CMPT 825
Natural Language Processing

Anoop Sarkar

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

n-grams

- A simple model of language
- Computes a probability for observed input
- Probability is likelihood of observation being generated by the same source as the training data
- Such a model is often called a *language model*

An example

- Let's consider an example we've seen before: *spelling correction*

… *was called a "stellar and versatile* ***acress*** *whose combination of sass and glamour has defined her* …

KCG model best guess is **acres**

An example

- A language model can take the context into account:
  - … *was called a "stellar and versatile **acress** whose combination of sass and glamour has defined her* …
  - … *was called a "stellar and versatile **acres** whose combination of sass and glamour has defined her* …
  - … *was called a "stellar and versatile **actress** whose combination of sass and glamour has defined her* …
- Each sentence is a sequence $w_1, \ldots, w_n$. Task is to find $P(w_1, \ldots, w_n)$.

Another example

physical Brainpower not plant is chief , now a 's asset , . firm
, a Brainpower not now chief asset firm 's is . plant physical ,
chief a physical , . firm not , Brainpower plant is asset 's now
not plant Brainpower now physical 's . a chief , asset firm , is
plant Brainpower is now , , not . firm a 's physical asset chief
physical is 's plant firm not chief . Brainpower now asset , , a
Brainpower , not physical plant , is now a firm 's chief asset .

Each sentence is a sequence $w_1, \ldots, w_n$.
Task is to find $P(w_1, \ldots, w_n)$.

How can we compute $P(w_1, \ldots, w_n)$

- Apply the *Chain rule*
- $P(w_1, \ldots, w_n) = P(w_1) \,.\, P(w_2 \mid w_1) \,.\, P(w_3 \mid w_1, w_2) \ldots P(w_n \mid w_1, \ldots, w_{n-1})$
- Each of these probabilities can be estimated (using frequency counts) from *training data*
- **But** we need to apply these probabilities on unseen *test data*
- The curse of dimensionality: **sparse data**

The Markov Assumption

a stellar and versatile ***acres*** *whose combination of*

$P(a)\,.\ P(stellar \mid a)\,.\ P(and \mid a, stellar)\,.$

$P(versatile \mid a, stellar, and)\,.$

$P(\textbf{acres} \mid a, stellar, and, versatile)\,.$

$P(whose \mid a, stellar, and, versatile, \textbf{acres}) \ldots$

a stellar and versatile ***acres*** *whose combination of*

$P(a)\,.\ P(stellar \mid a)\,.\ P(and \mid a, stellar)\,.\ P(versatile \mid stellar, and)\,.$

$P(\textbf{acres} \mid and, versatile)\,.\ P(whose \mid versatile, \textbf{acres}) \ldots$

n-grams

- *0*th order Markov model: $P(w_i)$ called a ***unigram*** model
- *1*st order Markov model: $P(w_i \mid w_{i-1})$ called a ***bigram*** model
- 2nd order Markov model: $P(w_i \mid w_{i-2}, w_{i-1})$ called a ***trigram*** model

n-grams

- How many possible distinct probabilities will be needed?, i.e. **parameter values**
- Total number of **word tokens** in our training data
- Total number of unique words: **word types** is our vocabulary size

n-gram Parameter Sizes

- Let V be the vocabulary, size of V is $|V|$
- $P(W_i = x)$, how many different values for W_i
  - $|V| = 3x10^3$
- $P(W_i = x \mid W_j = y)$, how many different values for W_i, W_j
  - $|V|^2 = 9x10^6$
- $P(W_i = x \mid W_k = z, W_j = y)$, how many different values for W_i, W_j, W_k
  - $|V|^3 = 27x10^9$

Parameter size

Corpus: said the joker to the thief

$|V| = 5$

Bigrams: max num of parameters = $|V|^2 = 25$

said | <bos>
the | said
joker | the
to | joker
the | to
thief | the

observed = $W_T = 5+1 << 25$