An Empirical Investigation of Discounting in Cross-Domain Language Models

Greg Durrett and Dan Klein
UC Berkeley
Discounting

\[ \text{Train} \]

1

\[ \text{Test} \]

\[ c(\text{test}) = 2 \]
Discounting

Train

1
2
3

c*(1) = 0.28

c*(2) = 0.97

c*(3) = 1.96

Test

“future count”
Experimental Setup

NYT 1995

Train

NYT 1995A

Test

NYT 1995B

[Church and Gale, 1991]
Future Counts

![Graph showing the relationship between future counts and trigram count in train. The graph has a linear trend line with data points plotted along it. The x-axis represents the trigram count in train, ranging from 0 to 20, and the y-axis represents future counts, ranging from 0 to 20. The trend line indicates a direct proportionality between the two variables.]
Empirical Discounts

Graph showing the empirical discounts against the trigram count in the training dataset.

Y-axis: Empirical discounts
X-axis: Trigram count in train
Experimental Setup

NY Times

NYT 1995

NYT 1995A

NYT 1995B

NYT 1996

NYT 1996

NYT 2006

NYT 2006
Empirical Discounts

Average empirical discount

Trigram count in train

NYT95B
Empirical Discounts

![Chart showing empirical discounts across trigram counts in train data]
Empirical Discounts

Changing the training year results in growing discounts
Experimental Setup

NYT 1995

NY Times

AFP 1995

Agence France-Presse

NYT 1995A

Train

NYT 1995B

Test

AFP 1995
Empirical Discounts

More discount growth as temporal divergence increases and source changes
Predicting Discount Growth

![Discount vs Divergence Diagram]

- Discount for count-30 trigrams
- Divergence

Note: The diagram shows a single point indicating the relationship between discount and divergence.
More divergence yields more discounting
A Growing Discounts LM

- Interpolated Kneser-Ney

\[
p(n\text{-gram}) = \frac{c(n\text{-gram}) - d}{c(\text{context})} + \ldots
\]

- Linearly Growing Discounts

\[
p(n\text{-gram}) = \frac{c(n\text{-gram}) - (d_1 + d_2c(n\text{-gram}))}{c(\text{context})} + \ldots
\]
Perplexity Results

(Parameters tuned to maximize held-out NYT perplexity)
Perplexity Results

In-Domain (NYT)

\[ d_2 = 0.083 \]

(Parameters tuned to maximize held-out NYT perplexity)
Perplexity Results

In-Domain (NYT)

\[ d_2 = 0.083 \]

(Parameters tuned to maximize held-out NYT perplexity)
Perplexity Results

In-Domain (NYT)

\[ d_2 = 0.083 \]

Out-of-domain (AFP)

\[ d_2 = 0.29 \]

(Parameters tuned to maximize held-out NYT perplexity)

Constant discounts (mod. KN)
Growing discounts (this work)
Conclusion

- Shape of discount must be changed (should grow with n-gram count) as corpora diverge

- Subtle cross-domain effects suggest using a qualitatively different model
Thank you!