Exponential models for natural language

Doug Beeferman, Adam Berger, John Lafferty
School of Computer Science
Carnegie Mellon University

Predicting the next word: The language modeling problem

- Shannon: The entropy of English is at most 1.6 bits per character
- The Shannon guessing game:

The	prodigal	son	wreaks	havoc	with	his	father's	Porsche.
3	1292	1	2341	1	8	4	6	194

- A stochastic language model $P(w_0|h)$ assigns to a word history h a probability distribution over the vocabulary.
- N-gram models, of the form $P(w_0|h) = P(w_0|w_{-n+1}, w_{-n+2}, \ldots, w_{-1})$, treat natural language as a stationary Markov process.
- But language is nonstationary!

Features

- For genericity, let's allow arbitrary *features* instead of just n-grams.
- Define a feature to be a function mapping histories h paired with the current word w_0 to a real value.
- For now let's restrict ourselves to binary-valued features (indicator functions).
- Possible features:

$f_{\mathrm{NA}}(h, w_0) = 1$ if the last noun in h is singular and w_0 is a singular verb; or 0 otherwise.	a syntactic feature for number agreement
$f_{\mathrm{C}}(h, w_0) = 1$ if w_0 has appeared in the current document already; or 0 otherwise.	a word cache feature
$f_{x,y}(h, w_0) = 1$ if the last two words in h are (x, y); or 0 otherwise.	a trigram feature for (x, y)

Choosing a model to fit the data: Maximum Entropy

Given:

- A training corpus of text with empirical distribution $\tilde{p}$
- A vector of features f
- Some *prior* model q_0 of the data

There are two natural probability distributions we might use as our model.

- Within the expectation-constrained set,

  $$\mathcal{P}(f, \tilde{p}) = \{ \; p : \; p[f] = \tilde{p}[f] \; \}$$

  where $p[f]$ denotes expectation with respect to $\tilde{p}(h)\, p(y \,|\, h)$, choose the distribution that has *maximum entropy* relative to q_0, *i.e.*

  $$q_\star^{\mathrm{ME}} = \operatorname*{arg\,min}_{p \in \mathcal{P}(f, \tilde{p})} D(p \,\|\, q_0) \,.$$

Choosing a model, cont'd: Maximum Likelihood

- Within the exponential family:

$$\mathcal{Q}(f, q_0) = \left\{ \frac{e^{\lambda \cdot f(h,y)}}{Z_\lambda(h)} q_0(y \,|\, h) : \ \lambda \in \mathbb{R}^n \right\}.$$

  choose the distribution that has *maximum likelihood* with respect to $\tilde{p}$, *i.e.*

$$q_\star^{\mathrm{ML}} = \arg\min_{q \in \mathcal{Q}(f, q_0)} D(\tilde{p} \,\|\, q).$$

- These criteria determine the same distribution!

$$q_\star = q_\star^{\mathrm{ML}} = q_\star^{\mathrm{ME}} \in \mathcal{P}(f, \tilde{p}) \cap \mathcal{Q}(f, q_0)$$

Trigger pair features

- The most widely used nonstationary features are *trigger pairs* that capture word associations over long distances.
- Formally, a trigger pair feature $f_i = f_{(x_i, y_i)}$ is an indicator function that is 1 if $w_0 = y_i$ and x_i appears somewhere in a window of the c previous words; or 0 otherwise.
- Tied to f_i is a weight λ_i, such that seeing x_i boosts the probability of seeing y_i within the next c words by a factor of e^{λ_i}.

WORD$_L$	WORD$_R$	λ	
Ms.	her	0.47	(weaker correlation)
I	me	0.77	
energy	gas	0.77	
committee	representative	0.85	
Japanese	Tokyo	0.86	
lieutenant	colonel	0.97	
board	board	0.98	
Soviet	missiles	1.04	
patients	drug	1.38	
AIDS	AIDS	3.87	(stronger correlation)

