6 Supplemental Material

6.1 Previous Works on Temporal Embeddings

We first introduce some notations to facilitate our discussion of time-specific embeddings in the remaining part. Suppose that the size of word vocabulary W is $|W|$ and word embeddings are of dimension m. The embedding matrix at time t is denoted as $\mathbf{V}^{(t)} \in \mathbb{R}^{m\times|W|}$, and $\mathbf{v}_{w,t} \in \mathbb{R}^m$ is the vector of word w at time t.

Existing approaches on time-specific embeddings can be divided into three categories: alignment of independently trained embedding, joint training of embeddings at different times and contextualized representations as time-sensitive sense embeddings. The first type of approaches include (Kulkarni et al., 2015), (Hamilton et al., 2016) and (Zhang et al., 2016). They pre-trained multiple sets of embeddings $\{\mathbf{V}^{(t)}\}_t$ for different times t independently. Then one set of embedding is projected to the space of another set so that two sets of embeddings are comparable.

Kulkarni et al. assumes that word vector spaces at different times are equivalent under linear transformations, and learns an alignment matrix between two sets of embeddings (Kulkarni et al., 2015). Furthermore, it assumes the local structure preservation in embeddings across time, and use word neighbors to learn the transformation matrix $\mathbf{R}$ for a word w from time t_1 to t_2. Suppose that $\mathbf{v}_{w,t}$ is the embedding of word w at time t, and $k\text{NN}(\cdot)$ gives the nearest words in the vector space.

$$\mathbf{R}_{w,t_1,t_2} = \underset{\mathbf{Q}}{\text{argmin}} \sum_{w_i \in k\text{NN}(\mathbf{v}_{w,t})} \|\mathbf{Q}\mathbf{v}_{w_i,t_1} - \mathbf{v}_{w_i,t_2}\|_2^2.$$

Based on the same assumption of space equivalence under the linear transformation, Hamilton et al. finds an alignment matrix $\mathbf{R}^{(t)} \in \mathbb{R}^{m\times m}$ consisting of basis vectors so that the mean square error between the transformed embedding at time t and embedding at time $t+1$ is minimized (Hamilton et al., 2016).

$$\mathbf{R}^{(t)} = \underset{\mathbf{Q}^T\mathbf{Q}=I}{\text{argmin}} \|\mathbf{Q}\mathbf{V}^{(t)} - \mathbf{V}^{(t+1)}\|_F.$$

Zhang et al. finds the linear transformation using anchor words whose meaning remains stable across time (Zhang et al., 2016). It requires expert knowledge to find these stable words, which limits its application to general corpora.

Different from aligning independently pre-trained embeddings, joint learning of time-stamped embeddings are shown to better capture semantic changes across time. (Bamler and Mandt, 2017; Yao et al., 2018). Bamler and Mandt uses a probabilistic language model to capture latent trajectories of word vectors across time (Bamler and Mandt, 2017). Their model is based on Bayesian skip-gram model, a probabilistic variant of word2vec. It learns embeddings $\mathbf{V}^{(t)}$ and context embeddings $\mathbf{U}^{(t)}$ at each time t. To align word embeddings across time, they add Gaussian assumption on the evolution of embeddings from time t to $t+1$. They assume that the probability of $\mathbf{U}^{(t+1)}$ conditioned on $\mathbf{U}^{(t)}$, $p(\mathbf{U}^{(t+1)}|\mathbf{U}^{(t)})$, follows Gaussian distribution. This Gaussian constraint prevents embeddings from growing large and enforces smooth vector trajectories.

Yao et al. used matrix factorization to learn an embedding matrix $\mathbf{V}^{(t)}$ and a context embedding matrix $\mathbf{U}^{(t)}$ from PPMI matrix $\mathbf{Y}^{(t)}$ with alignment constraints (Yao et al., 2018).

