# Chapter 3:

# Basics of Language Modelling

# Motivation

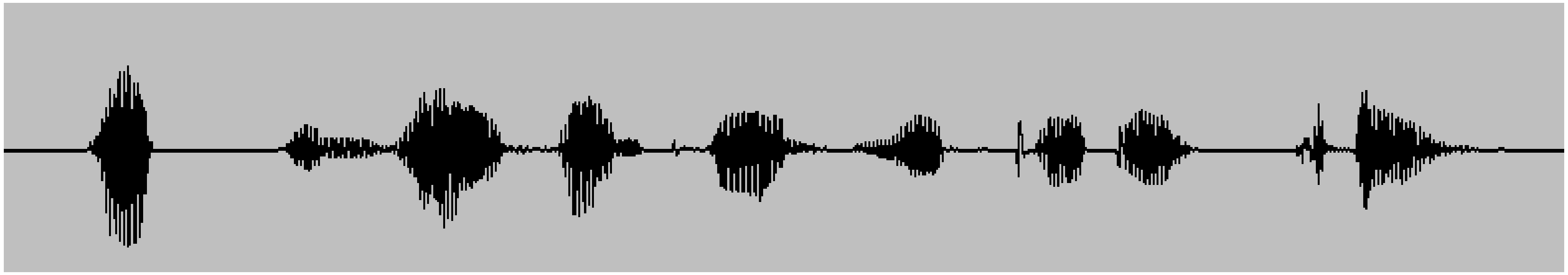
Language Models are used in

- Speech Recognition
- Machine Translation
- Natural Language Generation
- Query completion
- Caption Generation

Quality of LMS:

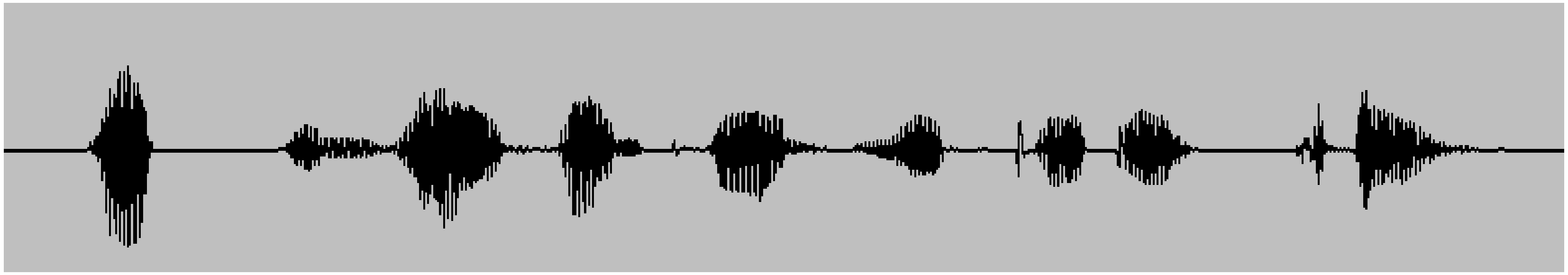
need a simple evaluation metric for fast turn-around times in experiments

# Test your Language Model



What's in your hometown newspaper ???

