CMPT-825
Natural Language Processing

Anoop Sarkar

http://www.cs.sfu.ca/~anoop

Human Supervision in Part of Speech Tagging

- In unseen data, we wish to find the part of speech tags:
  Input: *In 1994 , Hartnett said*
  Output: *In*_IN *1994*_CD ,_, *Hartnett*_NNP *said*_VBD

- The set of part of speech tags are decided by experts

- The experts also have to provide adequate amounts of data in which the part of speech tags have been listed for each word in context.

- This general approach is called **supervised learning** since the training data is provided by humans.

Trigram Models for Part of Speech Tagging

```
In_IN 1994_CD ,_, Hartnett_NNP said_VBD

THE_DT BONEYARD_NNP Northrop_NNP  Grumman_NNP ’s_POS modest_JJ
flight_NN museum_NN occupies_VBZ a_DT corner_NN of_IN one_CD of_IN
its_PRP\$ power-seat_NN adjusters_NNS ,_, door_NN trim_JJ now_RB
made_VBN in_IN South_NNP Korea_NNP ’s_POS antiquated_JJ coal-fired_JJ
power_NN plant_NN in_IN Canada_NNP ,_, to_TO a_DT 11.9_CD million_CD
mark_NN investment_NN in_IN Samsung_NNP ’s_POS Sachon_NNP plant_NN
in_IN Taiwan_NNP as_IN part_NN of_IN a_DT steam_NN turbine_NN ,_,
a_DT new_JJ high-yielding_JJ rice_NN plant_NN was_VBD reorganized_VBN
into_IN a_DT big_JJ expansion_NN of_IN a_DT fuel-fabrication_NN
plant_NN near_IN Nagoya_NNP in_IN Aichi_NNP Prefecture_NNP

From_IN October_NNP ,_, when_WRB they_PRP
do_VBP not_RB need_VB it_PRP
```

Borges gives a vague reference to some work by Franz Kuhn allegedly commenting on the classification of animals by a Chinese encyclopedia called the _Celestial Emporium of Benevolent Knowledge_.

>> ... animals are divided into:

(a) those that belong to the Emperor,
(b) embalmed ones,
(c) those that are trained,
(d) suckling pigs,
(e) mermaids,
(f) fabulous ones,
(g) stray dogs,
(h) those that are included in this classification,
(i) those that tremble as if they were mad,
(j) innumerable ones,
(k) those drawn with a very fine camel brush,
(l) others,
(m) those that have just broken a flower vase,
(n) those that resemble flies from a distance. <<
-- Jorge Luis Borges, "Other Inquisitions"

Part of Speech Tagging using Trigram Models

- Let the input sentence (word sequence) be $w_0, w_1, \ldots, w_n$
- Let the most likely tag sequence be $T^* = t_0^*, t_1^*, \ldots, t_n^*$
- In order to compare all possible tag sequences we build a probability model:

$$P(t_0, t_1, \ldots, t_n \mid w_0, w_1, \ldots, w_n)$$

Part of Speech Tagging using Trigram Models

- The best (or most likely) tag sequence is:

$$T^* = \underset{t_0, \ldots, t_n}{\text{arg max}} \; P(t_0, \ldots, t_n \mid w_0, \ldots, w_n)$$

$$P(t_0, \ldots, t_n \mid w_0, \ldots, w_n)$$

$$= \frac{P(w_0, \ldots, w_n \mid t_0, \ldots, t_n) \times P(t_0, \ldots, t_n)}{P(w_0, \ldots, w_n)} \text{(Bayes Rule)}$$

$$= P(w_0, \ldots, w_n \mid t_0, \ldots, t_n) \times P(t_0, \ldots, t_n)$$

Part of Speech Tagging using Trigram Models

$$
\begin{aligned}
& P(w_0, \ldots, w_n \mid t_0, \ldots, t_n) \\
&= P(w_0 \mid t_0) \times P(w_1 \mid t_1) \times \ldots \times P(w_n \mid t_n) \\
&= \prod_{i=0}^{n} P(w_i \mid t_i)
\end{aligned}
$$

$$
\begin{aligned}
& P(t_0, \ldots, t_n) \\
&= P(t_0) \times P(t_1 \mid t_0) \times P(t_2 \mid t_0, t_1) \times \ldots \times P(t_n \mid t_{n-2}, t_{n-1}) \\
&= P(t_0) \times P(t_1 \mid t_0) \times \prod_{i=2}^{n} P(t_i \mid t_{i-2}, t_{i-1})
\end{aligned}
$$

Part of Speech Tagging using Trigram Models

$$P(t_0, \ldots, t_n \mid w_0, \ldots, w_n)$$

$$= P(w_0, \ldots, w_n \mid t_0, \ldots, t_n) \times P(t_0, \ldots, t_n)$$

$$= \left(\prod_{i=0}^{n} P(w_i \mid t_i) \right) \times \left(P(t_0) \times P(t_1 \mid t_0) \times \prod_{i=2}^{n} P(t_i \mid t_{i-2}, t_{i-1}) \right)$$

$$= \prod_{i=0}^{n} P(w_i \mid t_i) \times P(t_i \mid t_{i-2}, t_{i-1})$$

Part of Speech Tagging using Trigram Models

- So, all we need to do to find the most likely tag sequence is to *train* the following two probability models:

$$P(w_i \mid t_i) \text{ and } P(t_i \mid t_{i-2}, t_{i-1})$$

- Easy to do if we have **training data** with word and tag sequences.

- All we need after we have the probability models is an algorithm to find the most likely tag sequence

- Use the algorithm used to find the best tag sequence in Hidden Markov Models: the *Viterbi* algorithm

Part of Speech Tagging using Trigram Models

- **Evaluation**: *train* your model on the training data, *test* on unseen test data to obtain best tag sequence for each word sequence.

- **Accuracy** is measured as the percentage of correct tags for words in the test data.

Brief History of Part of Speech Tagging

- Corpus building: English
  - Brown Corpus: 1979 (87 tags)
  - Penn Treebank Corpus: 1993 (45 tags)
  - British National Corpus (BNC): 1997
  - LOB corpus
- Other languages: Chinese, Czech, German, Korean, Turkish, . . .

Brief History of Part of Speech Tagging

- Models and Algorithms:
  - ngram models for tagging: Church 1988
  - extension of ngram model using HMMs: Xerox (Cutting et al) 1992
  - Transformation-Based Learning: Brill 1995
  - Maximum Entropy Models: Ratnaparkhi 1997
- Current work, using POS tagging to find phrases: *Chunking*