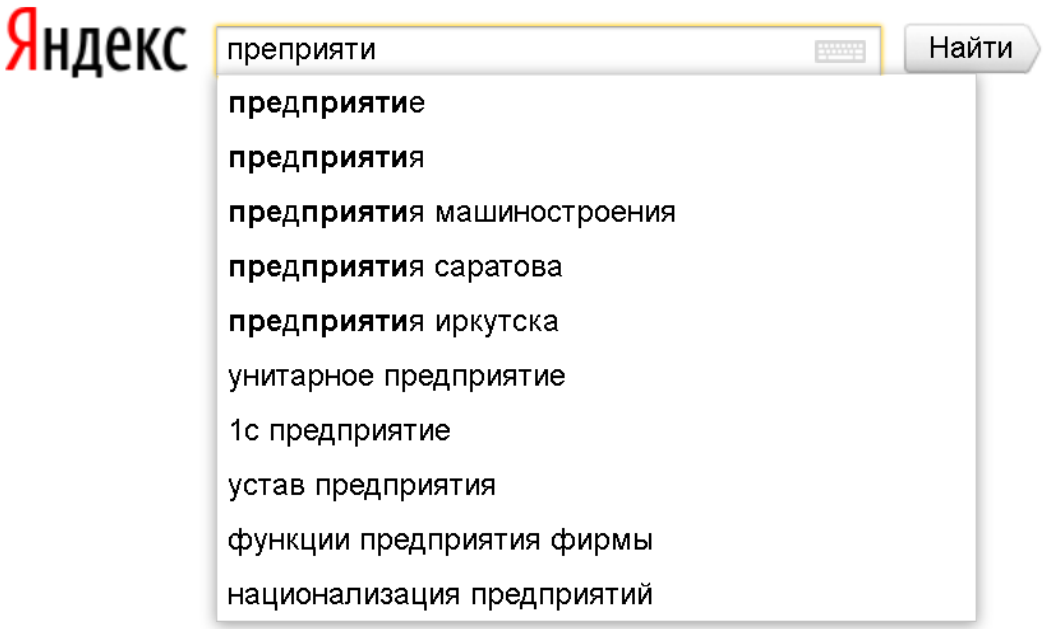
Задачи исправления ошибок написания

- Обнаружение ошибки
- Исправление ошибки
  - Автокоррекция
  - Предложение подсказки
  - Списки подсказок

