Semantic Ambiguity
Same trees have different meaning

- *Discourse*: The meeting is cancelled. Nicholas isn't coming into the office today.
- *Word senses*: bank (finance or river)
- *Quantifier scope*: Every child loves some movie

Structural (syntactic) Ambiguity
Different trees produce the same sentence.

- *Homophones*: blew and blue (particularly in speech)
- *Part of speech*: chair (noun or verb)
- *PP-attachement*: I saw a girl with a telescope
- *Reference*: John dropped the goblet onto the glass table and it broke

PP-Attachement
Put the block <u>in the box</u> <u>on the table</u> <u>in the kitchen</u>.

- *Put the block ((in the box on the table) in the kitchen)*
- *Put the block (in the box (on the table in the kitchen))*
- *Put ((the block in the box) on the table) in the kitchen*

n preposition phrases have $\mathrm{Cat}_n = \binom{2n}{n} - \binom{2n}{n-1} \approx \frac{4^n}{n^{3/2}\sqrt{\pi}}$ different parses.

Variability
Multiple different sentences can have the same meaning.

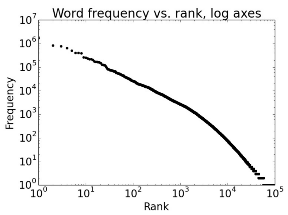
- He drew the house
- He made a sketch of the house
- He portrayed the house in his paintings

Zipf's Law
Infrequent words will make up the significant majority of the corpora. For this reason models have to estimate probabilities for things we rarely or never see in training.

$$f \times r \approx k$$

$$\log f = \log k - \log r$$

- f : frequency of a word
- r : rank of a word (if sorted by frequency)
- k : constant



Polysemy Homonymy
Homonymy: Two words with the same form but different meanings with different origins.
e.g. *orange*
Polysemy: Two words with the same form but different meanings with the same origin.
e.g. *bank*
Checking for Constiuency
These are groups of words or phrases that can

- be used with conjunction words like and, or, and but. (e.g. washed and peeled)
- substitute one word or phrase for another.
- appear in the frame "____ is/are/who/what/where/when/why/how ..." e.g.

They put the boxes in the basement.
In the basement is where *they put the boxes.*
Other Reasons

- Context Dependence
- Unknown representation
- Diversity in Language (POS order, number of noun cases (he/him in english, 10+ in russian))

Language is not determinsitic hence FSMs and CYK are useful.

Context/Coprpora

- Topic (sports, politics science)
- Mode of communication (speech, writing)
- Genre (news, fiction, scientific)
- Audience (formality, complexity)

Important to choose a corpus relevant to the task.

Human Annotation / Gold Lables
Often included in the *metadata* of a corpus.
Ofted used as **gold labels**, the best possible labels for a corpus. Gold labels are not always perfect, sources of errors include:

- Simple error (hitting the wrong button)
- Not reading the full context
- Not noticing an erroneous pre-annotation
- Forgetting a detail from the guidelines
- Cases not anticipated by or not fully specified in guidelines (room for interpretation)
- Ambiguity

To resolve we must consider

- Inter-annotator agreement
- Annotation guidelines
- Annotation tools

Sentiment Lexicon
A list of words with their associated sentiment (good / bad). Used as a tool in sentiment analysis, issues include ambiguity, sarcasm, context dependence.
Normalization/Pre-processing

- Tokenization
- Lowercasing
- Stopword removal
- Stemming
- Punctuation removal
- Spelling correction
- Sentence splitting
- Part of speech tagging
- Named entity recognition
- Parsing

Linguistics

In English, whole words are constructed by combining stems and affixes.

- **Stems**: base (dictionary) words (house, combine, eat, walk, ...)
- **Affixes**: changes the grammar of a word (prefixes, suffixes, infixes, and circumfixes)

Inflection (stem + grammar affix): no change the grammatical category (walk → walking)
Derivation (stem + grammar affix): change to grammatical category (combine → combination)
Compounding (stems together): dog, house → doghouse
Cliticisation: I've, we're, he's, ...

N-gram Models

- Text-generation
- Text-classification
- POS-tagging
- Named-entity recognition

Trigram equation

$$P_{MLE}\left(w_i \mid w_{i-2}, w_{i-1}\right) = \frac{C\left(w_{i-2}, w_{i-1}, w_i\right)}{C\left(w_{i-2}, w_{i-1}\right)}$$

Start/end of sentence tags and *costs* are used.

