A Compositional and Interpretable Semantic Space: Supplementary Material

Alona Fyshe,[1] Leila Wehbe,[1] Partha Talukdar,[2] Brian Murphy,[3] and Tom Mitchell[1]

[1] Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA

[2] Indian Institute of Science, Bangalore, India

[3] Queen's University Belfast, Belfast, Northern Ireland

afyshe@cs.cmu.edu, lwehbe@cs.cmu.edu, ppt@serc.iisc.in, brian.murphy@qub.ac.uk, tom.mitchell@cs.cmu.edu

1 Mechanical Turk Questions

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

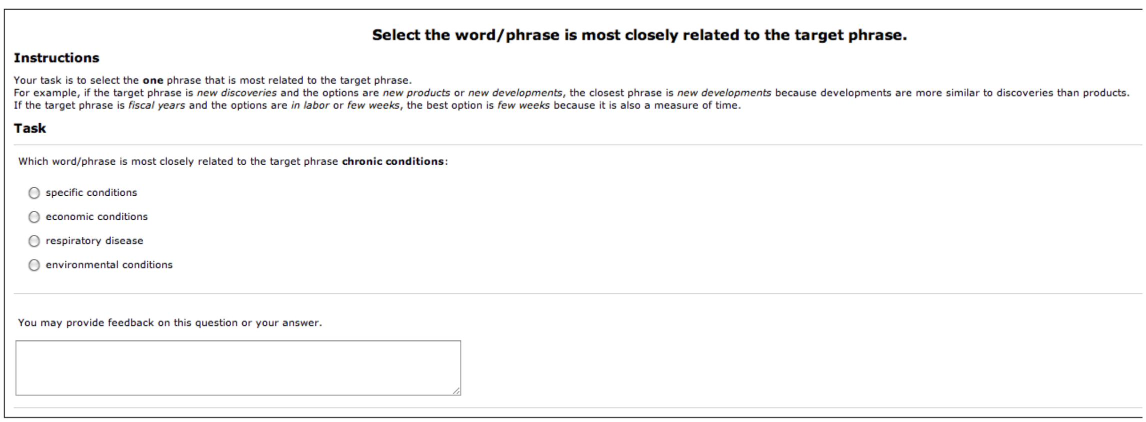
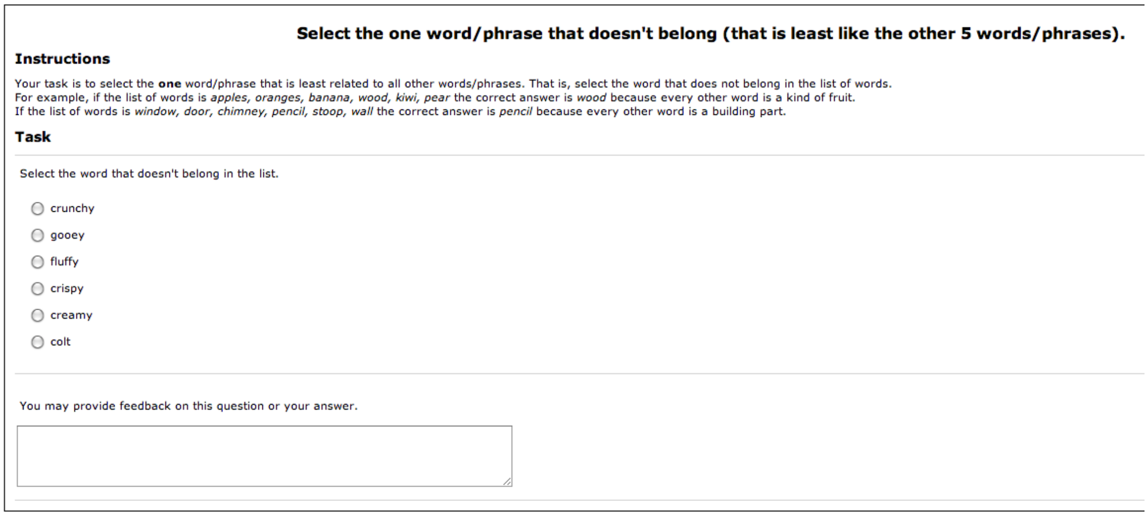


Figure 1: Screen shot for the Mechanical Turk question for determining if mis-ranked phrases are good approximations of the true phrase.



Select the one word/phrase that doesn't belong (that is least like the other 5 words/phrases).

Instructions

Your task is to select the **one** word/phrase that is least related to all other words/phrases. That is, select the word that does not belong in the list of words.
For example, if the list of words is *apples, oranges, banana, wood, kiwi, pear* the correct answer is *wood* because every other word is a kind of fruit.
If the list of words is *window, door, chimney, pencil, stoop, wall* the correct answer is *pencil* because every other word is a building part.