Сделать стартовой

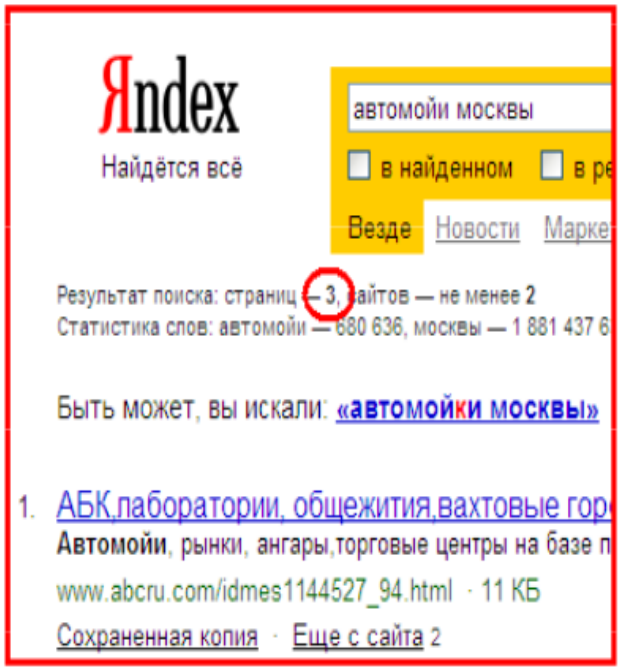
Войти в почту

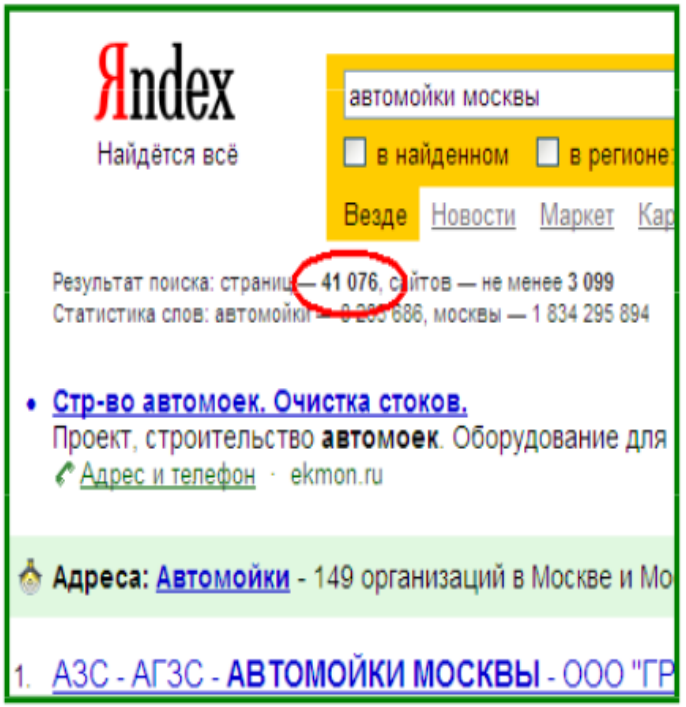
Яндекс: 15% запросов содержат ошибки



Запрос: автомойи москвы

[автомойки москвы]





(Байтин, 2009)

Типичные ошибки Поиск@Mail.ru:

- набор запроса в неправильной раскладке клавиатуры (например «zyltrc» вместо «яндекс»);
- недописанные запросы (например, «vko» или «знакомс»);
- ввод адреса сайта в поисковую строку вместо адреса (поисковую строку с адресной путают в каждом десятом запросе);
- самая распространенная опечатка: «одноклассники» (с одной буквой «с»); ее совершают в 3-5% случаев, но из-за высокой частоты данного запроса она лидирует;
- а самая частая опечатка: «агенство» (без буквы «т»); это слово вводится с ошибкой примерно в 30% случаев.
-

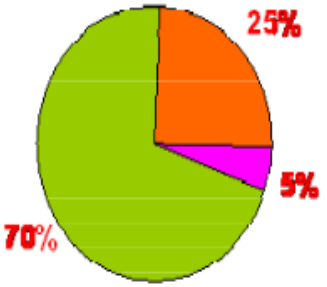
Источники ошибок в запросах

- **Случайные клавиатурные ошибки**

  приклбчение вместо *приключение*

- **Систематические когнитивные ошибки**
-
  - **Фонетические ошибки**
    игипет вместо *египет*
  - **Слитно-раздельное написание**
    фото ателье вместо *фотоателье*
  - **Заимствованные слова**
    фитнесс вместо *фитнес*
  - **Названия** фирм, брендов и т.п.
    лачетти вместо *лацетти*
  - **Марки** товаров
    Нокиа i_850_w вместо *Нокиа i850w*

Типы ошибок



- **Искажения в отдельных словах**
  - исчезновение буквы (*афавит* вместо *алфавит*)
  - вставка лишней буквы (*вврач* вместо *врач*)
  - замена буквы (*барабае* вместо *барабан*)
  - перестановка соседних букв (*притнер* вместо *принтер*)
- **Искажения в последовательности слов**
  - вставка пробела (*Вели кий Новгород* вместо *Великий Новгород*)
  - пропуск пробела (*КрасныйОктябрь* вместо *Красный Октябрь*)
- **Искажение смысла запроса**
  - контекстные ошибки (*вокруг меха* вместо *вокруг смеха*)
- **Латинский алфавит вместо русского**
  - транслитерация (*varezhka* вместо *варежка*)
- **Искаженная кодировка**
  - использование неправильной раскладки клавиатуры (*bnfkbz* вместо *италия*)

Типы ошибок написания

- Переход в несуществующее слово
  - Предприятие -> преприятие
- Переход в существующее слово:
  - Вокруг смеха -> вокруг меха

Данные для исправления ошибок

- 1. Правильных слов больше
  - Частотность опечатки обычно на порядок меньше частоты правильного слова
- 2. Ошибки повторяются
  - Повторяемость клавиатурных и фонетических ошибок очень высокая
- 3. Ошибки зависят от контекста
  - Корректность употребления слова зависит от контекста
- Запросы
  - Информация о частотности слов
  - Информация о сочетаемости слов
  - Информация о переформулировках слов

Переформулировки запросов

- Вот как исправляют запросы сами пользователи:

райфаззен	-> райффайзен
ретеил	-> ритейл
колбассоф	-> колбасофф
крбина	-> корбина
тинидозол	-> тинидазол
скаэкспресс	-> скайэкспресс

- Выбираем такие пары из пользовательских сессий и складываем их в словарь замен.

