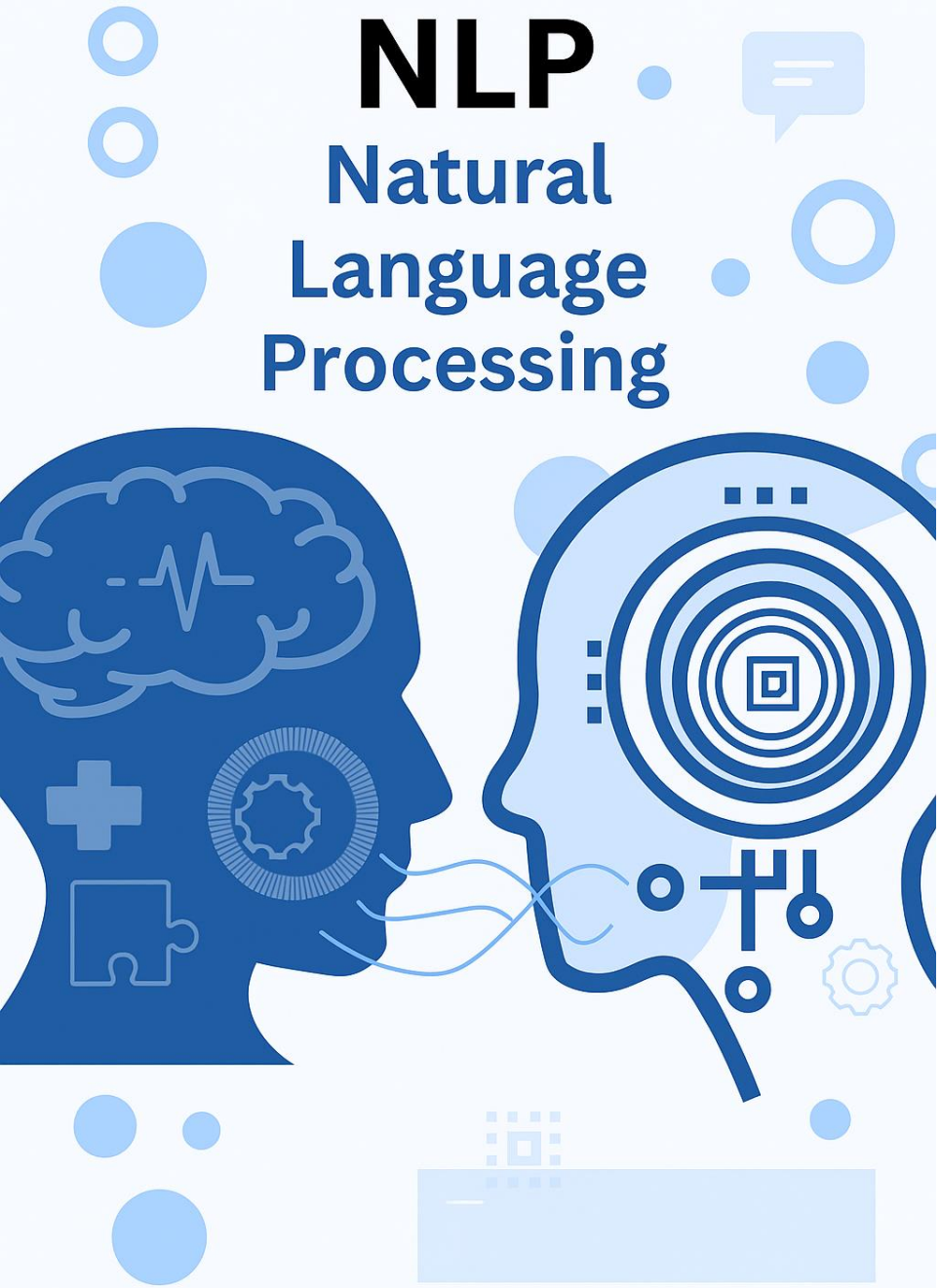

NATURAL LANGUAGE PROCESSING (NLP)

PMDS606L

MODULE 3

LECTURE 4

Dr. Kamanasish Bhattacharjee

Assistant Professor

Dept. of Analytics, SCOPE, VIT

EXTRINSIC EVALUATION

- The best way to evaluate the performance of a language model is to embed it in an application and measure how much the application improves.

- Extrinsic evaluation is the only way to know if a particular improvement in the language model (or any component) is really going to help the task at hand.

- Thus for evaluating n-gram language models that are a component of some task like speech recognition or machine translation, we can compare the performance of two candidate language models by running the speech recognizer or machine translator twice, once with each language model, and seeing which gives the more accurate transcription.

INTRINSIC EVALUATION

- Unfortunately, running big NLP systems end-to-end is often very expensive.
- Instead, it's helpful to have a metric that can be used to quickly evaluate potential improvements in a language model.
- An intrinsic evaluation metric is one that measures the quality of a model independent of any application.
- **Perplexity**: The standard intrinsic metric for measuring language model performance for n-gram language models.

MAIN GOALS OF EVALUATION

- Does the model assign **high probability** to valid sentences?
- Does it **generalize** well to unseen text?
- Is it **better** than other models for the same task?

PERPLEXITY

$$\begin{aligned}\text{perplexity}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}}\end{aligned}$$

Or we can use the chain rule to expand the probability of W:

$$\text{perplexity}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

UNIGRAM PERPLEXITY

$$\text{perplexity}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i)}}$$

BIGRAM PERPLEXITY

$$\text{perplexity}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

WALL STREET JOURNAL (WSJ)

- Trained unigram, bigram, and trigram grammars on **38 million words** from the **Wall Street Journal** newspaper.
- Then perplexity was computed for each of these models on a WSJ test set.

	Unigram	Bigram	Trigram
Perplexity	962	170	109