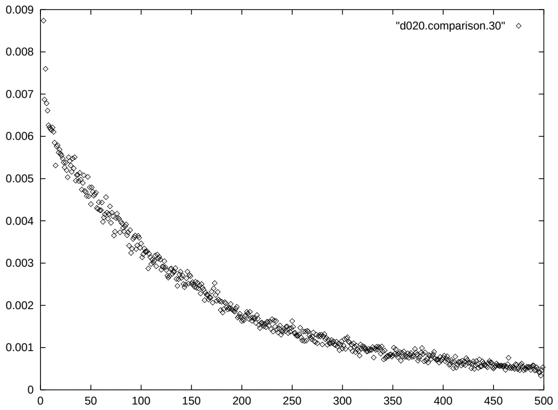
Trigram priors

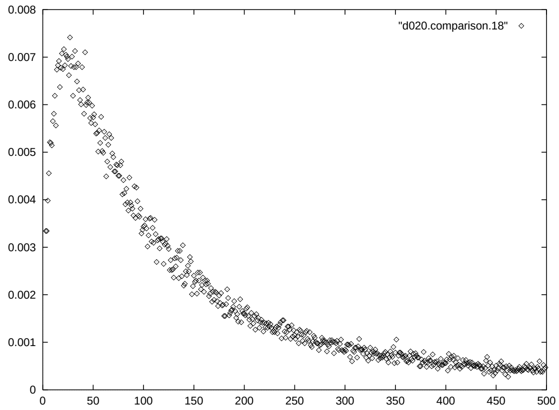
- We can train the weights λ iteratively with the Expectation Maximization (EM) algorithm.
- But these features alone do not suffice.
  - Expensive to train on large amounts of data.
  - Triggers do not capture short-range lexical constraints well.
- One idea is to throw a set of n-gram features into the mix. But this is inefficient!
- Solution: Use an n-gram model, which can be trained efficiently on gobs of data, as a prior.

$$q_0(y \mid h) = q(y \mid w_{-2}, w_{-1})$$

Extension: trigger distance modeling

- Observation: The strength of a word association depends on the time elapsed d since the triggering word was observed.
- Idea: Strengthen or weaken the contribution of trigger features f_i by a distance-dependent factor $d_k = p(d = k | f_i = 1)$.
- For non-self triggers, probability falls off exponentially.
- For self triggers, the peak probability occurs at a greater distance.

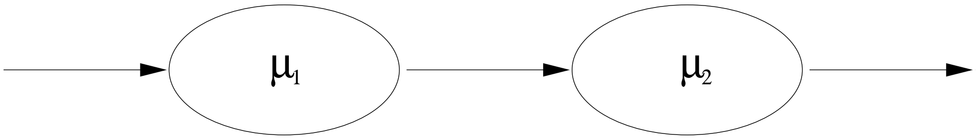


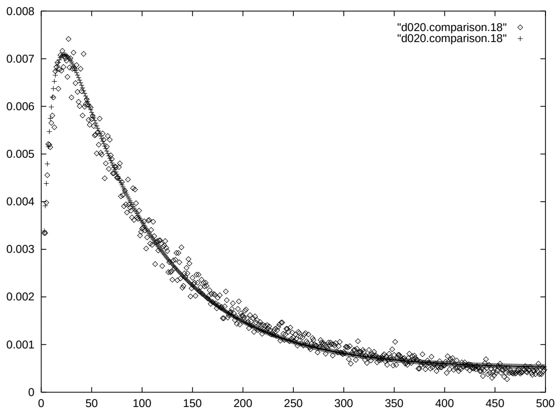


Two-stage exponential models

- The behavior for self triggers closely fits a *two-stage exponential* (Erlang) distribution.
- The form of the distance model is a convolution of two exponential models with different means.

$$P_{\mu_a,\mu_a}(k) = (1 - e^{-\mu_1})(1 - e^{-\mu_2}) \sum_{t=1}^{k} e^{-\mu_1 t}\, e^{-\mu_2 (k-t)}$$





Two-stage exponential models (cont'd)

- How do we choose two-stage exponential parameters μ_1, μ_2 that best fit some empirical data $p(d = k|f_i = 1)$?
- Treat the splitting point t between the stages as a hidden variable.
- Iteratively update initial (μ_1, μ_2) estimates with the EM algorithm:

$$
\begin{aligned}
\mu_1' &= \left(\sum_k \tilde{p}(k) \sum_{t=0}^{k} t \, p_{\mu_1,\mu_2}(t|k) \right)^{-1} \\
\mu_2' &= \left(\sum_k \tilde{p}(k) \sum_{t=0}^{k} (k - t) \, p_{\mu_1,\mu_2}(t|k) \right)^{-1}
\end{aligned}
$$

- Updates for one-stage exponential models (for non-self triggers) are similar.