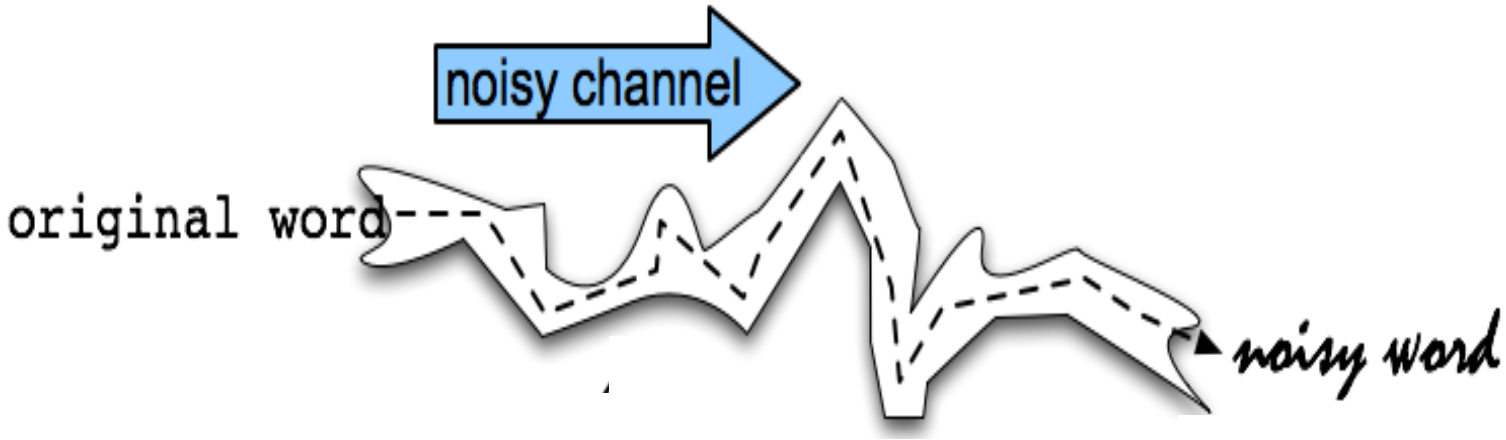
Несловарные ошибки

- Распознавание:
  - Любое слово, которое не найдено в словаре, - это ошибка
  - Чем больше словарь, тем лучше
- Исправление:
  - Порождение кандидатов: реальных слов, которые похожи на ошибку
  - Выбор наилучшего слова:
    - Сходство
      - По написанию - наикратчайшее редакционное расстояние (=расстояние Левенштейна)
      - По звучанию
    - Вероятность по методу зашумленного канала (noisy channel)

Ошибочное использование существующего слова

- Для каждого слова порождается множество кандидатов:
  - С похожим произношением
  - С похожим написанием
  - Текущее слово *w* включается в множество
- Выбор лучшего кандидата
  - Подход зашумленного канала
  - Использование контекста
  - *Flying form Heathrow to LAX → Flying from Heathrow to LAX*

Интуиция: зашумленный канал (noisy channel)



Noisy Channel + Правило Байеса

- Мы видим неправильно написанное слово x
- Найдем правильное слово $\hat{w}$

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid x)$$

$$= \underset{w \in V}{\operatorname{argmax}} \frac{P(x \mid w) P(w)}{P(x)}$$

$$= \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)$$

История: Модель Noisy channel предложена около 1990

- **IBM**
  - Mays, Eric, Fred J. Damerau and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing and Management*, 23(5), 517–522
- **AT&T Bell Labs**
  - Kernighan, Mark D., Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. Proceedings of COLING 1990, 205-210

Минимальное редакционное расстояние (расстояние Левенштейна)

- Измеряется в количестве минимальных редакционных операций, которые требуются для преобразования одного слова в другое:
- Операции:
  - Вставка
  - Удаление
  - Замена
  - Смена порядка расположения двух соседних букв

Расстояние 1 от слова *acress*

Error	Candidate Correction	Correct Letter	Error Letter	Type
acress	actress	t	-	deletion
acress	cress	-	a	insertion
acress	caress	ca	ac	transposition
acress	access	c	r	substitution
acress	across	o	e	substitution
acress	acres	-	s	insertion

Порождение кандидатов

- 80% ошибок находятся на расстоянии 1
- Почти все ошибки на расстоянии 2

- Также позволяется вставка пробела или дефиса
  - thisidea → this idea
  - inlaw → in-law

Из множества кандидатов нужно отобрать лучшего кандидата

Предположим, что список кандидатов создан. Вернемся к правилу Байеса

- Мы наблюдаем неправильно написанное слово
- Нужно найти правильное слово $\hat{w}$

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid x)$$

$$= \underset{w \in V}{\operatorname{argmax}} \frac{P(x \mid w) P(w)}{P(x)}$$

$$= \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)$$

Что такое $P(w)$?

