COMP 4360: Machine Learning
Menu Translation

Jacky Baltes
Autonomous Agents Lab
University of Manitoba
Winnipeg, Canada
R3T 2N2

Email: jacky@cs.umanitoba.ca
WWW: http://www.cs.umanitoba.ca/~jacky
http://aalab.cs.umanitoba.ca
Automatic Translation

- Some more comments about the assignment
- You are given some training data
  - "葱爆肉丁"; "sliced pork with scallion"
  - "葱爆肉丝"; "pork slices with vermicelli"
  - "葱爆羊肉"; "fried mutton slice with green scallion"
- Translate an expression such as
  "葱头牛肉丝"; "shredded beef with onion" into the other language
Machine Translation

● 3 Steps required to do the translation
  1. Segmentation
  2. Translation
  3. Ordering (Distortion)
● Segmentation
  ○ Break the original and other texts up into groups (words)
  ■ Chinese:
    ■ 葱头牛肉丝
  ■ Each character is a word (Approximation)
  ■ 葱头牛肉丝
  ■ More correct
  ■ English
    ■ shredded beef with onion
    ■ Separated by space
Translation

Now we would like to know the most likely translation for each segment.

Apply Naive Bayesian learning to find most likely translation for each segment $P(t_i|o_i)$.

Bayesian Theorem: $P(t_i|o_i) = P(o_i|t_i) \times P(t_i)/P(o_i)$

Apply to find a translation for $o=葱$.

Possibilities:
- sliced(1) pork(2) with(3) scallion(2) slices(1) vermicelli(1) fried(1) mutton(1) slice(1) green(1)
- $t=pork$: $P(葱|pork)=2/2$, $P(pork)=2/3$, $P(葱)=3/3$
- $t=mutton$: $P(葱|mutton)=1/1$, $P(mutton)=1/3$, $P(葱)=3/3$
- $t=scallion$: $P(葱|scallion)=2/2$, $P(scallion)=2/3$, $P(葱)=3/3$
- $t=sliced$: $P(葱|scallion)=1/1$, $P(sliced)=1/3$, $P(葱)=3/3$
Translation

- The word "with" causes a problem, since it has a probability of 1.0 in this example
- More/better training data would help
- Phrase based translation
  - Multiply probabilities for all word translations to give probability of translation of entire phrase
  - This assumes statistical independence of words in phrase
Distortion

- Distortion models the shifting of the segments with respect to each other.
- For example:

  - This is done by maintaining a conditional probability for each translation pair.
    - Use Bayesian Theorem to maximize distortion probability
      - \( P(葱|+2)=0.8, P(肉|-2)=0.7, \ldots \)
Evaluation

- 10% Holdout method
- Bag of words approach: order of words does not matter