Supplementary Material for Improving Word Meaning using Text, Feature Norms, and Images Together

September 16, 2013

Feature Norm Cleaning and Normalization

Subjects who met one of the following criteria were removed:

- Responses from known spammers were removed. Subjects were removed when they failed to provide the correct number of responses or ignored directions.
- Subjects who copied paragraphs from Wikipedia, Google, or our instructions were removed.

Collected feature norms were automatically cleaned using the following process:

1. Modal verbs (e.g., müssen, sollen, etc), prepositions, and forms of *sein* and *haben* were all corrected to all lowercase spellings.
2. Words which were spelled without umlauts or Eszett were normalized to the umlaut/Eszett spelling if another subject gave the exact same response with umlauts or eszetts.
3. We used a list of spelling corrections from a prior collection.
4. If two subjects gave the same response with and without the verbs haben or sein, such as "ist rot" and "rot", then we normalized to the form with the verb.

After these automatic procedures, we completed one pass where we manually corrected any additional spelling mistakes and typos. Finally, we semantically normalized the feature norms.

When semantically normalizing, we attempted to reduce the number of unique responses given by the subjects. For example, users may have given "ist aus Plastik" and "ist meist auf Plastik", and we normalized to the former response. Other examples include "ist grün oder weiß", which we normalized to two feature norms: "ist grün" and "ist weiß".

The normalizations we performed were somewhat adhoc and based on our own intuitions. In the interest of transparency, we release two versions of our data set: before and after semantic normalizations. A list of transformations we applied in normalization is also provided.

Inference Algorithm for 3D mLDA

Algorithm 1 lists our online variational inference algorithm for mLDA. Table 1 provides a listing and description of the variables and parameters in the algorithm.

Algorithm 1 Online Variational Inference for 3D mLDA.

```
Define $\rho_t = (1+t)^{-\kappa}$
Initialize $\lambda, \psi, \psi'$ randomly.
for $t = 0 \to$ do
    Randomly sample a minibatch of $S$ documents from the corpus.        ▷ E step:
    Initialize $\gamma_{dk} = 1$
    For each $d$ in minibatch:
    repeat
        $\phi_{d,i,k} \propto \exp\{E[\log \theta_{d,k}] + E[\log \beta_{k,w_i}] + E[\log \psi_{k,f_i}] + E[\log \psi'_{k,f'_i}]$
        $\gamma_{d,k} = \alpha + \sum_i \phi_{d,i,k} n_{d,i}$
    until $\gamma_{d,k}$ converges
    $\hat{\lambda}_{k,w} = \eta + \sum_d \sum_{w_i=w} n_{d,i}\phi_{d,i,k}$        ▷ M step:
    $\hat{\pi}_{k,f} = \mu + \sum_d \sum_{f_i=f} n_{d,i}\phi_{d,i,k}$
    $\hat{\pi}'_{k,f'} = \mu' + \sum_d \sum_{f'_i=f'} n_{d,i}\phi_{d,i,k}$
    $\lambda = (1-\rho_t) + \rho$
    $\pi = (1-\rho_t)\pi + \rho\pi$
    $\pi' = (1-\rho_t)\pi' + \rho\pi'$
end for
```

Var	Description	Param	Relavent Expressions
$n_{d,i}$	The count of the ith word in the dth document		
w_i	The ID of the ith word in a document		
f_i, f'_i	The IDs of the ith features in a document		
ϕ	Proportional to estimation of topic assignments for each document		
θ	Doc-topic distributions (Dirichlet)	γ	$\exp[\log \theta_{d,k}] = \Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^{K} \gamma_{d,j})$
β	Topic-word distributions (Dirichlet)	β	$\exp[\log \beta_{k,w}] = \Psi(\lambda_{k,w}) - \Psi(\sum_{j=1}^{V} \lambda_{k,j})$
ψ	Topic-feature$_1$ distributions (Dirichlet)	π	$\exp[\log \psi_{k,f}] = \Psi(\pi_{k,f}) - \Psi(\sum_{j=1}^{F} \pi_{k,j})$
ψ'	Topic-feature$_2$ distributions (Dirichlet)	π'	$\exp[\log \psi'_{k,f}] = \Psi(\pi'_{k,f}) - \Psi(\sum_{j=1}^{F} \pi'_{k,j})$

Table 1: Description of variables and parameters in the algorithm.