Supplementary Material

A Optimization Problem in Finding Hidden Topics

We first show that problem (2) is equivalent to the optimization problem (4). The reconstruction of word $\mathbf{w}_i$ is $\tilde{\mathbf{w}}_i$, and $\tilde{\mathbf{w}}_i = \mathbf{W}\tilde{\boldsymbol{\alpha}}_i$ where

$$\tilde{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i \in \mathbb{R}^K}{\operatorname{argmin}} \|\mathbf{w}_i - \mathbf{H}\boldsymbol{\alpha}_i\|_2^2. \tag{11}$$

Problem (11) is a standard quadratic optimization problem which is solved by $\tilde{\boldsymbol{\alpha}}_i = \mathbf{H}^\dagger \mathbf{w}_i$, where $\mathbf{H}^\dagger$ is the pseudoinverse of $\mathbf{H}$. With the orthonormal constraints on $\mathbf{H}$, we have $\mathbf{H}^\dagger = \mathbf{H}^T$. Therefore, $\tilde{\boldsymbol{\alpha}}_i = \mathbf{H}^T\mathbf{w}_i$, and $\tilde{\mathbf{w}}_i = \mathbf{H}\boldsymbol{\alpha}_i = \mathbf{H}\mathbf{H}^T\mathbf{w}_i$.

Given the topic vectors $\mathbf{H}$, the reconstruction error E is defined as:

$$\begin{aligned} E(\mathbf{H}) &= \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i} \|\mathbf{w}_i - \mathbf{H}\boldsymbol{\alpha}_i\|_2^2 \\ &= \sum_{i=1}^{n} \|\mathbf{w}_i - \mathbf{H}\mathbf{H}^T\mathbf{w}_i\|_2^2 \\ &= \|\mathbf{W} - \mathbf{H}\mathbf{H}^T\mathbf{W}\|_2^2, \end{aligned} \tag{12}$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$ is a matrix stacked by word vectors $\{\mathbf{w}_i\}_{i=1}^n$ in a document. Now the equivalence has been shown between problem (2) and (4).

Next we show how to derive hidden topic vectors from the optimization problem (4) via Singular Value Decomposition. The optimization problem is:

$$\begin{aligned} \min_{\mathbf{H}} \quad & \|\mathbf{W} - \mathbf{H}\mathbf{H}^T\mathbf{W}\|^2 \\ \text{s.t.} \quad & \mathbf{H}^T\mathbf{H} = \mathbf{I} \end{aligned}$$

Let $\mathbf{H}\mathbf{H}^T = \mathbf{P}$. Then we have:

$$\begin{aligned} \sum_{i=1}^{n} \|\mathbf{w}_i - \mathbf{P}\mathbf{w}_i\|^2 &= \sum_{i=1}^{n} (\mathbf{w}_i - \mathbf{P}\mathbf{w}_i)^T(\mathbf{w}_i - \mathbf{P}\mathbf{w}_i) \\ &= \sum_{i=1}^{n} (\mathbf{w}_i^T\mathbf{w}_i - 2\mathbf{w}_i^T\mathbf{P}\mathbf{w}_i + \mathbf{w}_i^T\mathbf{P}^T\mathbf{P}\mathbf{w}_i). \end{aligned}$$

Since $\mathbf{P}^T\mathbf{P} = \mathbf{H}\mathbf{H}^T\mathbf{H}\mathbf{H}^T = \mathbf{P}$, we only need to minimize:

$$\sum_{i=1}^{n} (-2\mathbf{w}_i^T\mathbf{P}\mathbf{w}_i + \mathbf{w}_i^T\mathbf{P}\mathbf{w}_i) = \sum_{i=1}^{n} (-\mathbf{w}_i^T\mathbf{P}\mathbf{w}_i).$$

It is equivalent to the maximization of $\sum_{i=1}^{n} \mathbf{w}_i^T\mathbf{P}\mathbf{w}_i$.

Let $\text{tr}(\mathbf{X})$ be the trace of a matrix $\mathbf{X}$, we can see that

$$\sum_{i=1}^{n} \mathbf{w}_i^T\mathbf{P}\mathbf{w}_i = \text{tr}(\mathbf{W}^T\mathbf{P}\mathbf{W}) = \text{tr}(\mathbf{W}^T\mathbf{H}\mathbf{H}^T\mathbf{W}) \tag{13}$$

$$= \text{tr}(\mathbf{H}^T\mathbf{W}\mathbf{W}^T\mathbf{H}) \tag{14}$$

$$= \sum_{k=1}^{K} \mathbf{h}_k^T\mathbf{W}\mathbf{W}^T\mathbf{h}_k \tag{15}$$

Eq. (14) is based on one property of trace: $\text{tr}(\mathbf{X}^T\mathbf{Y}) = \text{tr}(\mathbf{X}\mathbf{Y}^T)$ for two matrices $\mathbf{X}$ and $\mathbf{Y}$.