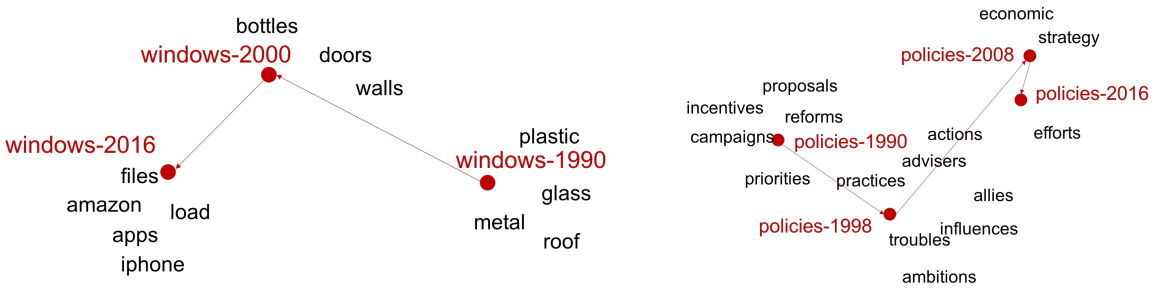
$$\begin{aligned}\mathbf{U}^*, \mathbf{V}^* = \underset{\mathbf{U}^{(t)},\mathbf{V}^{(t)}}{\text{argmin}} & \frac{1}{2}\sum_{t=1}^{T}\|\mathbf{Y}^{(t)} - \mathbf{V}^{(t)}{\mathbf{U}^{(t)}}^T\|_F^2 \\ & + \frac{\gamma}{2}\sum_{t=1}^{T}\|\mathbf{V}^{(t)} - \mathbf{U}^{(t)}\|_F^2 \\ & + \frac{\lambda}{2}\sum_{t=1}^{T}\|\mathbf{V}^{(t)}\|_F^2 + \frac{\tau}{2}\sum_{t=2}^{T}\|\mathbf{V}^{(t-1)} - \mathbf{V}^{(t)}\|_F^2 \\ & + \frac{\lambda}{2}\sum_{t=1}^{T}\|\mathbf{U}^{(t)}\|_F^2 + \frac{\tau}{2}\sum_{t=2}^{T}\|\mathbf{U}^{(t-1)} - \mathbf{U}^{(t)}\|_F^2,\end{aligned}$$

where terms $\|\mathbf{V}^{(t-1)} - \mathbf{V}^{(t)}\|_F^2$ and $\|\mathbf{U}^{(t-1)} - \mathbf{U}^{(t)}\|_F^2$ are alignment constraints on embeddings in neighboring time periods.

The third category of temporal embedding models are built upon pre-trained language models such as BERT (Devlin et al., 2018). These models are pre-trained on large corpus to learn representations for words in a given context, which can be taken as the sense embedding. Time-specific word semantics are treated as different senses of words, and thus are represented by the contextualized representations (Hu et al., 2019; Giulianelli et al., 2020).

6.2 Model and Optimization Problem

A model of static word embeddings is proposed by Arora et al., and provides a unified understanding of a group of embedding models including pointwise mutual information (PMI) method, word2vec



(a) Word "windows"

(b) Word "policies"

Figure 2: The trajectory of word embeddings over time.

and GloVe (Arora et al., 2016). It reveals that all these models train embeddings to estimate word co-occurrences in the training corpus. Suppose that $\mathbb{P}_s(w_1, w_2)$ is the co-occurrence probability of words w_1 and w_2 in context window of size s, $\mathbf{v}_{w_1}$ and $\mathbf{v}_{w_2}$ are word vectors. Their model states that

$$\log \mathbb{P}_s(w_1, w_2) = \frac{\|\mathbf{v}_{w_1} + \mathbf{v}_{w_2}\|_2^2}{2d} - 2\log Z + \gamma \pm \epsilon, \tag{12}$$

where d is the embedding dimension, Z is a constant, $\gamma = \log\left(\frac{s(s-1)}{2}\right)$ and ϵ is an error term. We consider the window size s to be a constant. Since the coefficient $\frac{1}{d}$ can be absorbed as a constant scale of the word vectors, the model suggests the approximation of the logarithm of word co-occurrence probability below:

$$\log \mathbb{P}(w_1, w_2) = \frac{1}{2}\|\mathbf{v}_{w_1} + \mathbf{v}_{w_2}\|_2^2 + \tau,$$

where constant $\tau = -2\log Z + \gamma$.

We propose a model for condition-specific word embeddings in condition c, where c can be time or location.

$$\log \mathbb{P}_c(w_1, w_2) = \frac{1}{2}\|\mathbf{v}_{w_1,c} + \mathbf{u}_{w_2,c}\|_2^2 + \tau.$$

