1 Mechanical Turk Questions

Figures 1-3 show the wording and format of the questions as presented to mechanical turk users.

![Mechanical Turk question screenshot](image)

Figure 1: Screen shot for the Mechanical Turk question for determining if mis-ranked phrases are good approximations of the true phrase.
Figure 2: Screen shot for the Mechanical Turk question used to determine if NNSE/CNNSE/SVD dimensions are interpretable and coherent.

Figure 3: Screen shot for the Mechanical Turk question used to determine if NNSE or CNNSE phrasal representations are consistent.
<table>
<thead>
<tr>
<th>Adjective</th>
<th>Noun</th>
<th>Phrase</th>
<th>Estimated Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>negative aspects</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>negative</td>
<td>aspects</td>
<td>negative aspects (observed)</td>
<td>negative aspects (estimated)</td>
</tr>
<tr>
<td>intruders, intrusions, overflows</td>
<td>facets, topics, different aspects</td>
<td>consequences, environmental consequences, serious consequences</td>
<td>facets, topics, different aspects</td>
</tr>
<tr>
<td>consequences, environmental consequences, serious consequences</td>
<td>underpinnings, arousal, implications</td>
<td>features, oddities, standard features</td>
<td>underpinnings, arousal, implications</td>
</tr>
<tr>
<td>instinctive, conditioned, oscillatory</td>
<td>features, oddities, standard features</td>
<td>intruders, intrusions, overflows</td>
<td>intruders, intrusions, overflows</td>
</tr>
<tr>
<td>indecent, unlawful, obscene</td>
<td>workings, truths, essence</td>
<td>facets, topics, different aspects</td>
<td>consequences, environmental consequences, serious consequences</td>
</tr>
<tr>
<td><strong>military aid</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>military</td>
<td>aid</td>
<td>military aid (observed)</td>
<td>military aid (estimated)</td>
</tr>
<tr>
<td>servicemen, commandos, military intelligence</td>
<td>guidance, advice, assistance</td>
<td>servicemen, commandos, military intelligence</td>
<td>guidance, advice, assistance</td>
</tr>
<tr>
<td>guerrilla paramilitary, anti-terrorist</td>
<td>mentoring, tutoring, internships</td>
<td>guidance, advice, assistance</td>
<td>servicemen, commandos, military intelligence</td>
</tr>
<tr>
<td>conglomerate, giants, conglomerates</td>
<td>award, awards, honors</td>
<td>compliments, congratulations, replies</td>
<td>mentoring, tutoring, internships</td>
</tr>
<tr>
<td>managerial, logistical, governmental</td>
<td>certificates, degrees, bachelor</td>
<td>training, appropriate training, advanced training</td>
<td>award, awards, honors</td>
</tr>
<tr>
<td>mankind, Palestinian people, Iraqi people</td>
<td>servicemen, commandos, military intelligence</td>
<td>conglomerate, giants, conglomerates</td>
<td>conglomerate, giants, conglomerates</td>
</tr>
<tr>
<td><strong>bad behavior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bad</td>
<td>behavior</td>
<td>bad behavior (observed)</td>
<td>bad behavior (estimated)</td>
</tr>
<tr>
<td>Great place, place, fantastic place</td>
<td>scholastic achievement, ethical behavior, behaviors</td>
<td>scholastic achievement, ethical behavior, behaviors</td>
<td>scholastic achievement, ethical behavior, behaviors</td>
</tr>
<tr>
<td>antithesis, affront, omen</td>
<td>dating, intimacy, courtship</td>
<td>intruders, intrusions, overflows</td>
<td>dating, intimacy, courtship</td>
</tr>
<tr>
<td>thankful, grateful, sorry</td>
<td>morphology, phylogeny, physiology</td>
<td>inconsistences, faults, flaws</td>
<td>morphology, phylogeny, physiology</td>
</tr>
<tr>
<td>goofy, crazy, fucking</td>
<td>psychosis, depression, disorder</td>
<td>comm, wildness, haunting</td>
<td>psychosis, depression, disorder</td>
</tr>
<tr>
<td>go-ahead, spanking, shift</td>
<td>invited, attitudes, encouraged</td>
<td>pasts, non-commercial use, mind-set</td>
<td>invited, attitudes, encouraged</td>
</tr>
</tbody>
</table>
2 CNNSE Algorithm

