Современные системы

Speech and Language Processing (3rd ed. draft)
Dan Jurafsky and James H. Martin
Глава 5, страницы 43-51

Епарская Анна 22508
25.12.2017

Проблемы с HCI* при написании

Если очень уверен в исправлении

- Автозамена

Менее уверен

- Дают лучшую коррекцию
- Дают список исправлений

Неуверен

- Просто пометить как ошибку

$$P(w|x) > P(x|x) \rightarrow \log P(w|x) - \log P(x|x) > 0$$

*HCI (англ. human-computer interaction) – Человеко-компьютерное взаимодействие

В каком случае предлагается коррекция w в более осторожной системе?

Методы на основе классификатора для коррекции правописания в реальном времени

- Вместо простой модели канала и языковой модели
- Используйте множество функций в классификаторе (следующая лекция).
- Создайте классификатор для определенной пары, например:

whether/weather

- “cloudy” within +- 10 words
- ___ to VERB
- ___ or not

Современный шумный канал

- Не перемножаем модели
- Вероятности не соразмерны
- Вместо этого: взвешивать их

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

- Узнайте λ из набора тестов разработки

Почему используют взвешенную комбинацию?

Усовершенствования модели каналов

Разрешить более радикальные изменения (Brill and Moore 2000)

Physical	fisikle
ph y s i c al	f i s i k le

p (f | ph) * p (i | y) * p (s | s) * p (i | i) * p (k | k) * p (al | le)

P (f | ph) -> P (f | ph, начало)

- ent → ant
- ph → f
- le → al

Преимущество модели Брилла Мура?

Модель фонетических ошибок

Метафон, используемый в GNU aspell

- Преобразование опечаток в метафон произношения

  «Отбросить дубликаты соседних букв, кроме C».

  «Если слово начинается с « KN »,« GN »,« PN »,« AE »,« WR », отбросить первую букву».

  «Отбросить "B", если после 'M' и если в конце слова»

- Найти слова, произношение которых на 1-2 расстояние редактирования от опечаток
- Список результатов оценки

  Взвешенное расстояние редактирования кандидата до опечатки

  Редактирование расстояния между произношением кандидата и произношением с ошибкой

ЧТО ТАКОЕ МЕТАФОН?

Усовершенствования модели каналов

Включить произношение в канал (Toutanova and Moore 2002)

- `actress и aktres → ae k t r ix s`

Задача дедупликации

Алгоритм Soundex (Knuth 1973, Odell and Russell 1922) - это более старый метод

- Jurafsky, Jarofsky, Jarovsky, Jarovski

Алгоритм Jaro-Winkler (Winkler, 2006).

Что решает задача дедупликации?

Список литературы

- Jurafsky D. & Martin J. H. Speech and Language Processing [электронный ресурс] . — Ch. 5, pp. 43-51 . — 2017.

  ACM. URL:

  https://web.stanford.edu/~jurafsky/slp3/5.pdf