Задача проверки правописания и зашумлённый канал

Практическая проверка правописания

Дятлова А. М.

18. 11. 2017

"Speech and Language Processing" (3rd ed. draft)
Dan Jurafsky and James H. Martin
Ch. 5, pp. 7-9

Ошибки правописания существующих слов

- Под землей живет **корт.** (переставленные буквы)
- Расскажи мне **об** себе. (неверное правило)
- Положи книгу на **сто.** (пропуск буквы)
- На лугу паслись **свечки.** (замена буквы)

25-40% орфографических ошибок — настоящие слова Kukich 1992

Вопрос

Какой процент орфографических ошибок — это настоящие существующие слова?

- ✓ 10 – 25 %
- ✓ 25 – 40 %
- ✓ 40 – 55 %

Решение ошибок существующих слов

- Для каждого слова в предложении:
  - Создается набор кандидатов:
    - Само слово
    - Все существующие слова, отличающиеся на 1 букву
    - Слова омофоны
- Выбор лучших кандидатов:
    - Модель зашумленного канала
    - Классификатор (определенный для конкретной задачи)

Вопрос

Как должен выглядеть набор кандидатов для ошибки:

"В день нужно съедать **гость** орехов"

Зашумленный канал для практической проверки правописания

- Учитывая предложение $w_1, w_2, w_3, \ldots, w_n$
- Создается набор кандидатов для каждого слова w_i
  - Кандидат $(w_1) = \{w_1, w'_1, w''_1, w'''_1, \ldots\}$
  - Кандидат $(w_2) = \{w_2, w'_2, w''_2, w'''_2, \ldots\}$
  - Кандидат $(w_n) = \{w_n, w'_n, w''_n, w'''_n, \ldots\}$
- Выберем последовательность W, которая максимизирует P(W)

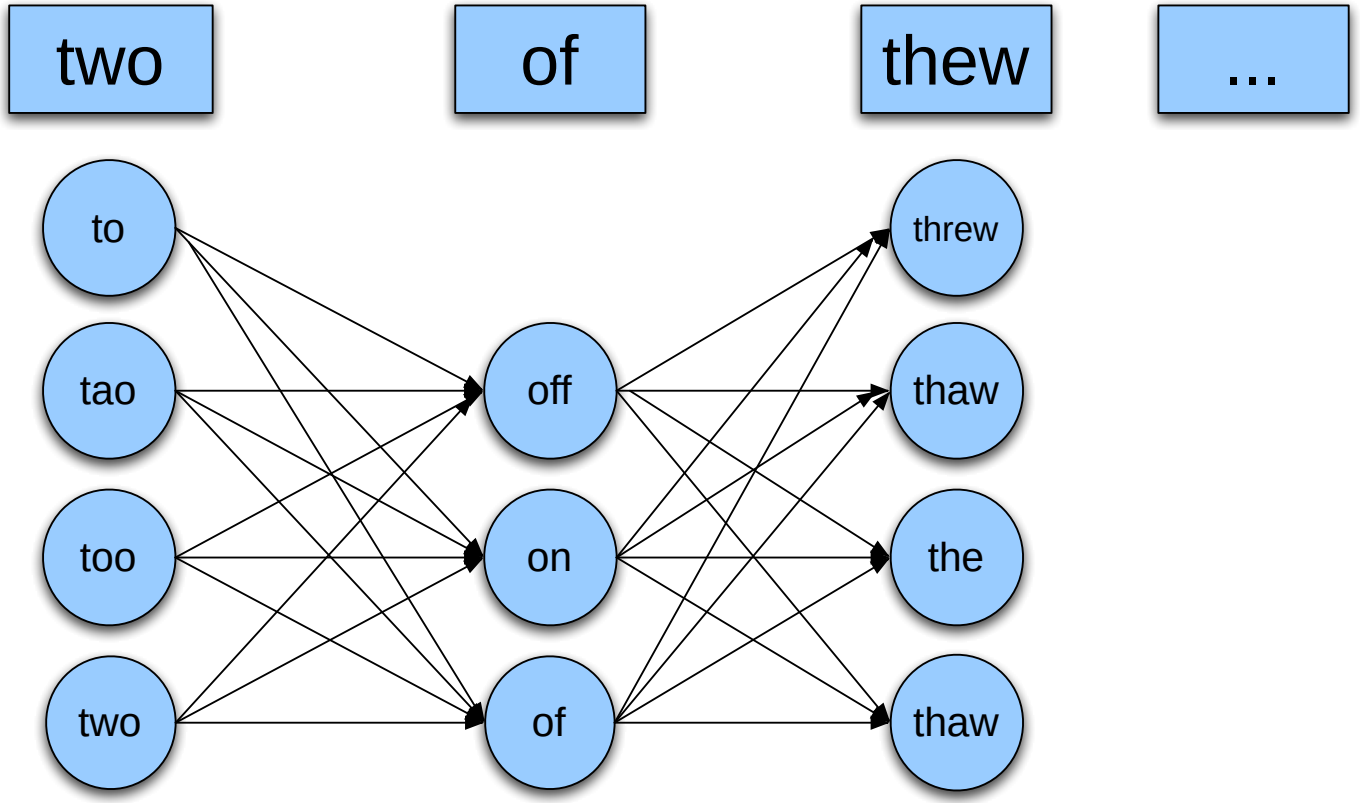
Вопрос

Какую нужно выбрать последовательность?

