What is Statistical Language Model?

- Statistical language models are probability P(w_1, w_2, .., w_n) defined on sequences of words w_1, w_2, .., w_n
- Many NLP applications
  - speech recognition, machine translation and information retrieval
- Estimating the probability becomes expensive and infeasible for long sequences (not enough data/storage)
- Smoothed N-gram models based on unigrams, bigrams, and trigrams are often used to approximate the language model:

  $P(w_1, w_2, .., w_n) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \ ... \times P(w_n|w_{n-1})$

How to derive a language model

1. Get a corpus of texts
2. Break texts into sentences, sentences into tokens
3. Count the frequencie of tokens, pairs, triples
4. Apply probability theory and information theory to estimate model parameters (prob. of tokens, pairs, triples)
    - MLE (Maximal Likelihood Estimation) = based on frequency, maximizing the prob. of sample, but gives zero prob. to unseen events
    - Smoothing = Dealing with low-count and unseen events (giving them non-zero probability)

Example – Case Restoration

- Input:
  - THE U.S. MILITARY ISSUED A PUBLIC APOLOGY TO THE PEOPLE OF A SHIITE MUSLIM NEIGHBORHOOD IN BAGHDAD ON THURSDAY …
- Output:
  - The U.S. military issued a public apology to the people of a Shiite Muslim neighborhood in Baghdad on Thursday …
- Many Likely Candidates for Output
  - the u.s. military issued a public apology to the people of a shiite muslim neighborhood in baghdad on thursday …
  - The u.s. military issued a public apology to the people of a shiite muslim neighborhood in baghdad on thursday …
- Which is more likely = Max Probability
  - P(the u.s. military issued a public apology to the people of a shiite muslim neighborhood in maghdad on thursday …)
  - P(The u.s. military issued a public apology to the people of a shiite muslim neighborhood in maghdad on thursday …)

Example – Filling in Articles and Prepositions

- In Alan Meryers (2005) Gateways to Academic Writing pp. 277, learners are asked to fill articles and prepositions in the blanks.
- The model T Ford was a fragile-looking automobile, but it became *the* most popular car in history. Henry Ford sold 16 million Model Ts _____ the years 1908 and 1928. The Model T is _____ immediate best-seller, not only because of its low price, but because it was _____ powerful car.

Example – using articles and prepositions correctly

- In Gateways to Academic Writing (Meryers 2005, pp. 277), learners are asked to fill articles and prepositions in the blanks.
- The model T Ford was a fragile-looking automobile, but it became ___*the*___ most popular car in history. Henry Ford sold 16 million Model Ts ___*in*___ the years 1908 and 1928. The Model T is ___*an*___ immediate best-seller, not only because of its low price, but because it was ___*the*___ powerful car.

How to Estimate Probability

- Counting words – (w, count(w)) (done in Lab 1)
- MLE (maximal likelihood estimator)
  - For all words w with count i ➔ P(w) = i / N
  - For all unseen words ➔ P(w) = 0

Smoothing

- Good-Turing Estimation
- V = Vocabulary Size

1. Counting word types with counts 1, 2, 3, …
   - $N = N_1 + 2\,N_2 + 3\,N_3 + \ldots$
   - $N_0 = V - N_1 - N_2 - N_3 - \ldots$ *(for 1gram)*
   - $N_0 = V^2 - N_1 - N_2 - N_3$ *(N_1, N_2 , N_3* are different *for 2gram)*
2. Compute $0^* = 1 \times N_1 / N_0$

   $r^* = (r+1)\,N_{r+1} / N_r$, for r = 1, k-1

   $r^* = r$, for r >= k
3. Adjusting counts and prob. with r*
   - For unseen word w (r = 0)

     $r^* = 0^* = N_1 / N_0$

     $P(w) = r^* / N = N_1 / (N_0\,N)$
   - For seen word w with count r < k (k = 10)

     $r^* = (r+1)\,N_{r+1} / N_r$

     $P(w) = r^* / N = (r+1)\,N_{r+1} / (N_r\,N)$

Re-normalization

- All probabilistic values must sum to 1
- Adjusted total
  - $N' = N_1 + 2\,N_2 + 3\,N_3 + \ldots + (k\text{-}1)\,N_{k-1} + k\,N_k + k\,N_k + (k+1)\,N_{K+1} + \ldots$
- Prob sum to $N' / N = (N + k\,N_k) / N$ (not 1.0)
- Normalization

  $P'(w) = N\,P(w) / (N + k\,N_k)$
- Do the same for bigram (w, w')
- Or directly

  $P(w) = r^* / (N + k\,N_k) = N_{r+1} / N_r / (N + k\,N_k)$ for r = 0, 1, ..., k-1

  $P(w) = r / (N + k\,N_k) = r / (N + k\,N_k)$ for r >= k

Estimating P(w) and P(w'|w)

- P(w) = r*/N
  -- r = count(w), go from r to r* (for unigram)
- P(w, w') = r*/N
  -- r = count (w, w'), go from r to r* (for bigram)
- P(w'|w) = P(w', w) / P(w)

Language Modeling Tools

- SRI Language Modeling Toolkit
  - http://www.speech.sri.com/projects/srilm/
- Tutorial
  - Introduction to SRILM by Berlin Chen
  - http://berlin.csie.ntnu.edu.tw/Courses/2005F-SpeechRecognition/Lectures2005F/SP2005F_Lecture08_SRILM%20Tutorial.pdf
- Python Wrapper for SRILM
  - http://www.isi.edu/~chiang/software/psrilm.tgz

Example of Running SRILM

- INPUT (科技想要什麼？)
  - 寫給 台灣 的 讀者 凱 文 · 凱利
  - 這 本 書 ， 可以 說 是 四十 年 前 在 台灣 扎 下 了 根基
  - 一九七二年 ， 二十 歲 的 我 離開 美國 紐澤西 ， 結束 平淡 的 生活 ， 展開 長長 的 旅程 ， 來到 一 個 完全 不同 的 世界 ， 也 就是 當時 正 開始 面臨 轉變 的 台灣
  - 我 第一 次 出國 ， 目的地 就 是 台灣
  - 眼前 所 見 ， 令 我 震驚 不能 自已
  - 那 個 時候 ， 最 主要 的 交通 工具 是 腳踏車 ， 現在 的 新 北市 則 是 一片片 稻田
  - 我 親眼 見證 ， 這 塊 活躍 的 土地 急速 現代化 ， 幾乎 每 天 都 在 改變 ， 從 第三世界 國家 提升到 名列 世界 富國
  - 台灣 ， 達成 了 不可能 的 任務

Example of Running SRILM

- Language Model
  - \data\
  - ngram 1=11589
  - ngram 2=71201
  - ngram 3=8893
  - \1-grams:
  - -5.149576　　一共　　-0.1150313
  - -4.149576　　一再　　-0.2703053
  - -3.516108　　一切　　-0.3589188
  - -3.631062　　一千　　-0.5536026

    ~ $\log_{10}$ (24/11589)　　(back off weight = 『除非 一千』查不到時的權重)

Example of Running SRILM

- \2-grams:
- -0.6187539　一共 有
- -0.4609967　一再 出現　　0.00103879
- -1.618754　一再 提出
- ...