Markov Assumption the probability of a word only depends of a fixed number N of previous words.

Order Higher order N-grams are more context-sensitive but suffer from sparsity, whereas lower order N-grams have reduced context but are less sparse.

Smoothing reduces sparse data problem. Smoothing methods below assign equal prob to all unseen events if interpolation and back-off are not used.

Add-α (Lidstone) Smoothing: Add α to all counts and normalise. We choose $\alpha < 1$ that minimises loss on the dev set. *NB:* $\alpha = 1 \implies$ Laplace smoothing.

- *Advantages*: simple, easy to implement
- *Disadvantages*: overestimates the probability of unseen events, assumes v is known

Let $v :=$ vocab size

$$P_{+\alpha}\left(w_i \mid w_{i-1}\right) = \frac{C\left(w_{i-1}, w_i\right) + \alpha}{C\left(w_{i-1}\right) + \alpha v}$$

Katz-Backoff: If $C(w_{i-1}, w_i) = 0$, back-off to lower order N-grams, using a *back-off weight α* to choose how the probability mass is distributed. Otherwise just use a discounted probability P^*.

- *Advantage*: Looking at smaller n-grams allows us to look at words in less context, allowing the model to generalise to new contexts easier.
- *Difference from Good-Turing*: Distribute mass across lower order n-grams using weights instead of uniformly across all unseen n-grams.

$P_{\mathrm{BO}}\left(w_3 \mid w_1, w_2\right)=$
$$\begin{cases}P^*\left(w_3 \mid w_1, w_2\right), & \text{if } C\left(w_1, w_2, w_3\right)>0 \\ \alpha_2 P_{\mathrm{BO}}\left(w_2 \mid w_1\right), & \text{otherwise.}\end{cases}$$

Interpolation: Combine estimates from all n-grams using weights λ_i.
Each λ_i must sum to 1. They can be *constant* or *context-dependent* (i.e. tuned using dev set).

$$\begin{aligned}\hat{P}\left(w_3 \mid w_1 w_2\right)= & \lambda_1 P\left(w_3\right)+\lambda_2 P\left(w_3 \mid w_2\right) \\ & +\lambda_3 P\left(w_3 \mid w_1 w_2\right)\end{aligned}$$

Good-Turing: Distribute probability mass unniformly across unseen n-grams.
$N_c :=$ number of n-grams seen c times
$N :=$ total seen n-grams
$c :=$ actual count
$c^* :=$ adjusted count
$P_c^* :=$ adjusted probability for an n-gram seen c times.
NB: If we don't know N_0, let $N_0 = N_1$ or for bigrams we can *estimate* with $N_0 = V^2 - N$.

$$c^*=(c+1) \frac{N_{c+1}}{N_c} \qquad P_c^*=\frac{c^*}{N}$$

Kneser-Ney: **no. 1 smoothing method!!**
Count how many times a word occured with a unique preceding word (distinct histories) and MLE.
Avoids bias in $P(york|new)$.

$$N_{1+}\left(\bullet w_i\right)=\left|\left\{w_{i-1}: c\left(w_{i-1}, w_i\right)>0\right\}\right|$$
$$P_{KN}\left(w_i\right)=\frac{N_{1+}\left(\bullet w_i\right)}{\sum_w N_{1+}(\bullet w)}$$

Evaluation for Classification

Extrinsic measure the performance of a system using a downstream application.
Intrinsic relies on measures inherent to the current task.

Metrics for Binary Classification

$$\text{accuracy}=\frac{correct}{total}=\frac{TP+TN}{TP+FP+TN+FN}$$

Doesn't work for imbalanced data i.e. mostly one class.

$$\text{precision} =\frac{\text{correct +ive tags}}{\text{total tagged}}=\frac{TP}{TP+FP}$$
$$\text{recall} =\frac{\text{correct +ive tags}}{\text{total +ive data}}=\frac{\text{TP}}{\text{TP + FN}}$$
$$F_1=\frac{2 \cdot P \cdot R}{P+R}$$

Confusion Matrix
Plot gold labels against output

		gold labels		
		urgent	normal	spam
system output	urgent	8 TP	10	1
	normal	5	60 TP	50
	spam	3	30	200 TP

Metrics for Multi Classification
Combine precision and recall micro and macro averaging.