# Test your Language Model

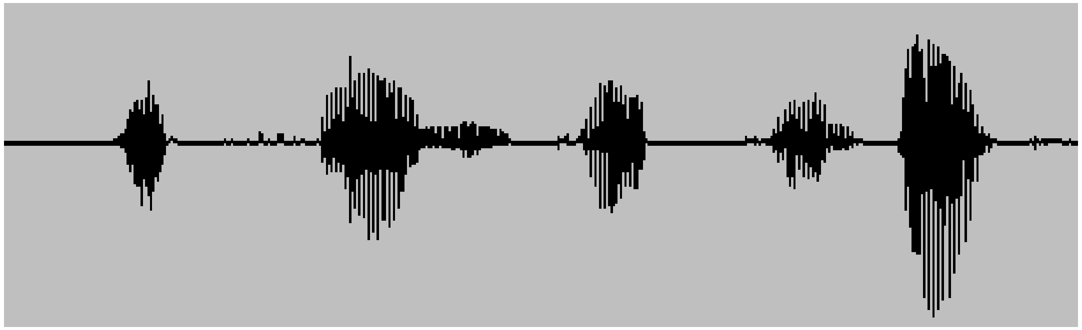


What's in your hometown newspaper today



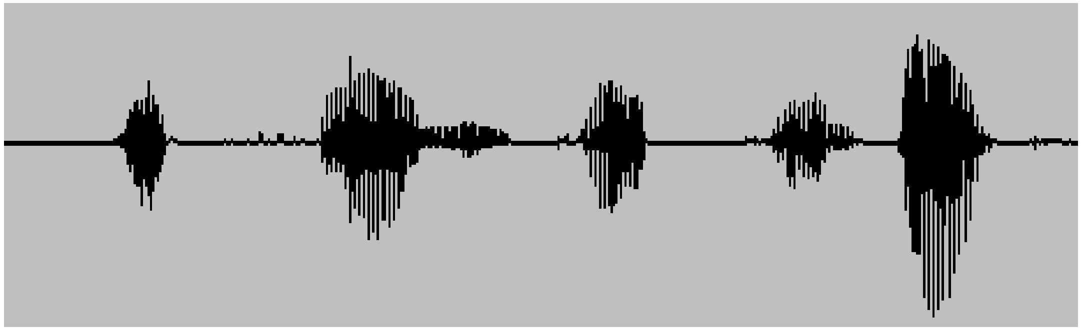
4

# Test your Language Model



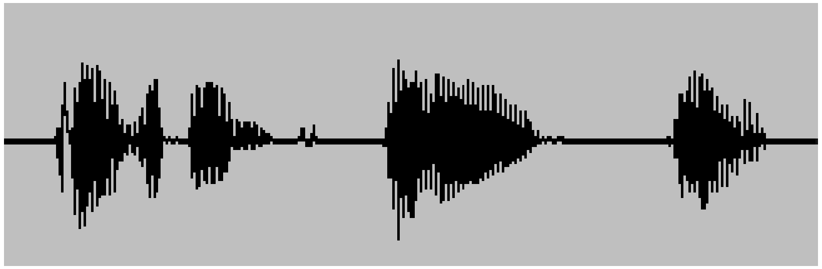
It's raining cats and ???

# Test your Language Model



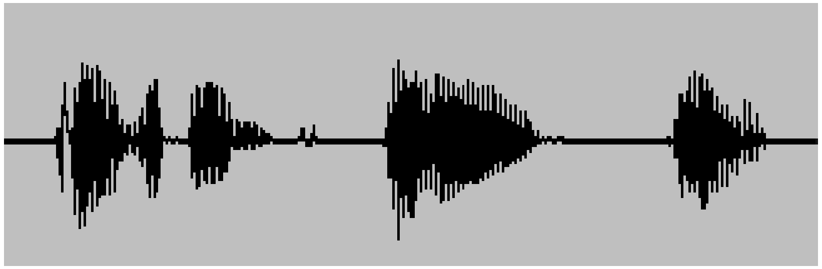
It's raining cats and dogs

# Test your Language Model



President Bill ???

# Test your Language Model



President Bill Gates

# Definition Language Model

A language model either:

- Predicts the next word w given a sequence of predecessor words h (history)

