Языковые модели

Задачи.
В общем: определить вероятность последовательности слов

- Исправление опечаток
- Генерация текста
- Машинный перевод
- Предсказание следующего слова

Данные

- Google Books n-gram viewer
- Microsoft Web Language Model API

Оценки моделей

Внешняя Extrinsic

Внутренняя Intrinsic

Коэффициент неопределённости. Перплексия

Хорошая модель приписывает более высокую вероятность тому предложению, которое действительно будет чаще встречаться в тексте, и наоборот

Обратная вероятность тестового набора, нормализованная на длину.
Как-то связана с энтропией

Логарифм:

$$\log PP(S) = -\frac{1}{n}\log P(w_1 w_2 \dots w_n) =$$
$$= -\frac{1}{n}\sum_{1}^{n}\log P(w_i|w_{i-1})$$

Формула:

Чем ниже, тем лучше

PP \in [1, |V|]

Log-Likelihood/Cross-Entropy

Хотим

- Coherency
- Diversity - разнообразие
  - При малом разнообразии можно сэмплировать слова с температурой
- Как можно сэмплировать
  - Лучшее — top-p % чтобы покрывать определённый процент вероятностной массы

Принцип работы

Статистические методы (считаем руками)

Нахождение вероятностей

Используя n-граммы

Произведение вероятностей + марковский процесс (избавляемся от зависимости от " прошлого")

Метод макс. правдоподобия

$$P_{MLE}(w_i|w_{i-1}) = \frac{f(w_i, w_{i-1})}{f(w_{i-1})}, f() - \text{частота}$$

Лучше юзать логарифмы вероятностей, как обычно

- Логарифмирование
- Сложение быстрее чем умножение

Марковское предположение: теперь слово не зависит от всего (возможно, большого) контекста, а только от n слов (поскольку у нас n-граммы)

Сглаживания

Сглаживание вероятностей по Лапласу.
Или "сглаживание + 1"

Модель должна обобщать

Всегда будут неизвестные слова

Если в тестовых новые слова, а в тренировочных нет, то модель будет ломаться

Формула:

$$P_{\text{Laplacian}} = \frac{f(w_i, w_{i-1}) + 1}{f(w_{i-1}) + \#D}, \text{где D} - \text{словарь}$$

Восстановление частот (после сглаживания)

Формула:

$$f^*(w_{i-1}w_i) = \frac{[f(w_{i-1}w_i) + 1] \times f(w_{i-1})}{f(w_{i-1}) + \#D}$$

Иногда работает плохо, более мягко использовать параметризованный вариант "+alpha"

Формула:

$$P_{\alpha\text{Laplacian}} = \frac{f(w_i, w_{i-1}) + \alpha}{f(w_{i-1}) + \alpha\#D}$$

Идея похожая на Лапласа: так же перекладывается часть вероятностной массы на неизвестные случаи

Сглаживание Гуда-Тьюринга

На примере: рыбалка. Поймали 10 карпов, 3 окуня, 2 сига, 1 форель, 1 лосось, 1 угорь = 18 рыб

1. Допустим, новая рыба будет редкой и её вероятность будет как наши пойманные редкие виды (форель, лосось и угорь), т.е. 3/18. Тогда для этих "редких" вероятность пересчитывается и будет ещё ниже

2. Наши переменные: n10 = 1, n3 = 1, n2 = 1, n1=3 (ni – частота частот для количества i). Т.о. оценка для нового вида рыбы будет равна:

$$P_{G-T}(\text{новый вид}) = \frac{n_1}{n}$$

3. Но теперь нам нужно подправить старые вероятности, f – старые вероятности:

$$f^* = \frac{(f+1)n_{f+1}}{n_f}$$

4. Таким образом, искомая вероятность будет равна:

$$P_{G-T}(\text{форель}) = \frac{n_1}{n} = \frac{f*}{n} = \frac{(f_{\text{форель}}+1)n_2}{n_1}/n = \frac{(1+1)1}{3}/18 = 2/3/18 = 1/27$$

В области больших k значения nk не идут подряд (некоторые значения ni нулевые (в столбиковой диаграмме есть "ямы")

Можно аппроксимировать какой-либо функцией (степенной, к примеру). А далее можем использовать эти значения

Другие сглаживания (практически идентичны вышеуказанным, мало используются)

- Witten-Bell — Учёт разнообразия продолжения
- Kneser-Ney — Учёт вероятности слова быть продолжением (смотрим на биграммы)

Обучаем вероятности (нейросетями)

Обучаем условные вероятности

Как обучить? Как задачу классификации

- Предсказываем следующее слово
- Кормим в НС контекст, получаем скрытое представление. Линейным слоем трансформим представление в размер словаря, вешаем софтмакс, побеждаем

Основная модель: модель "Left-to-Right"

- Слова идут слева направо и каждая последовательная пара зависит друг от друга (точнее правое слово от левого)
- Уходим от модели мешка слов
- Здесь мы как бы моделируем предложения
- Используем полученное вероятностное распределение и по нему генерируем предложение
- Отличие от человека должно быть МИНИМАЛЬНЫМ