The optimization problem (4) now can be rewritten as:

$$\begin{aligned} \max_{\{\mathbf{h}_k\}_{k=1}^K} \quad & \sum_{k=1}^{K} \mathbf{h}_k^T \mathbf{W}\mathbf{W}^T \mathbf{h}_k \\ \text{s.t.} \quad & \mathbf{h}_i^T \mathbf{h}_j = 1_{(i=j)}, \forall i, j \end{aligned} \tag{16}$$

We apply Lagrangian multiplier method to solve the optimization problem (16). The Lagrangian function L with multipliers $\{\lambda_k\}_{k=1}^K$ is:

$$\begin{aligned} L &= \sum_{k=1}^{K} \mathbf{h}_k^T \mathbf{W}\mathbf{W}^T \mathbf{h}_k - \sum_{k=1}^{K} (\lambda_k \mathbf{h}_k^T \mathbf{h}_k - \lambda_k) \\ &= \sum_{k=1}^{K} \mathbf{h}_k^T (\mathbf{W}\mathbf{W}^T - \lambda_k \mathbf{I}) \mathbf{h}_k + \sum_{k=1}^{K} \lambda_k \end{aligned}$$

By taking derivative of L with respect to $\mathbf{h}_k$, we can get

$$\frac{\partial L}{\partial \mathbf{h}_k} = 2(\mathbf{W}\mathbf{W}^T - \lambda_k \mathbf{I})\mathbf{h}_k = 0.$$

If $\mathbf{h}_k^*$ is the solution to the equation above, we have

$$\mathbf{W}\mathbf{W}^T \mathbf{h}_k^* = \lambda_k \mathbf{h}_k^*, \tag{17}$$

which indicates that the optimal topic vector $\mathbf{h}_k^*$ is the set of eigenvectors of $\mathbf{W}\mathbf{W}^T$.

The eigenvector of $\mathbf{W}\mathbf{W}^T$ can be computed using Singular Value Decomposition (SVD). SVD decomposes matrix $\mathbf{W}$ can be decomposed as $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, and $\mathbf{\Sigma}$ is a diagonal matrix. Because

$$\mathbf{W}\mathbf{W}^T\mathbf{U} = \mathbf{U}\mathbf{\Sigma}\Sigma^T\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{\Sigma}',$$

where $\mathbf{\Sigma}' = \mathbf{\Sigma}\mathbf{\Sigma}^T$, and it is also a diagonal matrix. As is seen, $\mathbf{U}$ gives eigenvectors of $\mathbf{W}\mathbf{W}^T$, and the corresponding eigenvalues are the diagonal elements in $\mathbf{\Sigma}'$.

We note that not all topics are equally important, and the topic which recover word vectors W with smaller error are more important. When $K = 1$, we can find the most important topic which minimizes the reconstruction error E among all vectors. Equivalently, the optimization in (16) can be written as:

$$\mathbf{h}_1^* = \underset{\mathbf{h}_1:\|\mathbf{h}_1\|=1}{\text{argmax}}\ \mathbf{h}_1^T\mathbf{W}\mathbf{W}^T\mathbf{h}_1 = \underset{\mathbf{h}_1:\|\mathbf{h}_1\|=1}{\text{argmax}}\ \lambda_1 \mathbf{h}_1^T\mathbf{h}_1 = \underset{\mathbf{h}_1}{\text{argmax}}\, \lambda_1 \tag{18}$$

The formula (18) indicates that the most important topic vector is the eigenvector corresponds to the maximum eigenvalue. Similarly, we can find that the larger the eigenvalue λ_k^* is, the smaller reconstruction error the topic $\mathbf{h}_k^*$ achieves, and the more important the topic is.

Also we can find that

$$\lambda_k^* = \mathbf{h}_k^{*T}\mathbf{W}\mathbf{W}^T\mathbf{h}_k^* = \|\mathbf{h}_k^{*T}\mathbf{W}\|_2^2.$$

As we can see, $\|\mathbf{h}_k^{*T}\mathbf{W}\|_2^2$ can be used to quantify the importance of the topic h_k, and it is the unnormalized importance score i_k we define in Eq. (6).

Henceforth, the K vectors in U corresponding to the largest eigenvalues are the solution to optimal hidden vectors $\{\mathbf{h}_1^*, \ldots, \mathbf{h}_K^*\}$, and the topic importance is measured by $\{\|\mathbf{h}_1^{*T}\mathbf{W}\|_2^2, \ldots, \|\mathbf{h}_K^{*T}\mathbf{W}\|_2^2\}$.

(a) Graph model to cluster senses

(b) Finite-state automaton as language analyzer

Figure 8: Topic Word Visualization to Interpret Hidden Topics

B Interpretation of Hidden Topics

Mathematically the hidden topics are orthonormal vectors, and do not carry physical meaning. To gain a deeper insight of these hidden topics, we can establish the connections between topics and words. For a given paper, we can extract several hidden topics $\mathbf{H}^*$ by solving the optimization problem (2).

For each word $\mathbf{w}_i$ in the document, we reconstruct $\tilde{\mathbf{w}}_i$ with hidden topic vectors $\{\mathbf{h}_k\}_{k=1}^K$ as below:

$$\tilde{\mathbf{w}}_i = \mathbf{H}\mathbf{H}^T \mathbf{w}_i$$

The reconstruction error for word $\mathbf{w}_i$ is $\|\mathbf{w}_i - \tilde{\mathbf{w}}_i\|_2^2$. We select words with small reconstruction errors since they are closely relevant to extract topic vectors, and could well explain the hidden topics. We collect these highly relevant words from papers in the same category, which are natural interpretations of hidden topics. The cloud of these topic words are shown in Figure 8. The papers are about graph modeling to cluster word senses in Figure 8(a). As we can see, topic words such as *graph*, *clusters*, *semantic*, *algorithms* well capture the key ideas of those papers. Similarly, Figure 8(b) presents the word cloud for papers on finite-state automaton and language analyzer. Core concepts such as *language*, *context*, *finite-state*, *transducer* and *linguistics* are well preserved by the extracted hidden topics.