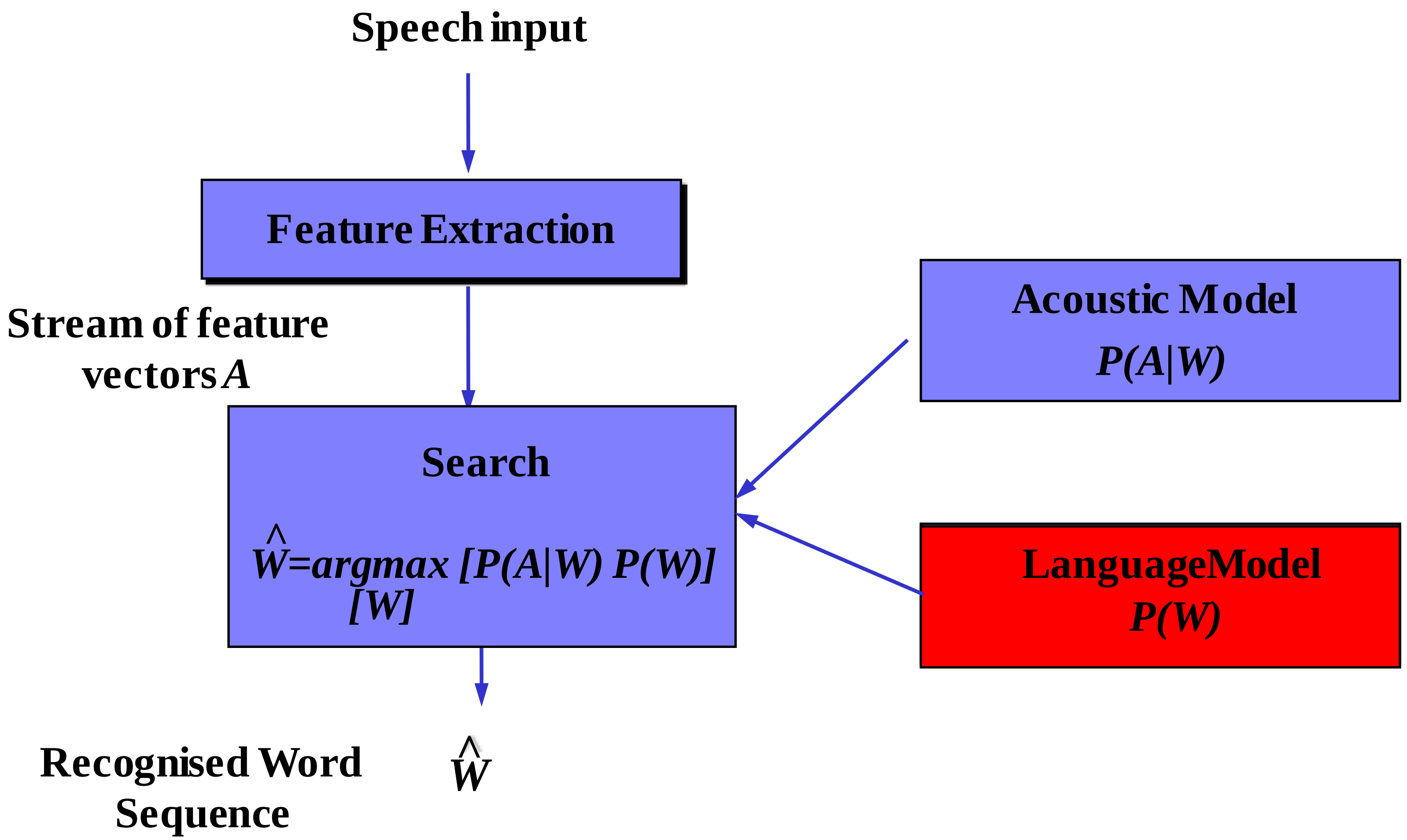
$$\mathrm{P}(\mathrm{w}_{\mathrm{i}} \mid h_i) = \mathrm{P}(\mathrm{w}_{\mathrm{i}} \mid w_1, \ldots w_{i-1}) \text{ with i : position in text}$$