Языковая статистическая модель

- Нужно собрать большой корпус.
- Пусть *C(w)* = # количество вхождений *w*

$$P(w) = \frac{C(w)}{T}$$

- Для запросов – корпусом может быть множество всех заданных пользователем запросов

Априорная вероятность униграмм

Частоты из 404,253,213 слов в современном корпусе английского языка (COCA)

word	Frequency of word	*P(w)*
actress	9,321	.0000230573
cress	220	.0000005442
caress	686	.0000016969
access	37,038	.0000916207
across	120,844	.0002989314
acres	12,874	.0000318463

Вероятность Channel model

- Вероятность ошибки, вероятность редактирования
- *Kernighan, Church, Gale 1990*

- *Ошибочное слово* $x = x_1, x_2, x_3 \ldots x_m$
- *Правильное слово* $w = w_1, w_2, w_3, \ldots, w_n$

- $P(x|w)$ = Вероятность перехода (редактирования)
  - (удаления/вставки/замена/перестановки)

Вычисление вероятности ошибки: confusion “matrix”

del[x,y]: количество (xy написано как x)

ins[x,y]: количество(x написано как xy)

sub[x,y]: количество(y написано как x)

trans[x,y]: количество(xy написано как yx)

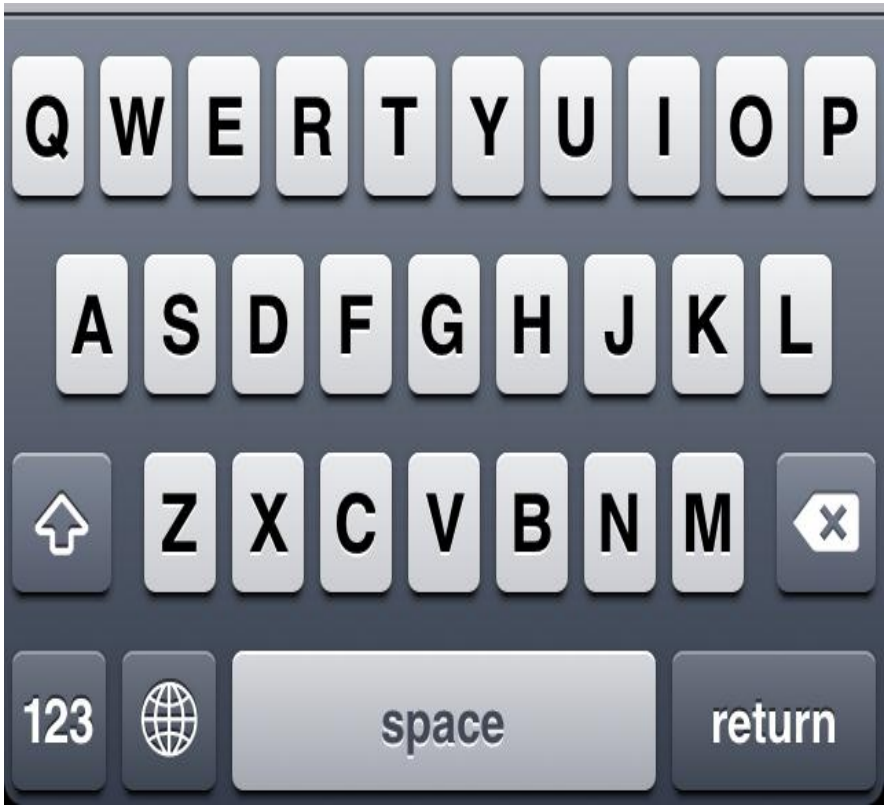
Вставка и удаление должны вычисляться как условные вероятности от предыдущего символа

Матрица количества замен Y на X

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Соседние клавиши на клавиатуре



Channel model

Kernighan, Church, Gale 1990

$$P(x|w) = \begin{cases} \frac{\text{del}[w_{i-1}, w_i]}{\text{count}[w_{i-1} w_i]}, & \text{if deletion} \\ \frac{\text{ins}[w_{i-1}, x_i]}{\text{count}[w_{i-1}]}, & \text{if insertion} \\ \frac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, & \text{if substitution} \\ \frac{\text{trans}[w_i, w_{i+1}]}{\text{count}[w_i w_{i+1}]}, & \text{if transposition} \end{cases}$$

Сглаживание: Правило Лапласа (Add-1)

- Если использовать только матрицу частот ошибок, то некоторые ошибки окажутся невозможными
- Поскольку их вероятность равна 0
- Простое решение: добавить 1, где |A| это алфавит символов и нормализовать

$$\text{If substitution, } P(x|w) = \frac{\text{sub}[x,w]+1}{\text{count}[w]+A}$$

Channel model for acress

Candidate Correction	Correct Letter	Error Letter	*x|w*	*P(x|w)*
actress	t	-	c|ct	.000117
cress	-	a	a|#	.00000144
caress	ca	ac	ac|ca	.00000164
access	c	r	r|c	.000000209
across	o	e	e|o	.0000093
acres	-	s	es|e	.0000321