Since we assume that $\mathbf{v}_{w,c} = \mathbf{v}_w \odot \mathbf{q}_c + \mathbf{d}_{w,c}$, we can substitute $\mathbf{v}_{w_1,c}$ with $\mathbf{v}_{w_1} \odot \mathbf{q}_c + \mathbf{d}_{w_1,c}$, and substitute $\mathbf{u}_{w_2,c}$ with $\mathbf{u}_{w_2} \odot \mathbf{q}_c + \mathbf{d}'_{w_2,c}$. We then have

$$\begin{aligned}
&\log \mathbb{P}_c(w_1, w_2) \\
&= \frac{1}{2}\|(\mathbf{v}_{w_1} \odot \mathbf{q}_c + \mathbf{d}_{w_1,c}) + (\mathbf{u}_{w_2} \odot \mathbf{q}_c + \mathbf{d}'_{w_2,c})\|^2 + \tau, \\
&= \frac{1}{2}\left(\|\mathbf{v}_{w_1} \odot \mathbf{q}_c + \mathbf{d}_{w_1,c}\|^2 + \|\mathbf{u}_{w_2} \odot \mathbf{q}_c + \mathbf{d}_{w_2,c}\|^2\right) \\
&+ (\mathbf{v}_{w_1} \odot \mathbf{q}_c + \mathbf{d}_{w_1,c})^T(\mathbf{u}_{w_2} \odot \mathbf{q}_c + \mathbf{d}'_{w_2,c}) + \tau, \\
&= \frac{1}{2}(\|\mathbf{v}_{w_1,c}\|^2 + \|\mathbf{u}_{w_2,c}\|^2 + 2\tau) + \\
&(\mathbf{v}_{w_1} \odot \mathbf{q}_c + \mathbf{d}_{w_1,c})^T(\mathbf{u}_{w_2} \odot \mathbf{q}_c + \mathbf{d}'_{w_2,c}).
\end{aligned} \tag{13}$$

Here $\frac{1}{2}\|\mathbf{v}_{w_1,c}\|^2$ can be taken as the bias related to words w_1 under condition c. Similar to GloVe, we introduce a bias term $b_{w_1,c}$, and $b_{w_1,c} = \frac{1}{2}\|\mathbf{v}_{w_1,c}\|^2 + \tau$. Similarly, we define bias $b'_{w_2,c} = \frac{1}{2}\|\mathbf{v}_{w_2,c}\|^2 + \tau$.

The logarithm of co-occurrence probability is:

$$\begin{aligned}
&\log \mathbb{P}_c(w_1, w_2) \\
&= (\mathbf{v}_{w_1} \odot \mathbf{q}_c + \mathbf{d}_{w_1,c})^T(\mathbf{v}'_{w_2} \odot \mathbf{q}_c + \mathbf{d}'_{w_2,c}) \\
&+ b_{w_1,c} + b'_{w_2,c}.
\end{aligned} \tag{14}$$

Our model has two sets of basic word vectors $\{\mathbf{v}_w\}$ and $\{\mathbf{u}_w\}$, two sets of deviation embeddings $\{\mathbf{d}_{w,c}\}$ and $\{\mathbf{d}'_{w,c}\}$, and two sets of bias terms $\{b_{w,c}\}$ and $\{b'_{w.c}\}$ for the vocabulary and context vocabulary respectively. All of these parameters are trained to minimize the difference between the real word co-occurrences and the estimated values.

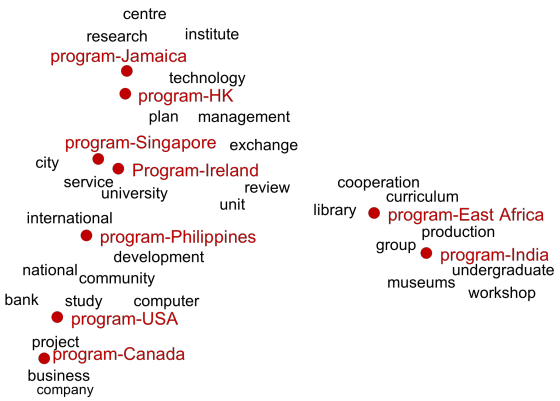


Figure 3: The trajectory of word embeddings over locations.

6.3 Word Embedding Trajectory

The trajectories of word *windows* and *policies* across time are shown in Fig. 2 (a) and (b). As we can see, *windows* was commonly used as an opening in houses to allow light and air before

the 20th century since its embedding has neighbors such as *glass* and *walls*. Recently it also refers to an operating system developed by Microsoft given its neighbors *files* and *load*. As for the word *policies*, it was relevant to *campaigns* and *reforms* in 1990, since many campaigns and reforms were launched in that cold-war era. Its meaning has shifted to be relevant to *economic* over time.

In Fig. 3, we plot the location-specific neighbors of word *program*. We note that *program* is a polysemous words with senses including *project*, *software* and *curriculum*. People in one region use it to refer to a sense that is different from another sense used in another region. The different senses can be inferred from its region-specific neighbors. For example, the *program* is meant as projects or business in Canada, while it is also related to computer software in USA. East Africa and India use it to refer to curriculum in the education domain.