確率的言語モデル

演習

吉野 幸一郎

2011/7/7

目次

- 言語モデルとは
- 文字言語モデル
- 単語 n-gram モデル
- 未知語モデルとスムージング
- エントロピー

目次

- 言語モデルとは
- 文字言語モデル
- 単語 n-gram モデル
- 未知語モデルとスムージング
- エントロピー

言語モデルとは

- 言語における文字列の傾向を確率的に表したもの

$$P(\text{最初のタスクである}) = P(\text{最初}, \text{の}, \text{タスク}, \text{で}, \text{あ}, \text{る})$$

- 自然な文に高い確率を付与する

$$P(\text{最初のタスクである}) > P(\text{タスクで最初でるあ})$$

- 頻出する文に高い確率を付与する

$$P(\text{コサイン距離に基づく}) > P(\text{うなぎなう})$$

（論文の場合）

ようするに　**言語現象を確率的に扱うためのモデル**

一般的な言語モデル

- $P(s) = P(c_1, c_2, \ldots, c_{n+1}) = \prod_{i=1}^{n+1}(c_i|c_0, \ldots, c_{i-1})$
- $P(\text{最初}, \text{の}, \text{タスク}, \text{で}, \text{あ}, \text{る}) =$
  $P(\text{最初}|BR)P(\text{の}|BR, \text{最初})P(\text{タスク}|BR, \text{最初})$ …

- c_0は文頭, c_{n+1}は文末を表す特殊文字
  - 2byteならBRやBT,3byteならTOSやEOSなど
  - c_0はc_1が文の最初である確率（1-gramではいらない）
  - c_{n+1}はc_nが文の終わりである確率

- N-gramモデルとは
  条件付き確率の条件を一定数だけ使うモデル

目次

文字言語モデル

- 日本語や中国語には明示的な単語境界がない
  - 単語言語モデルを作るのに単語分割が必要
- 単語言語モデルの場合未知語の問題が発生

- まずは文字単位で言語モデルを構築する
  - 文字0-gram言語モデル（最も単純なモデル）
    - 全ての文字に同じ確率を付与する
    - UTF-8では95221+1(文頭・文末文字を加える)

文字確率: $P_{c0}(c) = \frac{1}{95222}$

文確率: $P_{c0}(s) = \left(\frac{1}{95222}\right)^{n+1}$ 文末の確率分

文字1-gram言語モデル

- それぞれの文字に確率を与えるモデル
  - 履歴は利用しない（同じ文字には同じ確率）
  - 学習コーパス中の分布を確率として用いる

文字確率: $P_{c1}(c_i) = \frac{C(c_i)}{\sum_j C(c_j)}$

文確率: $P_{c1}(s) = \prod_{i=1}^{n+1} P_{c1}(c_i)$

文字n-gram言語モデル

- それぞれの文字に条件付き確率を与えるモデル
  - 条件は履歴（2-gramなら前1個分の文字）
  - 学習コーパス中の分布を確率として用いる
  - 以下は文字2-gramの例

文字確率: $P_{c2}(c_i|c_{i-1}) = \frac{C(c_{i-1},c_i)}{C(c_{i-1})}$

文確率: $P_{c2}(s) = \prod_{i=1}^{n+1} P_{c2}(c_i|c_{i-1})$

目次

単語1-gram言語モデル

- それぞれの単語に確率を与えるモデル
  - 履歴は利用しない（同じ単語には同じ確率）
  - 学習コーパス中の分布を確率として用いる
  - 基本的には単位が異なるだけで文字と同じ
  - 単位が単語なので、与えるコーパスは分割済である必要がある

単語確率: $P_{w1}(c_i) = \frac{C(w_i)}{\sum_j C(w_j)}$

文確率: $P_{w1}(s) = \prod_{i=1}^{n+1} P_{w1}(w_i)$

単語n-gram言語モデル

- それぞれの単語に条件付き確率を与えるモデル
  - 条件は履歴（2-gramなら前1個分の単語）
  - 学習コーパス中の分布を確率として用いる
  - 以下は文字2-gramの例

単語確率: $P_{w2}(w_i|w_{i-1}) = \frac{C(w_{i-1},w_i)}{C(w_{i-1})}$

文確率: $P_{w2}(s) = \prod_{i=1}^{n+1} P_{w2}(w_i|w_{i-1})$

目次

未知語（未知文字）の問題

- 未知語とは
  - 学習コーパスに出現しなかった単語（文字）の確率は計算できない
- 未知文字モデル
  - 文字1-gramの場合、コーパスに出現しなかった文字の確率は計算されない
  - 文字0-gramは出現文字数の上限があるのでバックオフまたは補間すれば対処可能
- 未知語モデル
  - 単語数に上限はない

バックオフと補完

- バックオフ
  - 確率がなかった場合に未知語モデルを使うこと
  - 例では1回以下の確率を削って未知語モデルに割当てる

$$P(w_i) = \begin{cases} P_{w1}(w_i) & if\ C(w_i) > 1 \\ \dfrac{\sum_{w_i \in C(w_i)=1} C(w_i)}{\sum_{w_i} C(w_i)} P_{unk}(w_i) & else \end{cases}$$

- 補間
  - 未知語モデルと組み合わせた確率を作ること

$$P(w_i) = (1-\alpha)P_{w1}(w_i) + \alpha P_{unk}(w_i)$$

αを補間係数と呼ぶ

補間係数の選び方

1. 適当に決める
2. ヘルドアウトセットでパラメータを推定
3. 絶対割引法
4. 線形法
5. Witten-Bell法　　etc...

今回は適当に決めます

このあたりは夏の音声認識講習会で詳しく習います

目次

エントロピー（モデルの評価）

- 確率的言語モデルを評価する方法（の１つ）

$$E = \frac{\sum_s -\log_2 P(s)}{\text{コーパスの文字数}}$$

- エントロピーの意味
  - そのモデルで圧縮をした場合、何ビット使うか
  - エントロピーが低いとモデルの予測精度が良い（それだけ候補が少ないということ）

カバレージ

- モデルの評価セットに対するカバー率

$$Cov = \frac{\sum_w I(C\ (x) > 0)}{\sum_w 1}$$

- 複雑なモデルほどカバレージは下がる
- (tri-gram cov. < bi-gram cov. < ...
  < char-zero-gram cov. = 100%)

課題

- コーパス
  - MPTコーパス（森先生のresearchのページから）
  - 200文を学習セットに　65文を評価セットにします

1. 65文の文字0-gramのエントロピーを測る
2. 65文の単語1-gramのエントロピーを測る
   - 補間係数は0.1とし線形補間で文字0-gramの未知語モデルを利用すること
3. 65文の単語2-gramのエントロピーを測る
   - できれば単語1-gramの線形補間＋文字0-gramの線形補間を行うこと（直接文字0-gramでもよい）
4. 3.の最適な補間係数を探す
   - 0.1刻みでエントロピーがどう変化するか確かめる

モデルのイメージ

入力

最初, の, タスク, で, あ, る

単語2-gram
言語モデル

P(最初|TOS)
P(の|最初)
P(タスク|の)
P(で|タスク)
P(あ|で)
P(る|あ)
P(EOS|る)

単語1-gram
言語モデル

P(タスク)
P(で)

文字0-gram
言語モデル

$P(タ) \times P(ス) \times P(ク) = \frac{1}{95222}$

注意点

- 確率モデルを扱う場合桁落ち・丸め誤差が問題となることが多いので、対数を取って利用する
  - 対数を取ると計算式が微妙に変わるので注意すること
  - 補間は対数を取る前にする必要がある