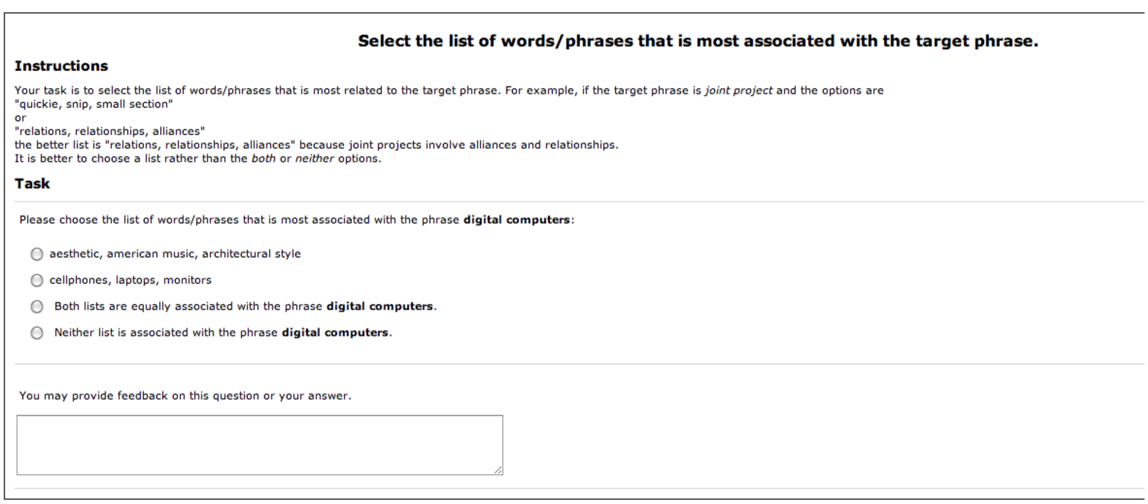
Task

Select the word that doesn't belong in the list.

- crunchy
- gooey
- fluffy
- crispy
- creamy
- colt

You may provide feedback on this question or your answer.

Figure 2: Screen shot for the Mechanical Turk question used to determine if NNSE/CNNSE/SVD dimensions are interpretable and coherent.



Select the list of words/phrases that is most associated with the target phrase.

Instructions

Your task is to select the list of words/phrases that is most related to the target phrase. For example, if the target phrase is *joint project* and the options are
"quickie, snip, small section"
or
"relations, relationships, alliances"
the better list is "relations, relationships, alliances" because joint projects involve alliances and relationships.
It is better to choose a list rather than the *both* or *neither* options.

Task

Please choose the list of words/phrases that is most associated with the phrase **digital computers**:

- aesthetic, american music, architectural style
- cellphones, laptops, monitors
- Both lists are equally associated with the phrase **digital computers**.
- Neither list is associated with the phrase **digital computers**.

You may provide feedback on this question or your answer.

Figure 3: Screen shot for the Mechanical Turk question used to determine if NNSE or CNNSE phrasal representations are consistent.

Table 1: A qualitative evaluation of CNNSE interpretable dimensions for several phrases and their constituent words. For each word or phrase the top 5 scoring dimensions are selected. Then, for each selected dimension the interpretable summarization is given, which reports the top scoring words in that dimension.

Adjective	Noun	Phrase	Estimated Phrase
negative aspects			
negative	aspects	negative aspects (observed)	negative aspects (estimated)
intruders, intrusions, overflows	facets, topics, different aspects	consequences, environmental consequences, serious consequences	facets, topics, different aspects
consequences, environmental consequences, serious consequences	underpinnings, arousal, implications	features, oddities, standard features	underpinnings, arousal, implications
instinctive, conditioned, oscillatory	features, oddities, standard features	intruders, intrusions, overflows	intruders, intrusions, overflows
indecent, unlawful, obscene	workings, truths, essence	facets, topics, different aspects	consequences, environmental consequences, serious consequences
postmodern, preconceived, psychoanalytic	key factors, key elements, main factors	contingencies, specific items, specific terms	features, oddities, standard features
military aid			
military	aid	military aid (observed)	military aid (estimated)
servicemen, commandos, military intelligence	guidance, advice, assistance	servicemen, commandos, military intelligence	guidance, advice, assistance
guerrilla paramilitary, anti-terrorist	mentoring, tutoring, internships	guidance, advice, assistance	servicemen, commandos, military intelligence
conglomerate, giants, conglomerates	award, awards, honors	compliments, congratulations, replies	mentoring, tutoring, internships
managerial, logistical, governmental	certificates, degrees, bachelor	training, appropriate training, advanced training	award, awards, honors
humankind, Palestinian people, Iraqi people	servicemen, commandos, military intelligence	conglomerate, giants, conglomerates	conglomerate, giants, conglomerates
bad behavior			
bad	behavior	bad behavior (observed)	bad behavior (estimated)
Great place, place, fantastic place	scholastic achievement, ethical behavior, behaviors	scholastic achievement, ethical behavior, behaviors	scholastic achievement, ethical behavior, behaviors
antithesis, affront, omen	dating, intimacy, courtship	intruders, intrusions, overflows	dating, intimacy, courtship
thankful, grateful, sorry	morphology, phylogeny, physiology	inconsistencies, faults, flaws	morphology, phylogeny, physiology
goofy, crazy, fucking	psychosis, depression, disorder	comm, wildness, haunting	psychosis, depression, disorder
go-ahead, spanking, shrift	invited, attitudes, encouraged	pasts, non-commercial use, mind-set	invited, attitudes, encouraged