Noisy channel probability for acress

Candidate Correction	Correct Letter	Error Letter	*x|w*	*P(x|w)*	*P(w)*	10^9 * *P(x|w)* * *P(w)*
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	.00078
caress	ca	ac	ac|ca	.00000164	.00000170	.0028
access	c	r	r|c	.000000209	.0000916	.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Noisy channel probability for acress

Candidate Correction	Correct Letter	Error Letter	*x|w*	*P(x|w)*	*P(w)*	10^9 **P(x|w)P(w)*
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	.00078
caress	ca	ac	ac|ca	.00000164	.00000170	.0028
access	c	r	r|c	.000000209	.0000916	.019
across	**o**	**e**	**e|o**	**.0000093**	**.000299**	**2.8**
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Учет контекста

- В итоге, наиболее вероятные исходные слова *actress* или *across*
- Учет контекста:
  - Учет условных вероятностей появления одного слова после другого

    $P(w_1 \ldots w_n) = P(w_1)P(w_2|w_1)\ldots P(w_n|w_{n-1})$
  - Вероятности насчитываются на некотором корпусе. Нужно насчитать:
    - $P(w_i)$ – вероятность униграмм
    - $P(w_n|w_{n-1})$ – вероятность биграмм

Подсчет вероятностей

- Для униграмм P(w) всегда ненулевое
  - Поскольку наш словарь построен на текстовой коллекции
- Но $P(w_k|w_{k-1})$ может быть нулевым.
- Нужно сглаживание
  - Можно применить сглаживание add-1 (как раньше в методе Байеса)
- Можно применить другой вид сглаживания:

$$P_{li}(w_k|w_{k-1}) = \lambda P_{uni}(w_к) + (1-\lambda)P_{bi}(w_k|w_{k-1})$$

$$P_{bi}(w_k|w_{k-1}) = C(w_k|w_{k-1}) / C(w_{k-1})$$

Учет биграмм

- “a stellar and versatile **acress** whose combination of sass and glamour…”
- Частоты из корпуса современного американского английского со сглаживанием add-1
- P(actress|versatile)=.000021
- P(across|versatile) =.000021

P(whose|actress) = .0010

P(whose|across) = .000006

- P(“versatile actress whose”) = .000021*.0010 = 210 x10^{-10}
- P(“versatile across whose”) = .000021*.000006 = 1 x10^{-10}

Учет биграмм

- “a stellar and versatile **acress** whose combination of sass and glamour…”
- Частоты из корпуса современного американского английского со сглаживанием add-1
- P(actress|versatile)=.000021 P(whose|actress) = .0010
- P(across|versatile) =.000021 P(whose|across) = .000006

- **P(“versatile actress whose”) = .000021*.0010 = 210 x10^{-10}**
- P(“versatile across whose”) = .000021*.000006 = 1 x10^{-10}

Исправление ошибок переходов в существующие слова

Ошибочные слова

- …leaving in about fifteen ***minuets*** to go to her house.
- The design ***an*** construction of the system…
- Can they ***lave*** him my messages?
- The study was conducted mainly ***be*** John Black.