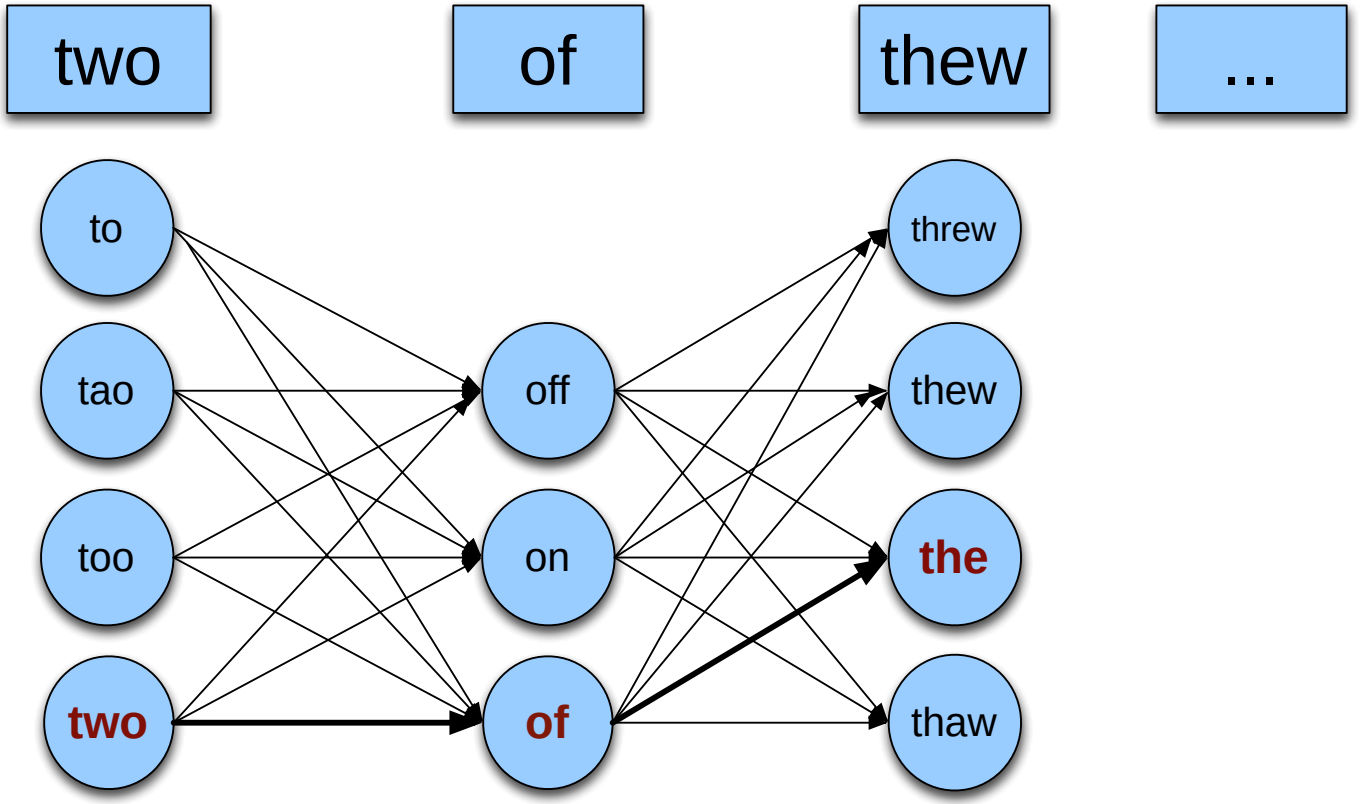
- ✓ которая минимизирует вероятность
- ✓ которая максимизирует вероятность

Зачем это нужно?

Зашумленный канал для практической проверки правописания



Зашумленный канал для практической проверки правописания



Упрощение: Одна ошибка в предложении

- Из всех возможных предложений 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
  - …
- Выберем последовательность W, которая максимизирует P (W)

Где взять вероятности

- Языковая модель
  - Униграмма
  - Биграмма
  - ….
- Модель канала
  - То же, что и для коррекции правописания "не-слов"
  - Плюс нужна вероятность отсутствия ошибки, P (слово | слово)

Вероятность ошибки

- Какова вероятность канала для правильно введенного слова?
- P("**the**" | "**the**")

- Очевидно, это зависит от приложения
  - .90 (1 ошибка в 10 словах)
  - .95 (1 ошибка в 20 словах)
  - .99 (1 ошибка в 100 словах)
  - .995 (1 ошибка в 200 словах)

Пример "thew" Питера Норвига

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

Norvig 2009

Вопрос

Почему в таблице был выбран именно вариант "**the**"?

Список литературы

Kukich, K. (1992). Techniques for automatically correcting words in text. ACM Computing Surveys, 24(4), 377–439.
URL: https://pdfs.semanticscholar.org/d204/4ca37a948fc34ea1f3f87e9090ec8bda4a33.pdf

Norvig, P. (2009). Natural language corpus data. In Segaran, T. and Hammerbacher, J. (Eds.), Beautiful data: the stories behind elegant data solutions. O'Reilly.
URL: http://vample.com/ebooks/Oreilly.Beautiful.Data.Jul.2009.pdf