or

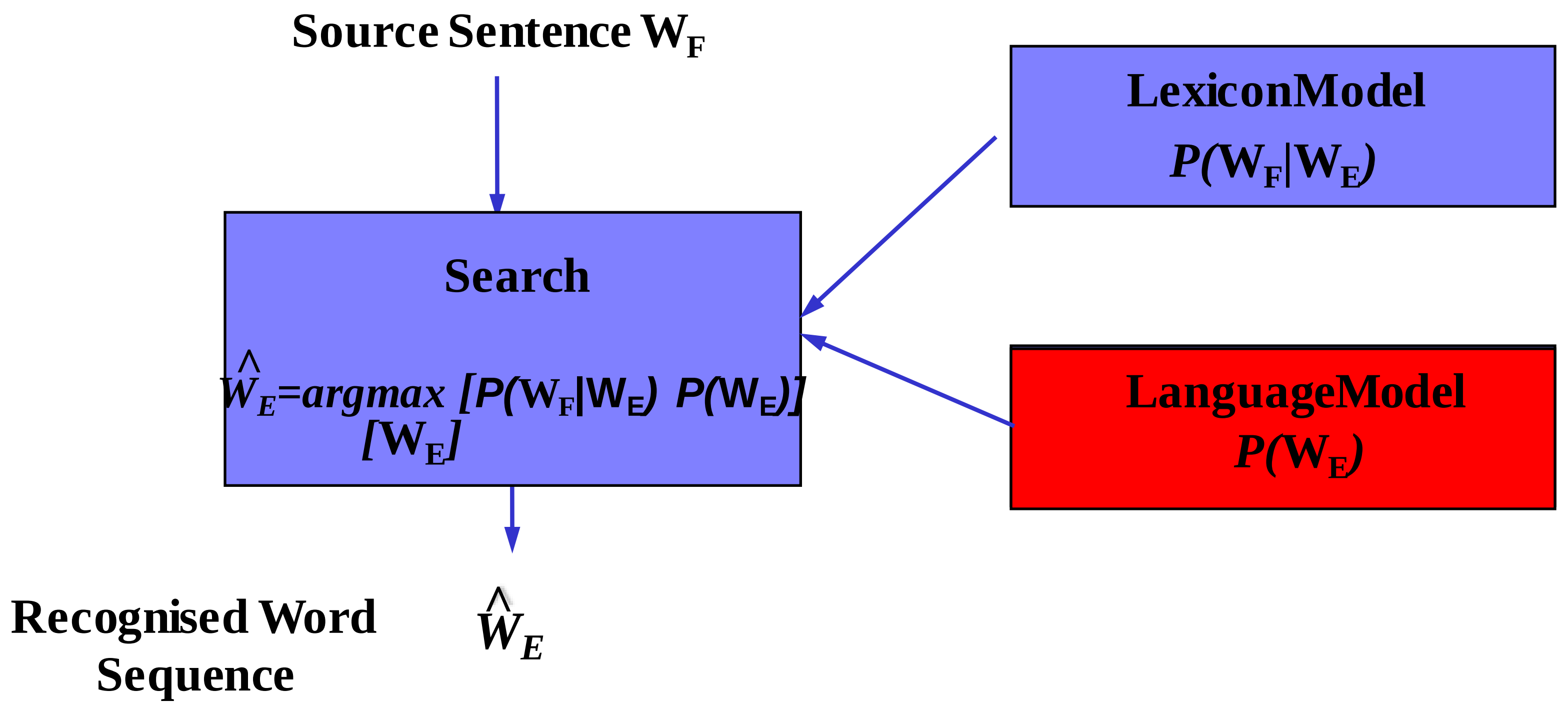
- Predicts the probability of a sequence of words