- 25-40% ошибок написания – реальные слова Kukich 1992

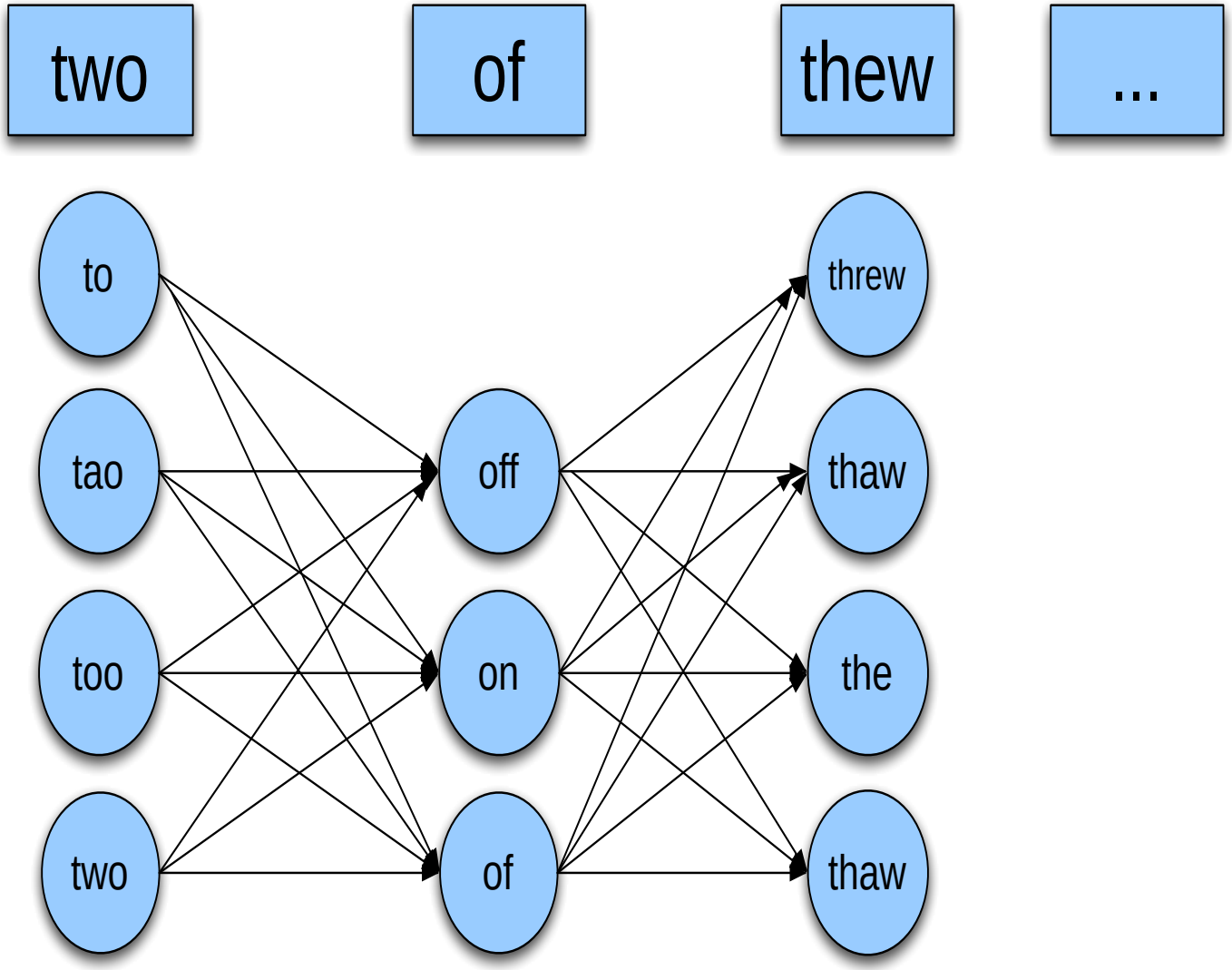
Решение проблемы ошибочных слов

- Для каждого предложения (фразы, запроса …)
  - Порождение списка кандидатов
    - Само слово
