3

CS197 Assignment 4: Progress Report

Project Member 1
Stanford Email
Project Member 2
Stanford Email
Project Member 3
Stanford Email

November 1, 2025

1 This week

1.1 Vector

1.2 Plan

1.3 Results

"'latex [11pt]article amsmath, amssymb, amsthm, mathtools bm geometry margin=1in
GloVe-Style Token–Feature Factorization for Morphene-Constrained Embeddings

2 GloVe-Style Factorization

2.1 Notation

- Token vocabulary: $\mathscr{T}$, $|\mathscr{T}| = T$.
- Feature (context) inventory: $\mathscr{F}$, $|\mathscr{F}| = F$.
- Token embedding matrix $U \in \mathbb{R}^{T \times k}$; u_t is row t.
- Feature embedding matrix $S \in \mathbb{R}^{F \times k}$; s_f is row f.
- Context statistics: PPMI_{tf} is the positive PMI between token t and feature f.
- Local morphological classes: $c \in \mathscr{C}$; a class c indexes a set of tokens $T_c \subseteq \mathscr{T}$.
- Morphological Laplacian: $L^{(c)} \in \mathbb{R}^{T \times T}$ supported only inside $T_c \times T_c$.

In contrast to fitting $\mathrm{PPMI} \approx VV^\top$ (token–token reconstruction), we directly factorize the *token–feature* matrix.

2.2 Objective

Let

$$X \equiv \text{PPMI} \in \mathbb{R}^{T\times F}.$$

A weighted GloVe-style objective is:

$$\min_{U,S} \sum_{(t,f)\in \text{nz}(X)} w_{tf} \left(u_t^\top s_f - X_{tf}\right)^2 + \lambda_{\text{morph}} \sum_{c\in\mathscr{C}} tr\big(U^\top L^{(c)} U\big) + \gamma\big(\|U\|_F^2 + \|S\|_F^2\big), \quad (1)$$

where w_{tf} is typically a monotone function of co-occurrence count used to smoothly discount extremely rare or extremely frequent pairs.
The first term directly aligns the embedding inner-product with PPMI. The morphology term keeps tokens in the same morphological class c close in embedding space. Regularization controls rank- and scale-pathologies.

2.3 Gradients

The gradients take the form $\nabla_{u_t} L = \sum_{f\in \text{nz}(X_{t\cdot})} 2\,w_{tf}\left(u_t^\top s_f - X_{tf}\right)s_f + 2\lambda_{\text{morph}} \sum_{c:\,t\in T_c} \left(L^{(c)}U\right)_t + 2\gamma u_t$,
$\nabla_{s_f} L = \sum_{t\in \text{nz}(X_{\cdot f})} 2\,w_{tf}\left(u_t^\top s_f - X_{tf}\right)u_t + 2\gamma s_f$.

2.4 Preservation of the Intended Structure

Morphological Intention. Cross-lingual *morphological* equivalence is enforced by the Laplacian term

$$\lambda_{\text{morph}} \sum_{c\in\mathscr{C}} tr(U^\top L^{(c)} U) = \lambda_{\text{morph}} \sum_{c} \sum_{(i,j)\in \text{edges}(c)} \|u_i - u_j\|_2^2.$$

Relationships among tokens in T_c are directly encoded; this includes cross-lingual effects when T_c contains members from more than one language. This *remains exactly the same* in the GloVe-style factorization.

Semantic Intention. Instead of reconstructing a *token–token* matrix $VV^\top$, the objective eq:glove recovers token relationships through the token–feature interactions $u_t^\top s_f$. Tokens are close when they share subword/context features (character n-grams, morph affixes). Thus we preserve the same semantic signals that originally lived in PPMI.

2.5 Computational Advantages

The matrix $X = \text{PPMI}$ is $T\times F$, typically sparse. We never materialize *any* $T\times T$ object. Instead, training cost is proportional to nonzeros of X:

$$cost = O\big(\text{nnz}(X)\,k\big) \quad per pass.$$

Memory is $O\big((T+F)k + \text{nnz}(X)\big)$.
By contrast, token–token reconstruction implicitly depends on $VV^\top$ (or $G = XX^\top$), requiring $O(T^2)$ space/time. Eliminating $VV^\top$ thus removes the quadratic blowup while keeping identical morphological constraints.

3 Mini-Batching

3.1 Token–Feature Sampling

Instead of summing over all $(t, f) \in \mathrm{nz}(X)$, sample a mini-batch $\mathscr{B}$:

$$\mathscr{B} \subset \mathrm{nz}(X), \qquad |\mathscr{B}| = B.$$

Then we optimize a stochastic objective

$$L_{\mathscr{B}}(U,S) = \sum_{(t,f)\in\mathscr{B}} w_{tf}\left(u_t^\top s_f - X_{tf}\right)^2 + \lambda_{\mathrm{morph}} \sum_c tr(U^\top L^{(c)} U) + \gamma(\|U\|_F^2 + \|S\|_F^2).$$

Since $L^{(c)}$ only touches tokens, we include its contribution at every gradient step (or sample edges within T_c).

Stochastic Update. For $(t, f) \in \mathscr{B}$, gradient updates:

$$u_t \leftarrow u_t - \eta \, \nabla_{u_t} L_{\mathscr{B}}, \qquad s_f \leftarrow s_f - \eta \, \nabla_{s_f} L_{\mathscr{B}},$$

with SGD/Adam; η is learning rate.

3.2 Mini-Batching via Edge Sampling in $L^{(c)}$

Each $L^{(c)}$ is sparse over $T_c \times T_c$. With a graph $G_c = (T_c, E_c)$,

$$tr(U^\top L^{(c)} U) = \sum_{(i,j)\in E_c} \|u_i - u_j\|_2^2.$$

We can sample edges $(i, j) \in E_c$ to approximate the Laplacian term:

$$\sum_c tr(U^\top L^{(c)} U) \approx \frac{|E_c|}{|\mathscr{E}_c|} \sum_{(i,j)\in\mathscr{E}_c} \|u_i - u_j\|_2^2,$$

where $\mathscr{E}_c$ is the mini-batch of edges from class c.

3.3 Benefits

- **Compute:** Each minibatch is $O(Bk)$, removing $O(T^2)$ dependence.
- **Memory:** No $T \times T$ matrices are formed.
- **Faithfulness:** Because $L^{(c)}$ is still applied to token embeddings U, cross-lingual morphological structure is preserved.
- **Low-resource focus:** We may oversample rare-lingual T_c or low-frequency (t, f) pairs to protect minority signal.

4 Conclusion

The GloVe-style factorization preserves the intent of the original formulation:

- Morphological structure is preserved through $L^{(c)}$ acting on token embeddings.
- Semantic similarity arises via token–feature PPMI interactions.
- Computational overhead is drastically reduced: no $T \times T$ objects; costs scale with $\mathrm{nnz}(X)$.
- Mini-batching follows naturally over token–feature pairs and optionally over morphological edges.

Thus the method remains faithful to its morphene-constrained objective while addressing scaling limits in both compute and memory.