$$\mathrm{P}(w_1, w_2, \ldots\ldots w_{N-1}, .w_N) = \mathrm{P}(W)$$

# Language Model in Speech Recognition (ASR)



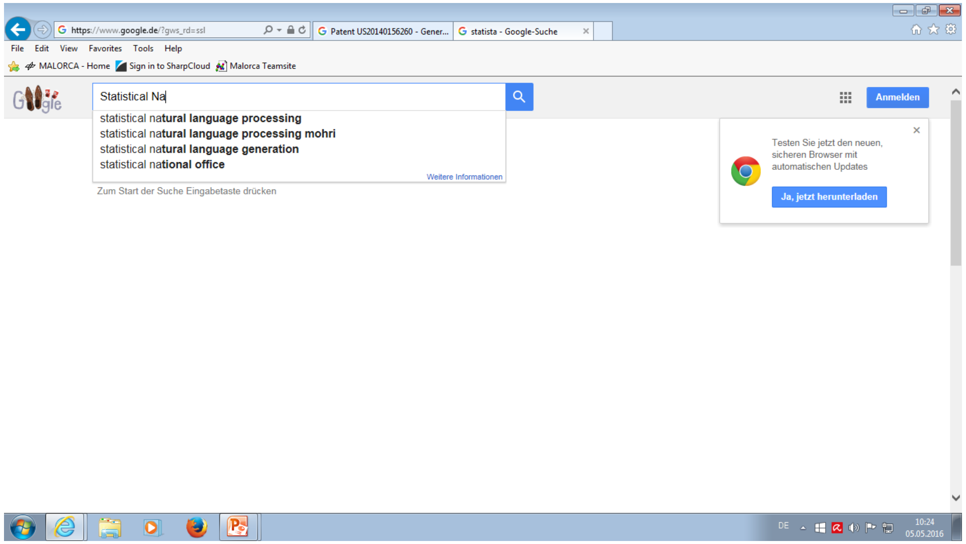
# Language Model in Machine Translation (MT)





11

# Sentence Completion




- Directly use of P(w|h)
- Needs good search algorithms



12

# The Markov Assumption: truncate the history

Cutting history to M-1 words:

$$P(w_i \mid w_1, w_2, \ldots\ldots w_{i-1}) = P(w_i \mid w_{i-M+1}, \ldots w_{i-1})$$

Special cases:

"Zerogram"
(uniform distribution)

$$P(w_i \mid w_1, w_2, \ldots\ldots w_{i-1}) \approx \frac{1}{W}$$

Unigram

$$P(w_i \mid w_1, w_2, \ldots\ldots w_{i-1}) \approx P(w_i)$$

Bigram

$$P(w_i \mid w_1, w_2, \ldots\ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

Trigram

$$P(w_i \mid w_1, w_2, \ldots\ldots w_{i-1}) \approx P(w_i \mid w_{i-2}, w_{i-1})$$

….

How would you measure the quality of an LM?



14

# Definition of Perplexity

Let $w_1$, $w_2$, … $w_N$ be an independent test corpus not used during training.

Perplexity

$$\boxed{PP = P(w_1, w_2, \ldots\ldots w_N)^{-1/N}}$$

Interpretation:
Normalized probability of test corpus

# Example: Zerogram Language Model

"Zerogram"
(uniform distribution) $\mathrm{P}(\mathrm{w_i} \mid w_1, w_2, \ldots\ldots w_{i-1}) \approx \frac{1}{\mathrm{W}}$

$$\mathrm{PP} = \mathrm{P}(w_1, w_2, \ldots\ldots w_N)^{-1/\mathrm{N}}$$

$$= \left(\prod_{i=1}^{N} \frac{1}{W}\right)^{-1/N}$$

$$= W$$

Interpretation: perplexity is the "average" de-facto size of vocabulary

# Alternate Formulation

Assumption:

Use an M-gram language model

History $h_i = w_{i-M+1} \ldots w_{i-1}$

Idea:

collapse all identical histories

# Alternate Formulation

Example:

M=2

Corpus:

“to be or not to be”

$h_2$=“to” $w_2$=“be”

$h_6$=“to” $w_6$=“be”

$\mapsto$ calculate P(“be”|“to”) only once and scale it by 2



18

# Alternate Formulation

$$\mathrm{PP} = \mathrm{P}(w_1, w_2, \ldots\ldots w_N)^{-1/N}$$

$$= \left(\prod_{i=1}^{N} P(w_i \mid h_i)\right)^{-1/N}$$

Use Bayes decomposition

Try to get rid of product

$$= \exp\left(\log\left(\prod_{i=1}^{N} P(w_i \mid h_i)\right)^{-1/N}\right)$$

$$= \exp\left(-\frac{1}{N}\log\left(\prod_{i=1}^{N} P(w_i \mid h_i)\right)\right)$$