Extension: context "relevance"

- Observation: Often we want to rely much more heavily on the active trigger set than on the short-range context to predict the next word.
- Idea: Strengthen or weaken the contribution of trigger features by a context-dependent factor.
- $R(w_1, w_2) = \Sigma_{w_3} \tilde{P}(w_3|w_1, w_2) \log \frac{P_{\text{exp}}(w_3|w_1,w_2)}{P_{\text{trigram}}(w_3|w_1,w_2)}$
- Defined similarly for single-word contexts.

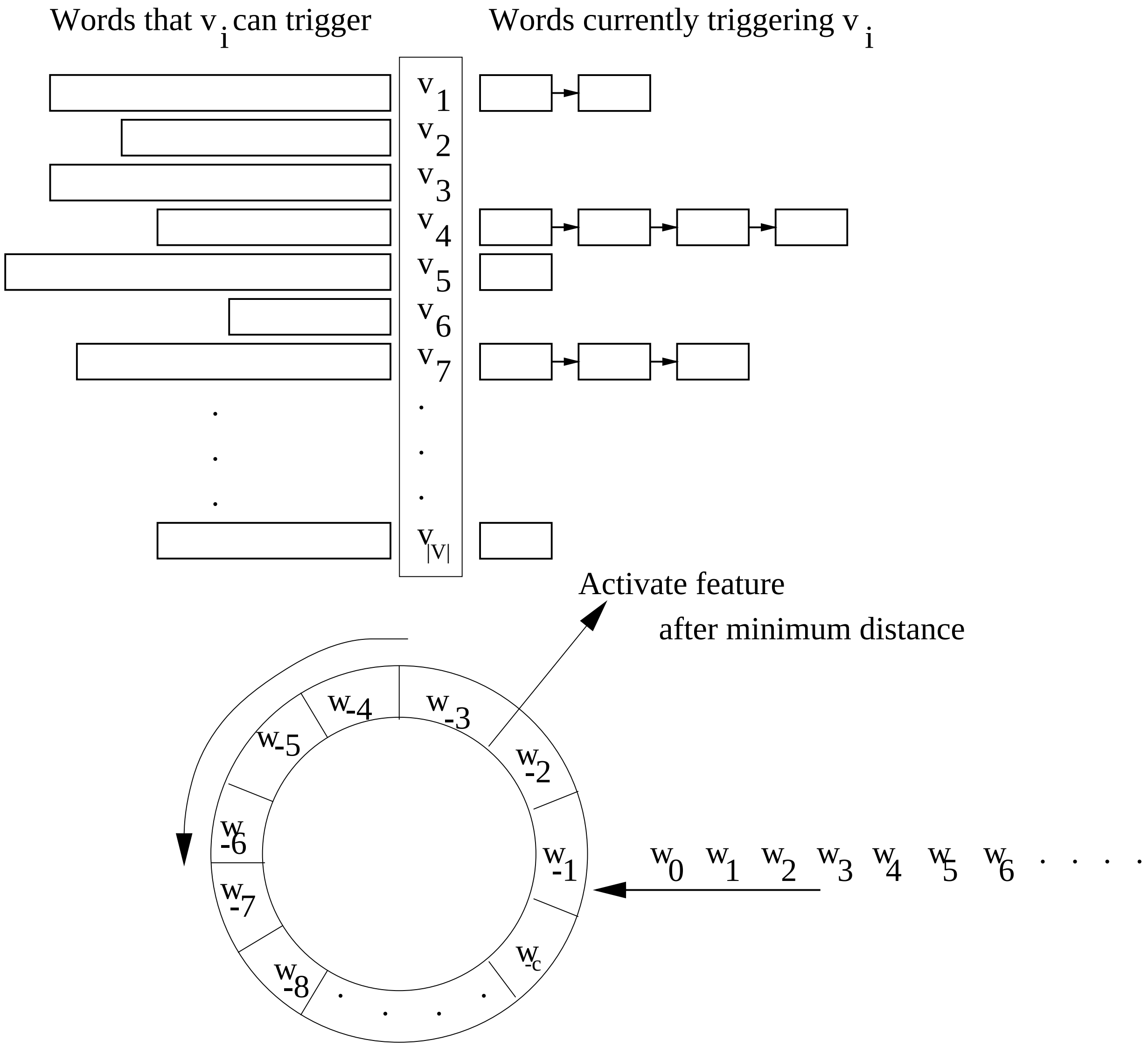
WORD	R(WORD)
MR.	1.402986
MS.	1.134188
MRS.	1.063191
DOCTOR	0.874050
WINTER	0.799196
EL	0.707947
SENATOR	0.696035
...	...
FIVE	-0.525790
EIGHTEEN	-0.569623
BILLION	-0.722252

