Ling/CSE 472: Introduction to Computational Linguistics

4/18/12

N-grams

Overview

- What are N-grams? (high-level)
- When are they useful?
- Evaluation
- Simple (un-smoothed) N-grams
- Training and test sets
- Unknown words
- N-grams and linguistic knowledge
- Next time: smoothing, back-off, interpolation

What are N-grams?

- A way of modeling the probability of a string of words.
- ... or the probability of the N+1st word being *w* given words 1-N.

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1})$$

- We'll come back to this in more detail in a moment

When are they useful?

- When would you want to know the relative probability of two strings?
- ... the relative probability of two words given the previous N words?
- ... the absolute probability of a string?

When are they useful?

- When would you want to know the relative probability of two strings?
    - Choosing among ASR candidates
    - Choosing among MT candidates
- ... the relative probability of two words given the previous N words?
    - Choosing among spell correction candidates
    - Predictive text (T9, AAC): Does your cell phone store an n-gram model?
- ... the absolute probability of a string? (kind of)
    - Spell checking: relative to a threshold
    - Language ID, authorship ID: relative to probability from some other model

Evaluating N-gram models

- What kinds of extrinsic evaluation are possible?
- What kinds of intrinsic evaluation are possible?
- What different kinds of models could you compare?

Evaluating N-gram models

- What kinds of extrinsic evaluation are possible?
    - ASR, MT, ...
- What kinds of intrinsic evaluation are possible?
    - Perplexity: Given an n-gram model trained on some training set, how well does it predict the test set? (i.e., what probability does it assign to the test set?)
- What different kinds of models could you compare?
    - Different: training data, smoothing/back-off techniques, higher-level tokens

Training and test sets

- Training and test sets must be distinct, otherwise probabilities will be artificially high

- This is just a special case of a more general reason: The purpose of test sets is to see if the method generalizes to unseen data

- Training and test sets are typically drawn from the same or comparable corpora

    - Why?

    - What if they aren’t?

Simple (un-smoothed) N-grams

- What we'd really like to calculate:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\dots P(w_n|w_1^{n-1})$$

$$= \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$

- But we'd never find a corpus big enough.

- Why not?

So: An approximation

- Markov assumption: The probability of a given word only depends on the previous word

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1})$$

- Is this assumption valid?

- (NB: That's a bigram model ... for a trigram model, we look at the previous two words, etc.)

Maximum Likelihood Estimates for bigram counts

- Bigram probability for a word *y* given a previous word *x*:

- Out of all the times you saw *x*, in what percentage was it followed by *y*?

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

Probability v. Frequency

- Probability: How likely something is to happen
- Frequency: How frequently something has happened in a set of observations
- Probability clearly influences frequency
- Frequency can be used to estimate probability (what else can?)
    - ... but they are not the same thing
- If a bigram never appears in a training corpus,
    - What is its observed frequency?
    - What is its probability?

Example

- What are the bigrams in the following mini corpus? What are their MLEs?

<s> How much wood would a wood chuck chuck if a wood chuck could chuck wood? </s> <s> As much wood as a wood chuck could if a wood chuck could chuck wood. </s>

- What probability does that bigram model assign to the following sentences?

<s> How much wood. </s>
<s> How much wood? </s>
<s> As much wood chuck chuck chuck wood. </s>
<s> How would a wood chuck chuck ? </s>

bigrams

- <s> How = 1/2
- How much = 1
- much wood = 1
- wood would = 1/8
- would a = 1
- a wood = 1
- wood chuck = 1/2
- chuck chuck = 1/7
- chuck if = 1/7
- if a = 1
- chuck could = 3/7
- could chuck = 2/3
- chuck wood = 2/7
- wood ? = 1/8
- ? </s> = 1
- <s> As = 1/2
- As much = 1/2
- wood as = 1/8
- as a = 1/2
- could if = 1/3
- wood . = 1/8
- . </s> = 1

Sentences

<s> How much wood. </s>
<s> How much wood? </s>
<s> As much wood chuck chuck chuck wood. </s>
<s> How would a wood chuck chuck ? </s>

1. 1/2 * 1 * 1 * 1/8 * 1 = 1/16
2. 1/2 * 1 * 1 * 1/8 * 1 = 1/16
3. 1/2 * 1/2 * 1 * 1/2 * 1/7 * 1/7 * 2/7 * 1 = 1/1372
4. 1/2 * 0 ... = 0

Counting things in a corpus

- Type/token distinction

- But what counts as a token? What are some cases where this is not obvious?

- And what counts as the same type? What are some cases where this is not obvious?

- Is there a single right answer?

Counting things in a corpus

- Type/token distinction
- But what counts as a token? What are some cases where this is not obvious?
    - Contracted forms, punctuation, hyphenated forms, words with spaces (*New York*), ...
- And what counts as the same type? What are some cases where this is not obvious?
    - Caps/non-caps, word-form/lemma, homographs, ...
- Is there a single right answer?
    - No: It depends on the application context

Unknown words

- What would a n-gram model trained as described so far say about the probability of a sentence with an unknown word in it?

- What could be done about that?

Unknown words

- What would a n-gram model trained as described so far say about the probability of a sentence with an unknown word in it? -- 0

- What could be done about that?

- Choose a vocabulary smaller than the actual V, and replace all other words with <UNK> -- Any words you might want to be sure to include in V?

- Or: Replace the first occurrence of every word in the training set with <UNK>

- Then: Estimate probabilities of <UNK> just like any other word

N-grams and linguistic knowledge

- Is an n-gram model a grammar?
- What kinds of information about a language does it capture?
- What kinds of information about a language does it miss?

N-grams and linguistic knowledge

- N-gram models are supposed to be “language-independent” in that they don’t require specific knowledge about a language to create --- just text.

- Would you expect them to work equally well across languages? Why or why not?

Overview

- What are N-grams? (high-level)
- When are they useful?
- Evaluation
- Simple (un-smoothed) N-grams
- Training and test sets
- Unknown words
- N-grams and linguistic knowledge
- Next time: smoothing, back-off, interpolation