# Alternate Formulation

$$= \exp\left(-\frac{1}{N}\log\left(\prod_{i=1}^{N} P(w_i \mid h_i)\right)\right)$$

$$= \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log\left(P(w_i \mid h_i)\right)\right)$$

$$= \exp\left(-\frac{1}{N}\sum_{w,h} N(w,h)\log\left(P(w \mid h)\right)\right)$$

N(w,h): absolute frequency of sequence h,w on test corpus

$$= \exp\left(-\sum_{w,h} f(w,h)\log\left(P(w \mid h)\right)\right)$$

f(w,h) relative frequency of sequence h,w on test corpus

# Alternate Formulation: Final Result

$$\boxed{\begin{aligned} PP &= P(w_1 \ldots w_N)^{-1/N} \\ &= \exp\left( -\sum_{w,h} f(w,h) \log\big(P(w \mid h)\big) \right) \end{aligned}}$$

Perplexity can also be calculated using conditional probabilities trained in the training corpus
and
relative frequencies from the test corpus.

# Alternate Formulation: Zerogram Example

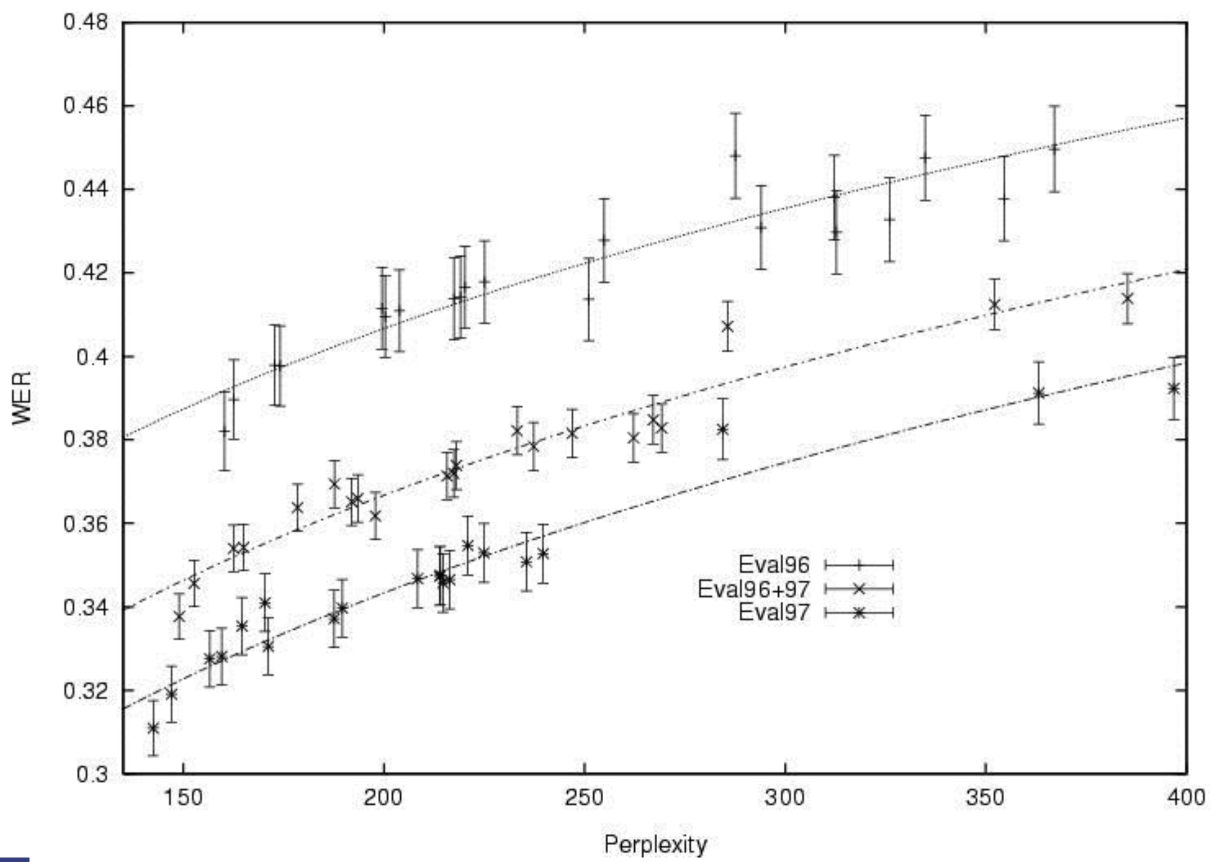
"Zerogram" $\mathrm{P}(\mathrm{w_i} \mid w_1, w_2, \ldots.. w_{i-1}) \approx \frac{1}{\mathrm{W}}$
(uniform distribution)