The form of the model

$$p(w_0 \mid h) = \frac{q_{\text{tri}(\text{w}_0|\text{w}_{-2},\text{w}_{-1})}}{Z(h)} * \exp\left(\Sigma_i \Sigma_k \left[\lambda_i + \log d_k(i) + R(w_{-2}, w_{-1})\right] f_i(w_{-k}, w_0)\right)$$

- $d_k(i)$: Distance penalty for feature i at distance k (bucketed).
- $R(w_{-2}, w_{-1})$: Relevance for bigram context (w_{-2}, w_{-1}).
- $f_i(w_{-k}, w_0)$: 1 if $x_i = w_{-k}$ and $y_i = w_0$ or 0 otherwise.
- $Z(h)$: Normalization

Data structures

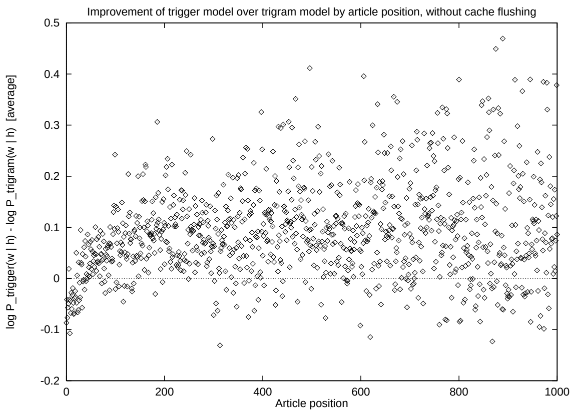


Selecting good trigger pairs

- By mutual information:
  $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.
- For a large training corpus, it's too expensive to compute $I(w_i, w_j)$ for all i and j.
- Solution: First determine all pairs that co-occur within some document in the corpus.
- Group triggers
  - Semantic classes: {`shave`} $\longrightarrow$ {`shave, cut, trim, slice`}
  - Morphological classes: {`shave`} $\longrightarrow$ {`shave, shaves, shaved shaving, shaver`}

Applications to "text tiling"

- Problem: Given a corpus of text with run-together documents, determine where the document boundaries are.
- Informedia
- Intuition: Predict a boundary where the topic of the next sentence becomes inconsistent with previous sentences.
- Informal solution: Look at where the contribution of trigger features suddenly starts to hurt more than help.



Text tiling: example

```
`` I was that deniable link between the U .S. government and the Contra
aid operation and i was supposed to be dropped like a hot rock when it
all came down '' , Colonel North said .

`` And I was willing to serve in that capacity .

I was not willing to become the victim of a criminal prosecution . ''

<art>

Fleet Financial group said its Fleet National Bank unit increased
its reserves against loans to less developed countries by forty million
dollars , causing the parent's second quarter earnings to fall forty
one percent .

For the second quarter Fleet earned nineteen point two million
dollars or thirty five cents a share compared with earnings of
thirty two point eight million dollars or sixty two cents a
share a year earlier .
```