    - Все существующие слова на небольшом редакционном расстоянии (1-2)
    - Слова, близкие по звучанию
    - Все это считается заранее
- Выбор лучшего кандидата
    - Noisy channel model

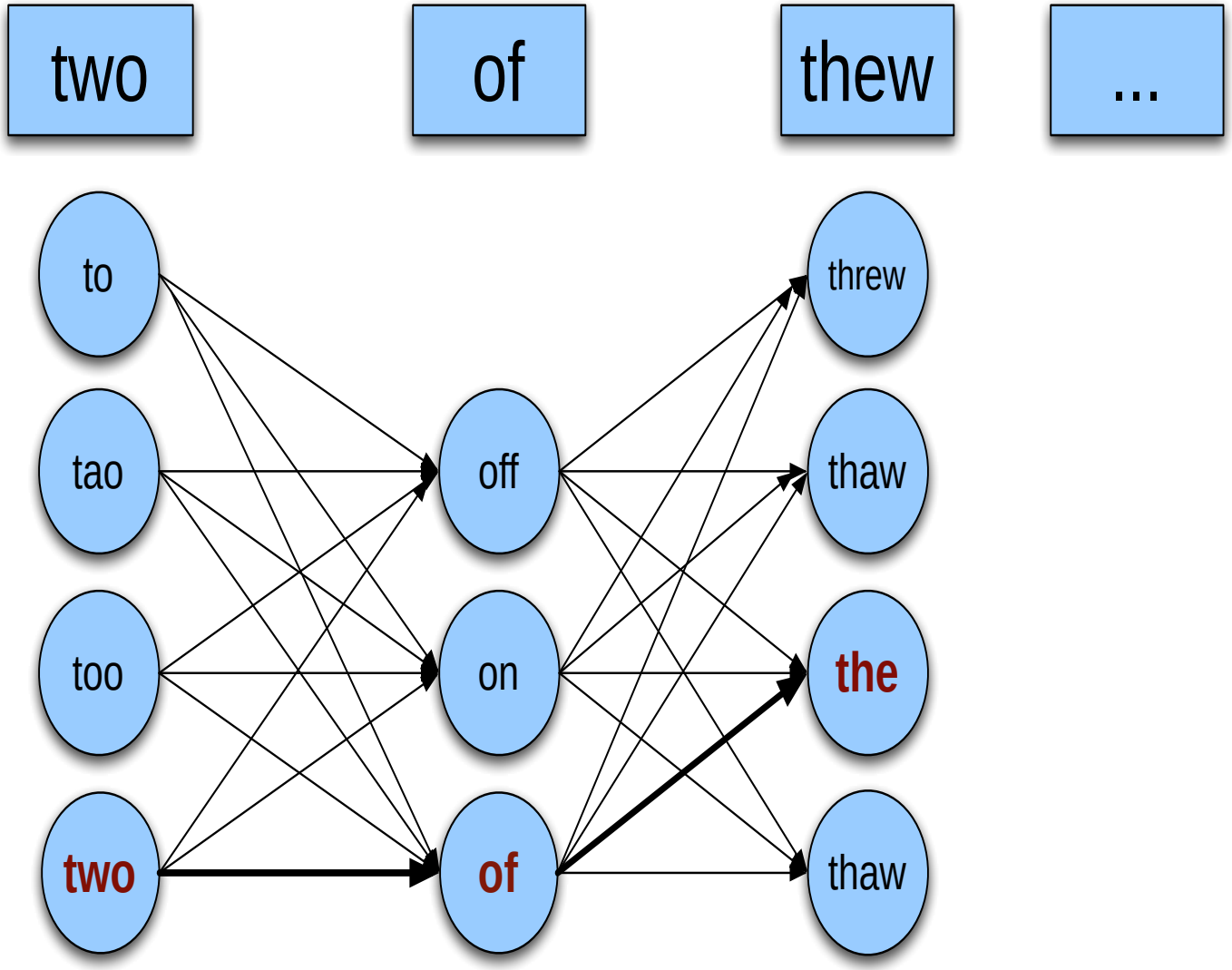
Noisy channel для исправления ошибочных словарных слов

