Evaluation of Language Models

Evaluation of Language Models

- Does the language model prefer good sentences over the bad ones?

Evaluation of Language Models

- Does the language model prefer good sentences over the bad ones?
- Does it assign higher probability to real or frequently observed sentences?

Evaluation of Language Models

- Does the language model prefer good sentences over the bad ones?
- Does it assign higher probability to real or frequently observed sentences?
- Does it assign lower probability to grammatically incorrect or rarely observed sentences?

Evaluation of Language Models

- Extrinsic Evaluation:

Evaluation of Language Models

- Extrinsic Evaluation:
  - Using some external source to evaluate our model

Evaluation of Language Models

- Extrinsic Evaluation:
  - Using some external source to evaluate our model
  - Train the language model

Evaluation of Language Models

- Extrinsic Evaluation:
  - Using some external source to evaluate our model
  - Train the language model
  - Use it for other tasks like machine translation, automatic speech recognition, spelling correction, etc

Evaluation of Language Models

- Extrinsic Evaluation:
  - Using some external source to evaluate our model
  - Train the language model
  - Use it for other tasks like machine translation, automatic speech recognition, spelling correction, etc

- Challenge with Extrinsic Evaluation:

Evaluation of Language Models

- Extrinsic Evaluation:
  - Using some external source to evaluate our model
  - Train the language model
  - Use it for other tasks like machine translation, automatic speech recognition, spelling correction, etc

- Challenge with Extrinsic Evaluation:
  - Time consuming, can take days or even months

Evaluation of Language Models

- Intrinsic Evaluation:

Evaluation of Language Models

- Intrinsic Evaluation:
  - Measure how good we are at modeling language

Evaluation of Language Models

- Intrinsic Evaluation:
  - Measure how good we are at modeling language
  - Gives a quick performance estimate

Evaluation of Language Models

- Intrinsic Evaluation:
  - Measure how good we are at modeling language
  - Gives a quick performance estimate
  - Commonly used evaluation metric: Perplexity

Perplexity

- How are you?
- I am doing great!
- Let’s play football

Perplexity

- How are you?
- I am doing great!
- Let’s play football

- High Probability

Perplexity

- How are you?
- I am doing great!
- Let’s play football

- High Probability
- Low Perplexity

Perplexity

- How are you?
- I am doing great!
- Let’s play football

- How are us?
- I doing am
- Can you does it?

- High Probability
- Low Perplexity

Perplexity

- How are you?
- I am doing great!
- Let’s play football

- How are us?
- I doing am
- Can you does it?

- High Probability
- Low Perplexity

- Low Probability

Perplexity

- How are you?
- I am doing great!
- Let’s play football

- How are us?
- I doing am
- Can you does it?

- High Probability
- Low Perplexity

- Low Probability
- High Perplexity

Perplexity

- Inverse probability of the test set

Perplexity

- Inverse probability of the test set

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

Perplexity

- Inverse probability of the test set
- Lower the perplexity, better the model

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

Thank You