Measuring model quality

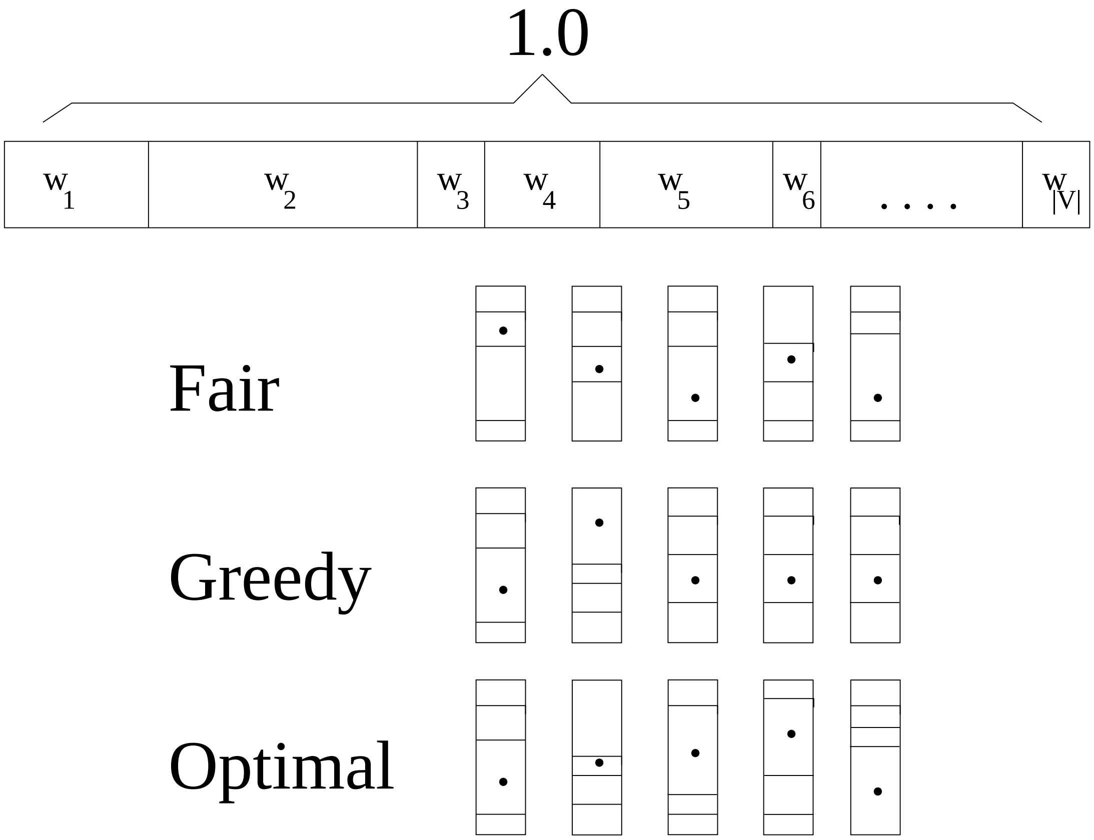
- A quantitative and objective way to assess a language model is to measure its *cross-entropy* on text that it wasn't trained on.

Experiment	*Perplexity*	*Reduction*
Baseline: trigram trained on 1M words	280.1	—
Trigram prior + 41,263 triggers	231.4	17.4%
Same as above + distance modeling	219.5	21.6%

- A qualitative and subjective way to assess a language model is to *generate* text by choosing words fairly from its distribution.

GENERATED BY THE TRIGRAM MODEL ALONE: `The creditor is a decode ago with its foreign exploration activities get related to its International Business Machines Corporation won contracts totaling thirteen percent .`
GENERATED BY THE EXPONENTIAL MODEL: `The creditor is a decade ago with its exposure to Mr. Gonzalez in the midst of a player in world` **`war`** `two lows after the` **`bloody`** `Sunday October thirty first .`

Text generation



A tough row to hoe

- We're slowly approaching the entropy estimates for English, but we have a long way to go. It's still not clear whether computers can outperform humans at this task.
- Syntactic knowledge should help, but it hasn't proven very useful yet.
- Exponential models give us a clean and principled way to experiment with new features.