- Дано предложение $w_1, w_2, w_3, \ldots, w_n$
- Множество кандидатов для каждого слова w_i
  - Candidate(w_1) = {w_1, w'_1 , w''_1 , w'''_1 ,…}
  - Candidate(w_2) = {w_2, w'_2 , w''_2 , w'''_2 ,…}
  - Candidate(w_n) = {w_n, w'_n , w''_n , w'''_n ,…}
- Нужно выбрать последовательностьW, которая максимизирует P(W)

Noisy channel для исправления замен на реальные слова



Noisy channel для исправления замен на реальные слова



Упрощение: Одна ошибка на предложение

- Все возможные предложения с заменой одного слова
  - w_1, **w’’$_2$**,w_3,w_4 two **off** thew
  - w_1,w_2,**w’$_3$**,w_4 two of **the**
  - **w’’’$_1$**,w_2,w_3,w_4 **too** of thew
  - …
- Нужно выбрать последовательностьW, которая максимизирует P(W)

Как получить вероятности

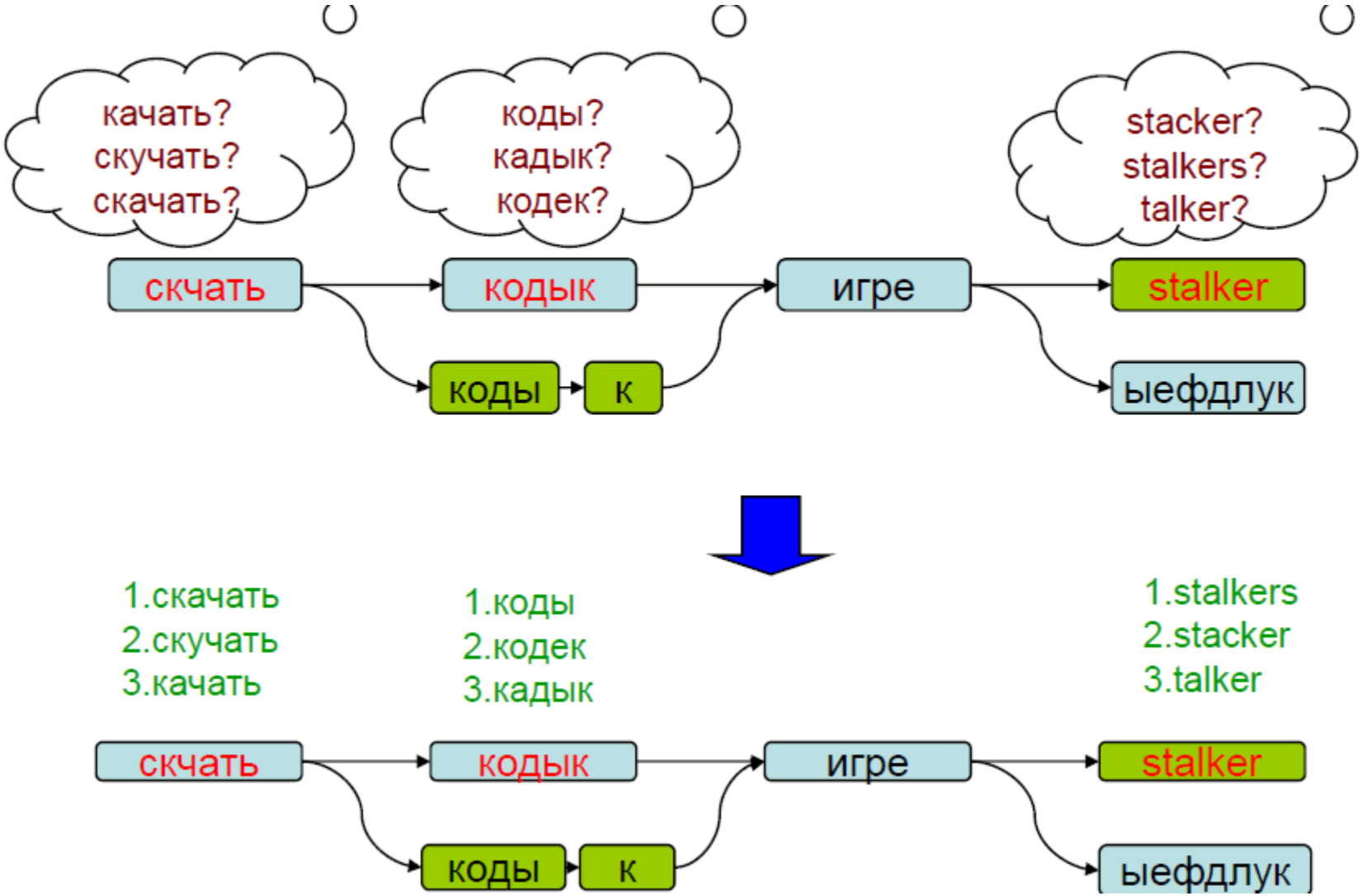
- Подсчет вероятностей по корпусу
  - Униграммы
  - Биграммы
  - И др.
- Channel model
  - То же самое, как для несловарной ошибки
  - Плюс нужна вероятность отсутствия ошибки: $P(w|w)$
    - *Оценивается вероятность отсутствия ошибки в слове*
      - 0.9-0.99

Пример с “thew”

x	w	x|w	P(x|w)	P(w)	10^9 P(x|w)P(w)
thew	the	ew|e	0.000007	0.02	144
thew	thew		0.95	0.00000009	90
thew	thaw	e|a	0.001	0.0000007	0.7
thew	threw	h|hr	0.000008	0.000004	0.03
thew	thwe	ew|we	0.000003	0.00000004	0.0001

Исправление ошибок запросов в Яндексе

Сортировка кандидатов



Выбор оптимального исправления

- Вероятность каждого варианта исправления:
- Вероятность каждого варианта исправления:
-
- $P(w_1,w_2,..,w_N) = \Pi_{i=1,N} (P(w_i|w_1w_{2...}w_{i-1}) * \Pi_{k=1,K} (P_{ok})$
- где
- $P(w_i|w_1w_{2...}w_{i-1})$ – условная вероятность слова w_i
- P_{ok} – вероятность k-ой ошибки
-
- **Опять условная вероятность! Слишком много вычислений…**

- Упрощаем: используем модель **двусловных сочетаний**

- Для запроса из 3-х слов вместо

$$P(w_1) * P(w_2|w_1) * P(w_3|w_1w_2)$$

- применяем:

скчать коды $P(w_2$ к $* P(w$ игре stalker

А.Байтин: удачные подсказки

одеяло стебаное	–> одеяло стеганое
лодки катра	–> лодки катера
квадратный мэтр	–> квадратный метр
выборы мера	–> выборы мэра
желательная резинка	–> жевательная резинка
грибница фараона	–> гробница фараона
вышел с ухой из воды	–> вышел сухой из воды

Было:	Стало:
гадостное настроение	радостное настроение

А. Байтин: неудачные подсказки

белявский	-> милявский
олбас	-> колбас
брендмауэры	-> брандмауэры
термису	-> термину
трассологическая	-> графологическая
любочка	-> юбочка
берег у моря	-> берегу моря
вход или выход	-> входили выход

А.Байтин: статистика (2009)

- **Находим ошибки в 10% запросов**
- **Точность исправления 75%**

- **Используем словарные базы:**

Список двусловий	**29M** сочетаний
Словарь	**2.7M** слов
Пользовательские замены	**190K** замен
Индекс кликабельности	**320K** замен

- **«Обслуживаем» службы Яндекса:**
- Поиск Блоги Новости Карты Маркет Картинки

- **Нагрузка кластера исправления**