Recall that NNSE seeks a lower dimensional sparse representation for $w$ words using the $c$-dimensional corpus statistics in a matrix $X \in \mathbb{R}^{w \times c}$. NNSE minimizes the following objective function:

$$\text{argmin}_{A,D} \frac{1}{2} \sum_{i=1}^{w} \|X_{i,:} - A_{i,:} \times D\|^2 + \lambda_1 \|A\|_1$$

subject to:

1. $D_{i,:}D_{i,:}^T \leq 1, \forall 1 \leq i \leq \ell$
2. $A_{i,j} \geq 0, 1 \leq i \leq w, 1 \leq j \leq \ell$

where $A_{i,j}$ indicates the entry at the $i$th row and $j$th column of matrix $A$, and $A_{i,:}$ indicates the $i$th row of the matrix. The solution includes a matrix $A \in \mathbb{R}^{w \times \ell}$ that is sparse, non-negative, and represents word semantics in an $\ell$-dimensional latent space. $D \in \mathbb{R}^{\ell \times c}$ is the encoding of corpus statistics in the latent space. The $L_1$ constraint encourages sparsity in $A$; $\lambda_1$ is a hyperparameter. Equation 2 constrains $D$ to eliminate solutions where the norm of $A$ is made arbitrarily small by making the norm of $D$ arbitrarily large. Equation 3 ensures that $A$ is non-negative. Together, $A$ and $D$ factor the original corpus statistics matrix $X$ in a way that minimizes reconstruction error while respecting sparsity and non-negativity constraints.

Consider a phrase $p$ made up of words $i$ and $j$. In the most general setting, the following composition constraint could be applied to the rows of matrix $A$ from Equation 1 corresponding to $p$, $i$ and $j$:

$$A(p,:) = f(A(i,:),A(j,:))$$

where $f$ is some composition function. The composition function constrains the space of learned latent representations $A \in \mathbb{R}^{w \times \ell}$ to be those solutions that are compatible with the composition function defined by $f$. Incorporating $f$ into Equation 1 we have:

$$\text{argmin}_{A,D,\Omega} \sum_{i=1}^{w} \frac{1}{2} \|X_{i,:} - A_{i,:} \times D\|^2 + \lambda_1 \|A\|_1 + \lambda_c \sum_{\text{phrase } p, \ p = (i,j)} (A(p,:) - f(A(i,:),A(j,:)))^2$$

Where each phrase $p$ is comprised of words $(i,j)$ and $\Omega$ represents all parameters of $f$ that may need to be optimized. We have added a squared loss term for the composition function, and a new regularization parameter $\lambda_c$ to weight the importance of respecting composition. We call this new formulation Compositional Non-Negative Sparse Embeddings (CNNSE).

In this work, we choose $f$ to be weighted addition because it has been shown to work well for adjective noun and noun noun composition [Mitchell and Lapata, 2010; Dinu et al., 2013], and because it leads to a formulation that lends itself well to optimization. Weighted addition is:

$$f(A(i,:),A(j,:)) = \alpha A(i,:) + \beta A(j,:).$$

This choice of $f$ requires that we simultaneously optimize for $A,D,\alpha$ and $\beta$.

We can further simplify the loss function by constructing a matrix $B$ that imposes the composition by addition constraint. $B$ is constructed so that for each phrase $p = (i,j)$:

$B_{(p,p)} = 1$, $B_{(p,i)} = -\alpha$, and $B_{(p,j)} = -\beta$. For our models, we use $\alpha = \beta = 0.5$, which serves to average the single word representations. The matrix $B$ allows us to reformulate the loss function from Eq 5:

$$\text{argmin}_{A,D} \frac{1}{2} \|X - AD\|_F^2 + \lambda_1 \|A\|_1 + \frac{1}{2} \lambda_c \|BA\|_F^2$$

This choice of $f$ requires that we simultaneously optimize for $A,D,\alpha$ and $\beta$.
Algorithm 1 CNNSE

