Some Remarks on Natural Language Processing

Nils Kujath

v2026.04

Degenerate Geometry of One-Hot Word Representations

- Let $V = \{w_1, \ldots, w_{|V|}\}$ be a finite vocabulary. The canonical (standard) basis of $\mathbb{R}^{|V|}$ is the indexed set $\mathcal{B}_{\text{can}} = \{\mathbf{e}_1, \ldots, \mathbf{e}_{|V|}\}$, where $\mathbf{e}_i$ denotes the i-th basis vector, i.e. the unique (one-hot) vector with a 1 in coordinate i and zeros elsewhere. The BoW construction $\boxed{\iota : V \to \mathbb{R}^{|V|}, \quad w_i \mapsto \mathbf{e}_i}$ identifies each word $w_i \in V$ with the canonical basis vector $\mathbf{e}_i \in \mathbb{R}^{|V|}$.
- Under the standard inner product on $\mathbb{R}^{|V|}$, all distinct word types are mutually orthogonal:

  $$\boxed{\mathbf{e}_u{}^\top \mathbf{e}_v = \delta_{uv} = \begin{cases} 1 & \text{if } u = v, \\ 0 & \text{if } u \neq v \end{cases}, \quad \forall u, v \in \{1, \ldots, |V|\}}.$$

  Fmr., the metric restricted to $\iota(V)$ is trivial; all word types are equidistant in $\mathbb{R}^{|V|}$: $\forall u, v \in \{1, \ldots, |V|\}$,

  $$\boxed{d\big|_{\iota(V)\times\iota(V)}(\mathbf{e}_u, \mathbf{e}_v) = \|\mathbf{e}_u - \mathbf{e}_v\| = \begin{cases} 0 & \text{if } u = v, \\ \sqrt{\underbrace{(1-0)^2}_{\text{pos. } u} + \underbrace{(0-1)^2}_{\text{pos. } v} + \underbrace{(0-0)^2 + \cdots + (0-0)^2}_{|V|-2 \text{ pos.}}} = \sqrt{2} & \text{if } u \neq v \end{cases}}.$$

- Consequently, the geometry induced by ι is degenerate: all distinct words are equally dissimilar, and no notion of graded semantic proximity can be expressed under the BoW representation.

From the Distributional Hypothesis to Word Embeddings

- Let $V = \{w_1, \ldots, w_{|V|}\}$ be a finite vocabulary. Let $\mathcal{D} = (t_1, t_2, \ldots, t_N)$ be a corpus of N tokens from V. Fix a context window size $k \in \mathbb{N}^+$. Define the context map $\mathcal{C}_k : \{n \in \mathbb{N} : k < n \leq N - k\} \to V^{2k}$ by:

$$\boxed{\mathcal{C}_k(n) = (t_{n-k}, \ldots, t_{n-1}, t_{n+1}, \ldots, t_{n+k})}.$$

  Note that positions $n \leq k$ and $n > N - k$ are excluded since k tokens of context are required on each side.

- For each $w_i \in V$, define the distributional profile of w_i in $\mathcal{D}$ as the multiset:

$$\boxed{\Delta_k(w_i) = \{\!\!\{\mathcal{C}_k(n) : n \in \{k{+}1, \ldots, N{-}k\},\ t_n = w_i\}\!\!\}}.$$

- The Distributional Hypothesis (see esp. Harris 1954 and Firth 1957) asserts that w_i and w_j are semantically similar if they appear in a similar contexts, that is, if $\Delta_k(w_i) \approx \Delta_k(w_j)$. The previous slide has shown that the degenerate geometry of BoW representations precludes any graded notion of similarity. However, comparing multisets over V^{2k} directly also seems intractable. The goal is therefore to find a map:

$$\boxed{\phi : V \to \mathbb{R}^m \; (m \ll |V|) \quad \text{s.t.} \quad \Delta_k(w_i) \approx \Delta_k(w_j) \quad \text{is operationalised as} \quad \phi(w_i) \approx \phi(w_j) \text{ in } \mathbb{R}^m}.$$

  That is, ϕ embeds the discrete set V into (the so-called embedding space) $\mathbb{R}^m$ such that distributional similarity in $\mathcal{D}$ is faithfully compressed into geometric proximity. (Note: We will discuss later why we desire $m \ll |V|$.)

Count-Based Word Embeddings

- Let $V = \{w_1, \ldots, w_{|V|}\}$ be a finite vocabulary and $k \in \mathbb{N}^+$ the selected size of the context window. We could define a co-occurrence matrix $M \in \mathbb{N}^{|V|\times|V|}$ where M_{ij} is the number of times w_j appears in a context window of size k around w_i in the corpus $\mathcal{D} = (t_1, \ldots, t_N)$:

  $$\boxed{M_{ij} = \sum_{n=k+1}^{N-k} \underbrace{\mathbf{1}[t_n = w_i]}_{1 \text{ if center is } w_i} \cdot \sum_{\substack{l=n-k \\ l\neq n}}^{n+k} \underbrace{\mathbf{1}[t_l = w_j]}_{1 \text{ if context slot is } w_j}}\,.$$

  The outer sum ranges over all valid center positions $n \in \{k{+}1, \ldots, N{-}k\}$; the inner sum scans the $2k$ surrounding context slots.
- Recall the context map $\mathcal{C}_k(n) = (t_{n-k}, \ldots, t_{n-1}, t_{n+1}, \ldots, t_{n+k})$ for $n \in \{k{+}1, \ldots, N{-}k\}$, and the distributional profile $\Delta_k(w_i) = \{\!\!\{\mathcal{C}_k(n) : n \in \{k{+}1, \ldots, N{-}k\},\ t_n = w_i\}\!\!\}$ from the previous slide. The co-occurrence matrix M is a lossy compression of the distributional profiles Δ_k over D: ordering within each context tuple is discarded, and only co-occurrence frequencies are retained.
- In M, each row $\mathbf{m}_i = (M_{i1}, \ldots, M_{i,|V|}) \in \mathbb{R}^{|V|}$ is already a representation of w_i that reflects distributional similarity: words with similar co-occurrence patterns have similar row vectors. However, these rows live in $\mathbb{R}^{|V|}$, not the $\mathbb{R}^m$ with $m \ll |V|$ sought on the previous slide.

References

Firth, John R. 1957. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, 1–32. Oxford: Basil Blackwell.

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2–3). 146–162.