$$PP = \exp\left(-\sum_{w,h} f(w,h)\log\big(P(w \mid h)\big)\right) = \exp\left(-\sum_{w,h} f(w,h)\log\left(\frac{1}{W}\right)\right)$$

$$= \exp\left(-\log\left(\frac{1}{W}\right)\sum_{w,h} f(w,h)\right) = \exp\big(\log(W)*1\big) = W$$

Identical result

# Perplexity and Error Rate




Perplexity is correlated to word error rate

Power law relation

# Quality Measure: Mean Rank

- Definition:
  - Let w follow h in the test text
  - Sort all words after a given history according to p(w|h)
  - Determine position of correct word w
  - Average over all events in the test text



24

# Alternative: Average Rank

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | The | are | **to** | know | the | issues | necessary |
| 2 | This | will | | have | this | problems | data |
| 3 | One | the | | understand | these | **the** | information |
| 4 | Two | would | | do | problems | | above |
| 5 | A | also | | get | any | | other |
| 6 | Three | do | | the | a | | time |
| 7 | Please | **need** | | use | problem | | people |
| 8 | In | | | provide | them | | operators |
| 9 | **We** | | | insert | **all** | | tools |
| . | | | | . | | | . |
| . | | | | . | | | . |
| . | | | | . | | | . |
| 93 | | | | request | | | factors |
| 94 | | | | respond | | | facts |
| 95 | | | | supply | | | I |
| 96 | | | | write | | | jobs |
| 97 | | | | me | | | MVS |
| 98 | | | | **resolve** | | | old |
| . | | | | | | | . |
| . | | | | | | | . |
| . | | | | | | | . |
| 1636 | | | | | | | mailroom |
| 1637 | | | | | | | marketplace |
| 1638 | | | | | | | provision |
| 1639 | | | | | | | reception |
| 1640 | | | | | | | shop |
| 1641 | | | | | | | **important** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | role | and | **the** | **next** | be | meeting | of |
| 2 | thing | from | | | **two** | months | . |
| 3 | that | in | | | | years | |
| 4 | to | to | | | | meetings | |
| 5 | contact | are | | | | to | |
| 6 | parts | with | | | | week | |
| 7 | point | were | | | | **days** | |
| 8 | for | requiring | | | | | |
| 9 | **issues** | still | | | | | |
| . | | . | | | | | |
| . | | . | | | | | |
| . | | . | | | | | |
| 61 | | being | | | | | |
| 62 | | during | | | | | |
| 63 | | i | | | | | |
| 64 | | involved | | | | | |
| 65 | | would | | | | | |
| 66 | | **within** | | | | | |

Task:
guess the next word

Metric:
Measure how many attempts it takes

# Quality Measure: Mean Rank

| Measure | Correlation with ASR error rate |
|---|---|
| Perplexity | 0.955 |
| Mean Rank | 0.957 |

Mean rank equally good in predicting ASR performance

# Summary

- Language models predict the next word
- Applications:
  - Speech Recognition
  - Machine Translation
  - Natural Language Generation
  - Information retrieval
  - Sentence Completion
- Perplexity:
  - Measures quality of language model