**Input:** $X, B, \lambda_1, \lambda_c$

Randomly initialize $A, D$

prevL $\leftarrow 0$

curL $\leftarrow \frac{1}{2} \| X - AD \|_F^2 + \lambda_1 \| A \|_1 + \frac{1}{2} \lambda_c \| BA \|_F^2$

while (prevL - curL) $\leq$ prevL*10^{-3} do

$A \leftarrow \text{ADMM}(D, X, B, \lambda_1, \lambda_c)$

$D \leftarrow \text{gradientDescent}(D, X, A)$

prevL $\leftarrow$ curL

curL $\leftarrow \frac{1}{2} \| X - AD \|_F^2 + \lambda_1 \| A \|_1 + \frac{1}{2} \lambda_c \| BA \|_F^2$

end while

return $A, D$

where $F$ indicates the Frobenius norm. $B$ acts as a selector matrix, subtracting from the latent representation of the phrase the average latent representation of the phrase’s constituent words.

We now have a loss function that is the sum of several convex functions of $A$: squared loss, $L_1$ regularization and the composition constraint. This sum of sub-functions is the format required for the alternating directions method of multipliers (ADMM) (Boyd 2010). ADMM substitutes a dummy variable $z$ for $A$ in the sub-functions:

$$\arg\min_{A,D} \frac{1}{2} \| X - AD \|_F^2 + \lambda_1 \| A \|_1 + \frac{1}{2} \lambda_c \| BA \|_F^2$$

st: $A = z_1$

$A = z_c$

$D_{i,i}D_{i,i}^T \leq 1, \forall 1 \leq i \leq \ell$

$A_{i,j} \geq 0, 1 \leq i \leq w, 1 \leq j \leq \ell$

Equations 9 and 10 ensure that the dummy variables match $A$; ADMM uses an augmented Lagrangian to incorporate and relax these new constraints. The augmented Lagrangian for the above optimization problem above is:

$$L_\rho(A, z_1, z_c, u_1, u_c) = \frac{1}{2} \| X - AD \|_F^2 + \lambda_1 \| A \|_1 + \frac{1}{2} \lambda_c \| BA \|_F^2 +$$

$u_1(A - z_1) + u_c(A - z_c) + \frac{\rho}{2} (\| A - z_1 \|_2^2 + \| A - z_c \|_2^2)$

We optimize for $A, z_1$ and $z_c$ separately, and then update the dual variables (see Algorithm 2 for solutions and updates). ADMM has nice convergence properties for convex functions, as we have when solving for $A$. Code for ADMM is available online. ADMM is used when solving for $A$ in the Online Dictionary Learning algorithm, solving for $D$ remains unchanged from the NNSE implementation (see Algorithm 1).

Algorithm 2 ADMM solution for augmented Lagrangian in equation 13

Input: $D, X, B, \lambda_1, \lambda_c$

{Lagrangian parameter}
\[ \rho \leftarrow 1 \]

{Dummy Variables}
\[ z_1 \leftarrow 0_{w,\ell} \]
\[ z_c \leftarrow 0_{w,\ell} \]

{Dual Variables}
\[ u_1 \leftarrow 0_{w,\ell} \]
\[ u_c \leftarrow 0_{w,\ell} \]
\[ dti \leftarrow DD^T + 2 \ast \rho \ast I_m \]

while not converged do

\[ A \leftarrow (XD^T + \rho(z_1 + z_c) - (u_1 + u_c)) / dti \]
\[ z_c \leftarrow (\rho \ast A + u_c) / (\lambda_c \ast (B' \ast B) + \rho \ast I_w) \]
\[ \gamma \leftarrow A + u_1 / \rho \]
\[ \kappa \leftarrow \lambda_1 / \rho \]

{Soft Threshold Operator for $L_1$ constraint} \{(a)_+ \ is \ shorthand \ for \ max(0,a)\}
\[ z_1 = (\gamma - \kappa)_+ - (-\gamma - \kappa)_+ \]

{Update Dual Variables}
\[ u_1 = u_1 + \rho \ast (A - z_1) \]
\[ u_c = u_c + \rho \ast (A - z_c) \]

end while

return $A$

References

