Words as Sources: Entropy and Redundancy in the Embedding Space

Language can be seen as a great communication network whose nodes are words, each acting as a *source of information*. They can also be seen as outputs of world-word mappings, where each word encodes a concept or object in the external world. In this case we may see their semantic contentn as a signal transmitted through a noisy channel.

In NLP representation learning, the meaning is formalized as a *word embedding*—a point $\boldsymbol{w}_i \in \mathbb{R}^d$—a small coordinate in a high-dimensional semantic manifold. The embedding compresses the uncertainty of linguistic form into the geometry of meaning. If two words, such as *big* and *large*, fall into the same region of this space, they share an information channel; their redundancy mirrors synonymy.

Imagine binning the embedding space into cells $\{\mathcal{E}_1, \ldots, \mathcal{E}_K\}$, as if discretizing a continuous signal. Each cell collects all words whose vectors fall within it, and therefore represents a coarse semantic class. If we define

$$p(B = k) = \frac{1}{|V|} \sum_{i=1}^{|V|} \mathbf{1}_{\boldsymbol{w}_i \in \mathcal{E}_k}, \tag{1}$$

then the entropy of the binned representation,

$$H(B) = -\sum_{k=1}^{K} p(B = k)\, \log p(B = k), \tag{2}$$

measures how widely dispersed meanings are across the semantic manifold. Low entropy indicates tight clustering—a language rich in redundancy; high entropy signals a lexicon stretched across conceptual space.

From an information-theoretic perspective, we can regard the embedding function

$$f : V \to \mathbb{R}^d, \qquad f(w_i) = \boldsymbol{w}_i, \tag{3}$$

as an *encoder* in a semantic channel:

$$w_i \longrightarrow \boldsymbol{w}_i \longrightarrow \hat{w}_i.$$

Here the discrete lexical source W emits symbols w_i with probability $p(w_i)$, the encoder maps them to continuous signals $\boldsymbol{w}_i$, and a decoder attempts to reconstruct the intended concept $\hat{w}_i$. The process carries mutual information

$$I(W; \boldsymbol{W}) = H(W) - H(W \mid \boldsymbol{W}), \tag{4}$$

quantifying how much of the lexical uncertainty is preserved in the embedding geometry. When two words have nearly identical embeddings, the channel becomes degenerate: the same semantic code corresponds to multiple lexical sources, and synonymy appears as a natural form of compression.

In this view, the embedding space is not merely a storage of vectors but a semantic communication medium. Each word is a transmitter, each embedding a signal, and each synonym a sign of the system's efficiency— a reminder that language, like any channel, trades entropy for meaning.

TODO:

Relation to Carnap–Bar-Hillel

- Unit: continuous word embeddings (vectors) vs logical sentences/propositions over state-descriptions.
- Measure: Shannon entropy/mutual information in a communication channel vs logical content $\mathrm{I}(\varphi) = -\log P(\varphi)$.
- Probabilities: corpus- and task-induced empirical distributions vs a priori logical probability.
- Structure: geometric similarity and redundancy in vector space vs set-theoretic entailment and inclusion.
- Outcomes: channel efficiency and compression (synonymy as redundancy) vs semantic content and confirmation; no analogue of the Bar-Hillel–Carnap paradox (tautology=0, contradiction=max).

Describe shortly