Prediction and part-of-speech tagging

Lecture overview:

1. corpora in NLP
2. word prediction
3. part-of-speech (POS) tagging
4. evaluation in general, evaluation of POS tagging

Corpora

Changes in NLP research over the last 10-15 years are largely due to increased availability of electronic corpora.

- **corpus**: text that has been collected for some purpose
- **balanced corpus**: texts representing different genres
- **tagged corpus**: a corpus annotated with POS tags
- **treebank**: a corpus annotated with parse trees
- specialist corpora — e.g., collected to train or evaluate particular appplications
  - Wizard of Oz experiment: human pretends to be a computer

Prediction

Guess the missing words:

Illustrations produced by any package can be transferred with consummate __ to another.

Wright tells her story with great __.

Prediction is relevant for:

- language modelling for speech recognition:
  e.g., using **N-grams**
  alternative to finite state grammars, suitable for large-scale recognition
- word prediction for communication aids
  e.g., to help enter text that's input to a synthesiser
- text entry on mobile phones etc
- OCR, spelling correction, text segmentation
- estimation of entropy

bigrams

A probability is assigned to a word based on the previous word:

$$P(w_n|w_{n-1})$$

where w_n is the nth word in a sentence.

Probability of a sequence of words:

$$P(W_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

Probability is estimated from counts in a training corpus:

$$\frac{C(w_{n-1}w_n)}{\Sigma_w C(w_{n-1}w)}$$

i.e. count of a particular bigram in the corpus divided by the count of all bigrams starting with the prior word.

Calculating bigrams

⟨s⟩ good morning ⟨s⟩ good afternoon ⟨s⟩ good afternoon ⟨s⟩ it is very good ⟨s⟩ it is good ⟨s⟩

sequence	count	bigram prob
<s>	5	
<s> good	3	.6
<s> it	2	.4
good	5	
good morning	1	.2
good afternoon	2	.4
good <s>	2	.4
morning	1	
morning <s>	1	1
afternoon	2	
afternoon <s>	2	1
it is	2	1
is very	1	.5
is good	1	.5
very good	1	1

Practical application

- word prediction: guess the word from initial letters (user confirms at each point)
- speech recognition: maximize likelihood of a sequence (implemented using the Viterbi algorithm)

Problems because of **sparse data**:

- smoothing: distribute 'extra' probability between rare and unseen events
- backoff: approximate unseen probabilities by a more general probability, e.g. unigrams

Part of speech tagging

They can fish.

POS lexicon fragment:

```
they PNP
can VM0 VVB VVI NN1
fish NN1 NN2 VVB VVI
```

CLAWS 5 tagset:

```
NN1 singular noun
NN2 plural noun
PNP personal pronoun
VM0 modal auxiliary verb
VVB base form of verb
VVI infinitive form of verb
```

1. They_PNP can_VM0 fish_VVB ._PUN
2. They_PNP can_VM0 fish_NN2 ._PUN
3. They_PNP can_VVB fish_NN2 ._PUN

Stochastic POS tagging

A tag is assigned based on the lexical probability plus the sequence of prior tags.

> They used to can fish in those towns. But now few people fish in these areas.

```
They_PNP used_VVD to_TO0 can_VVI
fish_NN2 in_PRP those_DT0
towns_NN2 ._PUN But_CJC now_AV0
few_DT0 people_NN2 fish_VVB
in_PRP these_DT0 areas_NN2 ._PUN
```

sequence	count	bigram prob
DT0	3	
DT0 NN2	3	1
NN2	4	
NN2 PRP	1	0.25
NN2 PUN	2	0.5
NN2 VVB	1	0.25

Assigning probabilities

Slightly more complex than word prediction, because looking at words and tags.

Prob of tag sequence T, given word sequence W. Applying Bayes theorem:

$$P(T|W) = \frac{P(T)P(W|T)}{P(W)}$$

$P(W)$ is constant:

$$P(T|W) = P(T)P(W|T)$$

estimate $P(T)$ as $P(t_i|t_{i-1})$ (bigrams)

estimate $P(W|T)$ as $P(w_i|t_i)$ (i.e., the probability of each word given its tag)

- Actual systems use trigrams — smoothing and backoff are critical.
- Unseen words: use all possible *open class* tags, possibly restricted by morphology.

Evaluation of POS tagging

- percentage of correct tags
- one tag per word (some systems give multiple tags when uncertain)
- over 95% for English (but punctuation unambiguous)
- baseline of most common tag gives 90% accuracy
- different tagsets give slightly different results: utility of tag to end users vs predictive power (an open research issue)

Evaluation in general

Training data and test data

test data must be kept unseen, often 90% training and 10% test data

Baselines

Ceiling

human performance on the task, where the ceiling is the percentage agreement found between two annotators (*interannotator agreement*)

Error analysis

error rates are unevenly distributed

Reproducibility

Representative corpora and data sparsity

- test corpora have to be representative of the actual application
- POS tagging and similar techniques are not always very robust to differences in genre
- balanced corpora may be better, but still don't cover all text types
- communication aids: extreme difficulty in obtaining data, text corpora don't give good prediction for real data