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:

$$\underset{A,D}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{w} \left\| X_{i,:} - A_{i,:} \times D \right\|^2 + \lambda_1 \left\| A \right\|_1 \tag{1}$$

$$\text{st: } D_{i,:} D_{i,:}^T \leq 1, \forall\ 1 \leq i \leq \ell \tag{2}$$

$$A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell \tag{3}$$

where $A_{i,j}$ indicates the entry at the ith row and jth column of matrix A, and $A_{i,:}$ indicates the ith 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 ℓ-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; λ_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:

$$A_{(p,:)} = f(A_{(i,:)}, A_{(j,:)}) \tag{4}$$

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:

$$\begin{aligned} \underset{A,D,\Omega}{\operatorname{argmin}} \sum_{i=1}^{w} \frac{1}{2} \left\| X_{i,:} - A_{i,:} \times D \right\|^2 + \lambda_1 \left\| A \right\|_1 + \\ \lambda_c \sum_{\substack{\text{phrase } p, \\ p=(i,j)}} \left(A_{(p,:)} - f(A_{(i,:)}, A_{(j,:)}) \right)^2 \end{aligned} \tag{5}$$

Where each phrase p is comprised of words (i, j) and Ω 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 λ_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 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:

$$f(A_{(i,:)}, A_{(j,:)}) = \alpha A_{(i,:)} + \beta A_{(j,:)} \tag{6}$$

This choice of f requires that we simultaneously optimize for A, D, α and β.

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:

$$\underset{A,D}{\operatorname{argmin}} \frac{1}{2} \left\| X - AD \right\|_F^2 + \lambda_1 \left\| A \right\|_1 + \frac{1}{2} \lambda_c \left\| BA \right\|_F^2 \tag{7}$$

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 - currL) $\leq$ prevL$*10^{-3}$ **do**
  $A \Leftarrow$ ADMM$(D, X, B, \lambda_1, \lambda_c)$
  $D \Leftarrow$ 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:

$$\underset{A,D}{\operatorname{argmin}} \frac{1}{2}\|X - AD\|_F^2 + \lambda_1\|z_1\|_1 + \frac{1}{2}\lambda_c\|Bz_c\|_F^2 \tag{8}$$

$$\text{st: } A = z_1 \tag{9}$$

$$A = z_c \tag{10}$$

$$D_{i,:}D_{i,:}^T \leq 1, \forall\ 1 \leq i \leq \ell \tag{11}$$

$$A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell \tag{12}$$

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:

$$L_\rho(A, z_1, z_c, u_1, u_c) = \frac{1}{2}\|X - AD\|_F^2 + \lambda_1\|z_1\|_1 + \frac{1}{2}\lambda_c\|Bz_c\|_F^2 + \\ u_1(A - z_1) + u_c(A - z_c) + \frac{\rho}{2}(\|A - z_1\|_2^2 + \|A - z_c\|_2^2) \tag{13}$$

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[1]. 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).

[1] `http://www.stanford.edu/~boyd/papers/admm/`

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}
$u_1 \Leftarrow 0_{w,\ell}$
$u_c \Leftarrow 0_{w,\ell}$
$dti \Leftarrow DD^T + 2 * \rho * I_m$
while not converged **do**
  $A \Leftarrow \left(XD^T + \rho(z_1 + z_c) - (u_1 + u_c)\right)/dti$
  $z_c \Leftarrow (\rho * A + u_c)/(\lambda_c * (B' * B) + \rho * I_w)$
  $\gamma \Leftarrow A + u_1/\rho$
  $\kappa \Leftarrow \lambda_1/\rho$
  {Soft Threshold Operator for L_1 constraint} {$(a)_+$ is shorthand for max(0,a)}
  $z_1 = (\gamma - \kappa)_+ - (-\gamma - \kappa)_+$
  {Update Dual Variables}
  $u_1 = u_1 + \rho * (A - z_1)$
  $u_c = u_c + \rho * (A - z_c)$
end while
return A

References

Stephen Boyd. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1): 1–122, 2010. ISSN 1935-8237. doi: 10.1561/2200000016.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. General estimation and evaluation of compositional distributional semantic models. In *Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, 2013.

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–429, November 2010. ISSN 1551-6709. doi: 10.1111/j.1551-6709.2010.01106.x.